



# Deliverable D3.5

## Final evaluation report

**The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 287678.**


| Participant no. | Participant organisation name | Part. short name | Country |
|---|---|---|---|
| 1 (Coordinator) | University of Edinburgh | UEDIN | UK |
| 2 | Aalto University | AALTO | Finland |
| 3 | University of Helsinki | UH | Finland |
| 4 | Universidad Politécnica de Madrid | UPM | Spain |
| 5 | Technical University of Cluj-Napoca | UTCN | Romania |

| | |
|---|---|
| **Project reference number** | FP7-287678 |
| **Proposal acronym** | SIMPLE4ALL |
| **Status and Version** | Complete, proofread, ready for delivery: version 1 |
| **Deliverable title** | Final evaluation report |
| **Nature of the Deliverable** | Report (R) |
| **Dissemination Level** | Public (PU) |
| **This document is available from** | `http://simple4all.org/publications/` |
| **WP contributing to the deliverable** | WP3 |
| **WP / Task responsible** | WP3 / T3.5 |
| **Editor** | Tuomo Raitio (AALTO ) |
| **Editor address** | tuomo.raitio@aalto.fi |
| **Author(s), in alphabetical order** | Paavo Alku, Roberto Barra-Chicote, Jaime Lorenzo-Trueba, Juan M. Montero, Tuomo Raitio |
| **EC Project Officer** | Leonhard Maqua (effective from 12th March 2014; project month 29) |

# Abstract

This deliverable provides a summary of several evaluations of our synthesis system and describes experiments in the synthesis of expressive speech conducted in WP3. A new vocoding method was developed, based on predicting time-domain glottal excitation waveforms directly from acoustic features with a deep neural network (DNN). The method was first evaluated with normal speaking style and then used with adaptation for the synthesis of breathy, normal, and Lombard speech. The new vocoder reached the quality level of our existing GlottHMM vocoder and in synthesis of Lombard speech the DNN-based method outperformed the current GlottHMM system. A joint evaluation study was conducted with researchers of a parallel EC-funded project on synthesis of laughter. The topic was challenging, as expected, and showed that for statistical synthesis of laughs some of the existing vocoders are not robust enough in the presentation of speech parameters. Evaluation studies were also conducted to better understand the perceptual importance of the periodic and aperiodic components of the vocoder excitation. Evaluation results obtained with a specifically designed generalised vocoder indicate that in combining periodic excitation waveforms with a periodic noise sequences, the impact of spectral weighting is essential while the role of temporal noise envelope is small. Finally, two methods for emotion transplantation, based on cross-speaker extrapolation and cross-speaker model adaptation, were studied. Evaluations conducted in the synthesis of emotional speech indicate, for example, a very clear preference (reaching as high as 96% for happiness, an average of 87% preference across all the emotions) for the emotional synthesiser we have developed.

# Contents

# 1 Introduction

This deliverable describes our latest work on synthesis evaluation, done in task T3.5 "Expressive speech synthesis. Evaluation". The research conducted involves evaluations of many versions of the system, as stated in the project plan. These multiple evaluations are reported in this document as follows. First, our new vocoding system, based on predicting the glottal excitation waveform with a deep neural network (DNN) directly from acoustic features, is described by reporting the evaluation results from two experiments. Second, a vocoder evaluation on a challenging style of vocalisation, laughter, is described. Third, evaluation of the impact of the periodic and aperiodic components of the vocoder excitation on synthesis quality is reported. Fourth, emotion transplantation using two techniques is described. Finally, conclusions regarding these evaluations are given.

# 2 Using deep neural networks in modelling of the voice source in vocoding

In our previous vocoding studies, we generated the excitation waveform of the synthesiser by using either a glottal pulse waveform estimated from real speech [1, 2] or by combining several pulses with principal component analysis (PCA) [3]. As an alternative, we devised a method in which the time-domain voice source waveform at the synthesis stage is predicted with a deep neural network (DNN) directly from acoustic features. The flow chart of the method is shown in Fig. 2.0a and the technique consists of the following main steps. First, acoustic features and the glottal flow signal are estimated from each frame of the speech database. Acoustic features consist of the following components: energy, fundamental frequency, harmonic-to-noise ratio, voice source spectrum, and vocal tract spectrum. Pitch-synchronous glottal flow time-domain waveforms are extracted, interpolated to a constant duration, and stored (in a codebook) as the training set for the DNN. Then, a DNN is trained to map from acoustic features to these duration-normalised glottal waveforms. At synthesis time, acoustic features are generated from a statistical parametric model, and from these, the trained DNN predicts the glottal flow waveform.

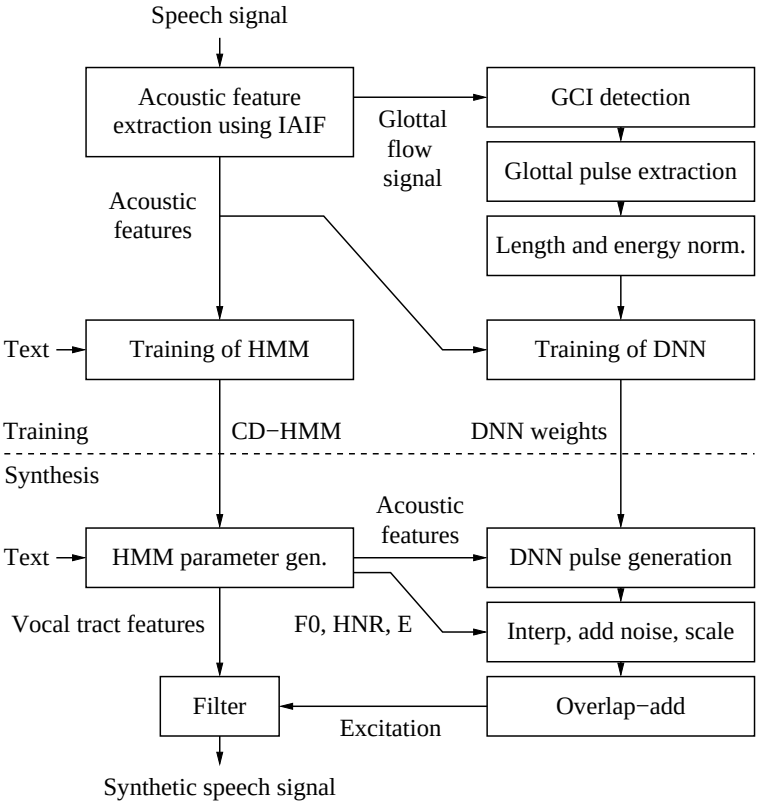



Figure 2.0a: The novel vocoder utilising DNN-based prediction of the voice source waveform was evaluated in two subjective listening tests which are briefly described below (from the publication in Appendix 6.1).

## 2.1 DNN-based voice source modelling: Comparison to the conventional GlottHMM with normal speaking style

Two Finnish speech databases, one of a male speaker and the other one of a female speaker, were used in the experiment. The male voice comprises 600 sentences (approx. 1 h) and the female database comprises 500 sentences. Waveforms for both voices are sampled at 16 kHz. GlottHMM [2, 3] was used for extracting the acoustic features and the glottal flow signal. Glottal flow pulse codebooks were constructed for both databases in order to train the DNN-based voice source model. The codebooks contained 203,172 or 203,768 pulses for the male or

female speaker, respectively. Additionally, smaller codebooks were constructed for both speakers from 20 sentences of speech material, in order to implement the alternative method in which the DNN output is used to select a natural pulse from the codebook; these codebooks consisted only of 7,495 and 8,131 pulses in order to minimise computational cost at synthesis time. The standard HTS 2.1 method [4] was used for training the HMM-based system.

The DNN took as input a 47-dimensional vector composed of the extracted acoustic speech features. The target output of the network was a 400 sample duration-normalised glottal flow pulse. In order to determine the optimal number of layers and hidden units for the DNN, six different systems were trained by varying the number of hidden layers (from 1 to 3) and the number of units per layer (from 800 to 1200). Unsupervised restricted Boltzmann machine (RBM) pre-training was tried for one of the configurations. 200,000 training examples were used for training with 3,000 examples held out for cross-validation. The results showed that the best performance was achieved with 2 hidden layers and 1000 units per hidden layer, with RBM pre-training slightly improving performance.

An online subjective evaluation was carried out to assess the proposed method. Three different vocoding techniques were compared: 1) Conventional GlottHMM synthesis [2] using a single natural glottal flow pulse, of which spectrum is matched according to the voice source LSF, 2) DNN-based voice source modelling, and 3) DNN-based voice source model used as a target cost for selecting natural glottal flow pulses from a small library. A comparison category rating (CCR) test was used with a discrete, seven-point scale ranging from $-3$ to 3. A group of 26 people (15 Finnish and 11 non-Finnish) participated in the evaluation.

Results of the evaluation are shown in Fig. 2.1a indicating that there were no statistically significant quality differences between the three methods. Given the fact that this was the first experiment on a new, greatly different vocoding principle, we consider the result encouraging: the quality achieved was rated equal to that of an established baseline. This first result (published in the paper in Appendix 6.1) motivated us to develop the method further and to conduct new evaluations on different speaking styles described next .

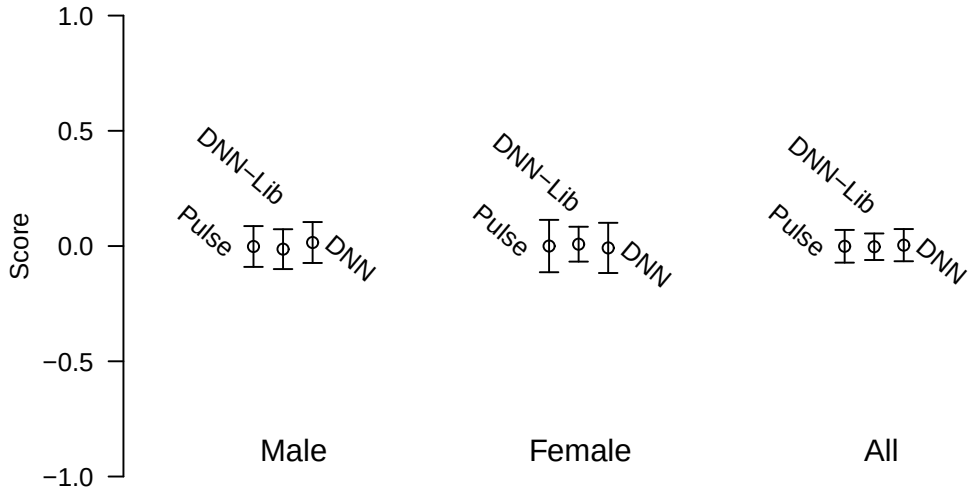



Figure 2.1a: Results of the evaluation of the DNN-based voice source modelling methods in comparison to a method using natural glottal flow waveforms (from the publication in Appendix 6.1).

## 2.2 DNN-based voice source modelling: Evaluation with varying vocal effort

Our second experiment on DNN-based modelling (published in the paper in Appendix 6.2) of the voice source used data, one male voice and one female voice, from two speech corpora [5]. For both speakers, three different vocal

effort levels were utilised: breathy, normal, and Lombard. The normal style consists of 1450 sentences, comprising approximately two hours of speech for both speakers. Lombard speech was elicited by playing babble noise with 80 dB SPL to the speaker's ears through headphones while recording, and feeding back the speakers own voice through headphones, corresponding to a level of speaking in a normal room without headphones. The Lombard style consists of 300 sentences. The breathy speaking style was elicited by increasing the level of the speakers feedback through headphones as well as instructing the subjects to speak softly without whispering. 200 sentences were read in the breathy style. The recording and processing of the speech data are described in more detail in [5].

The DNN was trained as described in Sec. 2.1 except that a 2-hidden-layer structure was used with 100 and 200 neurons in the first and the second hidden layers, respectively. Within a reasonable number of training epochs, the network achieved much lower errors than the DNN architecture described in the first experiment.

The HMM training and adaptation procedures were identical to the experiments done in [5]. The training of the normal voices followed the standard HTS method [4]. Speech features described were extracted using the GlottHMM vocoder [2] and delta, and delta-delta features were added. Semi hidden Markov models were used as acoustic models, and features were trained in individual streams except that the vocal tract LSFs and energy were trained together.

In order to create the low and high vocal effort voices (breathy and Lombard), the normal voice models were adapted with the constrained structural maximum a posteriori linear regression combined with maximum a posteriori adaptation (CSMAPLR+MAP) technique [6]. The speaker-dependent voice source model DNNs were trained using all speech material including breathy, normal, and Lombard speech.

Subjective evaluations were conducted using three vocal effort levels. A high-quality mean glottal flow pulse excitation scheme was selected for a reference baseline system, which has been successfully used in synthesising speech with varying vocal effort [5]. The baseline system uses a style-specific mean glottal flow pulse for each of the three styles [5] (corresponding to the PCA-based excitation in [7]), and a spectral matching scheme [2, 5], where a pole-zero filter is used to filter the excitation signal in order to apply the desired spectral properties defined by the generated voice source spectrum. Two types of tests were conducted to compare the proposed and the baseline systems: a comparison category rating (CCR) test was conducted to evaluate the speech quality, and a similarity test was conducted in order to assess the speaker and style similarity between the two methods. 14 native Finnish listeners participated in both tests.

The mean scores of the quality test are shown for each vocal effort level with 95% confidence intervals in Fig. 2.2a. Only for Lombard speech is the difference between the two methods statistically significant, with the proposed method being rated higher in quality.

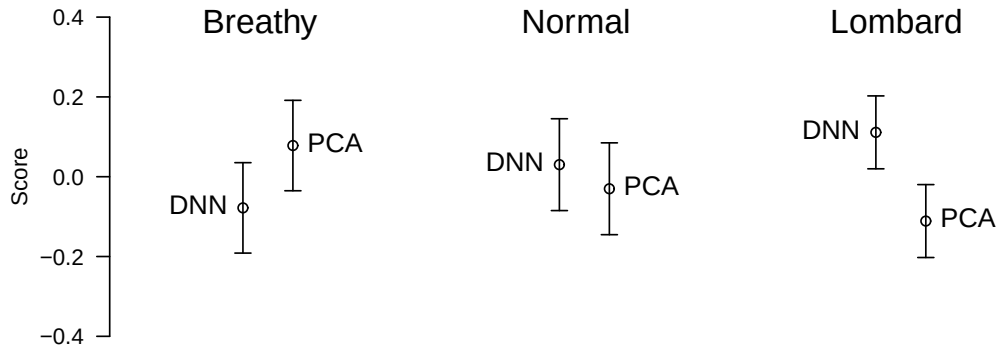



Figure 2.2a: Results of the quality test comparing DNN and PCA-based excitation methods with breathy, normal, and Lombard speech (from the publication in Appendix 6.2).

# 3 Vocoder evaluation for HMM-based synthesis of laughter

Laughter is non-verbal vocalisation that plays an essential role in our daily conversations. It conveys information about emotions and fulfils important social functions. Integrating laughter into a speech synthesis system can bring the synthetic speech closer to natural human conversational speech. Synthesis of laughter is, however, challenging and there is little previous research on the topic, except for a few studies [8, 9, 10, 11].

Bridging the gap between knowledge on human laughter and its use by avatars is a goal of a parallel EC-funded project, ILHAIRE (no. 270780). Since researchers in SIMPLE4ALL had connections to colleagues working on the ILHAIRE project, it was possible to conduct an experiment on laughter synthesis jointly between the two projects. This investigation is briefly described below.

The study on laughter synthesis focused on comparing four vocoders that are commonly used in HMM-based speech synthesis. The selected vocoders were: 1) Impulse train-excited mel-cepstrum-based vocoder (MCEP), 2) STRAIGHT [12, 13] with mixed excitation, 3) Deterministic plus stochastic model (DSM) [14], and 4) GlottHMM [2]. All vocoders use the source-filter principle for synthesis, and thus there are two components that mostly differ among the systems: the type of spectral envelope extraction and representation, and the method for modelling and generating the excitation signal.

A subjective evaluation was carried out to compare the performance of the vocoders in synthesising natural laughs. For each vocoder, two types of samples were used: a) copy-synthesis, which consists of extracting the parameters from a laugh signal and re-synthesising the same laugh from the extracted parameters; b) HMM-based synthesis, where an HMM-based system is trained from a laughter database and laughs are then synthesised using the models corresponding to phonetic transcriptions of natural laughter. Copy-synthesis can be seen as the theoretically best synthesis that can be obtained with a particular vocoder, while HMM-based synthesis shows the current performance that can be achieved when synthesising new laughs. Natural human laughs were also included in the evaluation as a reference. The data of the evaluation consisted of one male (64 laughs) and one female (54 laughs) voice from the AVLaughterCycle database. The subjective evaluation was carried out using a web-based listening test, where listeners were asked to rate the synthesised laughter signals on a 5-point Likert scale [15]. Participants could listen to the laugh as many times as they wanted and were asked to rate its naturalness on a 5-point Likert scale where only the highest (completely natural) and lowest (completely unnatural) options were labelled.

Overall, the results show that all vocoders perform relatively well in copy-synthesis. However, in HMM-based laughter synthesis, all synthesised laughter voices were significantly lower in quality than in copy-synthesis. The evaluation results revealed that two vocoders (MCEP, DSM) using rather simple and robust excitation modelling performed best, while two other vocoders (STRAIGHT, GlottHMM) using more complex analysis, parameter representation, and synthesis suffered from the statistical modelling. These findings (published in the paper provided as Appendix 6.3) suggest that the robustness of parameter extraction and representation is a key factor in laughter synthesis, and increased efforts should be directed to enhancing the robust estimation and representation of the acoustic parameters of laughter. More details on the comparisons conducted can be found in [16].

# 4 Evaluation of the periodic and aperiodic components in excitation modelling

In order to reduce buzziness, various excitation models have been proposed for HMM-based speech synthesis in recent years. In different vocoders, excitation (of voiced speech) is typically modelled by a mixture of periodic and aperiodic components. There is, however, little knowledge on how the two components separately contribute to the naturalness and quality of the synthesis. In order to address this research question, a generalised mixed excitation modelling tool was built to study how the synthesis quality is affected by the following three excitation factors: periodic waveform, noise spectral weighting, and noise time envelope.

The workflow of this generalised vocoder is displayed in Fig. 4.0b. The periodic contribution of excitation $e_p(t)$ is obtained from a specific waveform whose duration is adapted to the current $F_0$ value, and which is then filtered using some aperiodicity measurements. The aperiodic excitation component, $e_a(t)$, is generated from white

Gaussian noise that is spectrally modified with the aperiodicity measurements but also temporally modulated using a given time-envelope. Note that all this processing is done pitch-synchronously. The two components $e_p(t)$ and $e_a(t)$ are then summed up and the pitch-synchronous windowed frames are overlap-added. The resulting excitation is finally filtered to get the speech signal. The three main factors (periodic waveform, noise spectral weighting and noise envelope) impacting the performance of this generalised excitation model are next described.

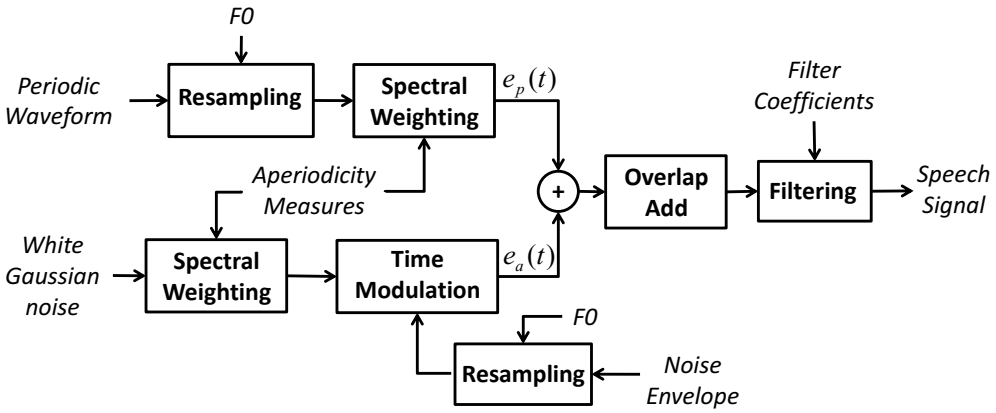



Figure 4.0b: Workflow of generalised vocoder using mixed excitation (from publication in Appendix 6.4).

## 4.1 Periodic waveform

Three variants for the periodic waveform were used: i) Dirac impulse as used in the simplest vocoders; ii) a natural excitation residual frame; iii) speaker-dependent eigenresidual as proposed in [14]. Note that the choice of the natural residual frame was not arbitrary and resulted from the consideration of several criteria: a) having a low pitch to avoid as much as possible up-sampling to the target $F_0$ (as this will cause energy holes in high frequencies); b) its amplitude spectrum must be as flat as possible to avoid artefacts due to residual resonances; c) having a clear discontinuity at the glottal closure instant (GCI).

## 4.2 Spectral weighting

In order to reduce buzziness caused by an overly strong harmonic structure, it has proven beneficial in HMM-based synthesis to adopt an approach in which both the periodic and aperiodic component may coexist [17]. This can be implemented by using, for example, a multiband approach where the energy of the periodic and aperiodic component is controlled for each frequency band by aperiodicity measurements [13]. Four options for spectral weighting were investigated: i) the aperiodic component is discarded and the excitation consists only of the periodic contribution; ii) use of a static maximum voiced frequency $F_m$ fixed to 4 kHz as is done in [18] and [14]; iii) use of dynamic $F_m$ value estimated using the algorithm described in [19]; iv) use of the HNR measurements proposed in [3].

## 4.3 Envelope for noise modulation

In addition to the spectral characteristics, also the temporal properties of the aperiodic component may affect the synthesis quality. The motivation for considering this possibility stems from the observation that the time distribution of the aperiodic component is not uniform over the fundamental period but rather time-synchronised with the different sections of the glottal cycle. Three temporal noise envelopes were studied: i) uniform distribution; ii) the triangular window proposed in [19]; iii) the speaker-dependent Hilbert envelope proposed in the DSM approach [14].

### 4.4 Subjective evaluation

Subjective evaluation was performed in three separate steps in order to uncover the effect of each component and also their possible interactions. The idea was to first select the best noise spectral weighting according to a subjective evaluation among the four systems. Then, the best spectral weighting method according to the first evaluation is used to study the effect of the noise time envelope, in which three systems are evaluated. Finally, in the third test, both the best noise spectral weighting and the best time envelope are used in the study of the effect of the periodic waveform, in which three systems are compared. Comparison Category Rating (CCR) tests were used in order to determine the quality difference between the systems.

### 4.5 Results

Results clearly indicate that: i) the spectral weighting is an essential feature as it leads to the greatest perceptual differences; ii); incorporating a noise model during the production of voiced sound is crucial. This can be efficiently achieved based on HNR measures or using a maximum voiced frequency; iii) the perceptual impact of the noise envelope seems to be negligible; iv) it is necessary to adapt the periodic waveform according the speaker's $F_0$ range as it will affect the excitation phase properties. These conclusions (published in the paper in Appendix 6.4) should be carefully considered when designing new excitation models. As a result, we believe that future research efforts should focus on new strategies to weight the energy of both periodic and aperiodic components in several spectral bands, as well as on a better understanding of the phase information in the periodic waveform.

# 5 Emotion transplantation

One of the advantages of Parametric Speech Synthesis over Unit Selection methods is the capability to modify paralinguistic and extralinguistic properties of the synthetic speech (such as the speaking style, emotion or age) by using interpolation or adaptation techniques. However, these techniques have been used within a certain multi-emotion or multi-style corpus, they have not been used for non-averaged cross-speaker modifications.

Task 3.5 in this project addresses the design and test of cross-speaker expression transplantation methods. Generally speaking, the objective is to take information from one corpus containing both expressive and non-expressive recordings of one or several speakers and to use those pieces of information to add expressiveness to non-expressive models of other speakers from other corpora.

The first technique evaluated is based on cross-speaker extrapolation: post-processing the output speech parameters using an appropriate set of extrapolation weights for the different streams of the speech synthesiser.

The second approach is based on cross-speaker model adaptation: learning the linear adaptation functions from the expressive source corpus and applying those functions to the non-expressive models of the other corpus.

## 5.1 Cross-speaker emotion extrapolation

Although it would be possible to interpolate emotional speech models of an emotional voice and a non-emotional voice using the standard technique, the similarity to the target speaker would decrease, as the target speaker would be mostly identified as the emotional source speaker.

However, we can modify the basic interpolation method and thus propose cross-speaker extrapolation of acoustic emotional patterns to other new target speakers (for whom we have no emotional speech training data): the acoustic emotional patterns can be learnt as deviations (linear functions) of emotional speech models of a source speaker from his or her neutral model. The details of the proposed methods can found paper on emotion extrapolation provided as Appendix 6.5. In order to minimise speech artefacts, the extrapolation was limited to the stable central states of the models.

In the evaluation experiments we tested several linear extrapolation factors (such as 0.5, 0.75, 1.0 or 1.25) and an ad-hoc configuration (with different weights for the different streams and emotions). The extrapolated emotions were sadness, anger, fear and surprise. The MOS-based Speech Quality (SQ) of the neutral synthetic source and target voices were 3.2 and 3.4 respectively; the Emotion Identification Rates (EIR) were 69% and 86% and the Speaker Identification Rates (SIR) were 69% and 93%.

When extrapolating emotions, SQ decreases however large the extrapolation factor is, because the models are deviating more and more from the trained target neutral models (Figure 3 in Appendix 6.5). Only k = 0.5 extrapolation factor obtains no statistically significant different SQ values when compared to the SQ of the source speaker. The adhoc extrapolation scheme slightly reduced this SQ degradation (equivalent to 0.75, in spite of being greater on average).

Regarding the Emotional Strength (ES) perceived by the listeners, we have normalised the scores on a per listener basis, to minimise the bias. The obtained ES scores are significantly higher than the neutral ones however large the extrapolation factor k is (Figure 4 in Appendix 6.5).

EIRs are shown in Figure 5 in Appendix 6.5. Using extrapolation schemes with k higher than 0.5, sadness and fear synthetic speech of the transformed speaker are identified (69% using k equal to 1.0 for both emotions). However, EIR is low for surprise and anger, although the adhoc scheme achieves a rate of 53% for anger).

SIRs are shown in Figure 6 in Appendix 6.5. The transformed speaker is notably identified as the target speaker or at least as another speaker different from the source speaker. The ad-hoc scheme obtains a good compromise between a high target SIR (40%), a lower source SIR (23%) and a 37% identified as other speaker (neither option).

## 5.2 Cross-speaker emotion transplantation

Some instability problems when extrapolating to other speakers could be due to the use of a post-processing technique based on linear transformations which are not estimated during the training process, but computed afterwards.

A transplantation method that has been introduced lately is based on Cluster Adaptive Training (CAT), a projective adaptation technique. As such, it is only capable of producing speaker models based on linear combinations of the original training speaker models. The main advantage of this approach is that – because the produced model is always a combination of preexisting training models – the process is extremely robust, outputting very high quality speech. On the other hand, the level of expressive strength or speaker similarity cannot be guaranteed as the transplantation 'reach' is highly constrained.

The new proposed emotion transplantation method uses a cascade of adaptations (one for transforming the speaker and another one for transforming the emotion) to lessen speech quality degradation, while using the adaptation functions as pseudo-rules for modifying the speaker models. As a result, the new method can control the expressive strength while maintaining reasonable speech quality and speaker identifiability when compared to non-transplanted expressive synthetic speech. The method has three steps: to adapt the reference emotion from an average voice model obtained by applying Speaker Adaptive Training (SAT); to adapt the target speaker model and the target emotion from the reference emotion; and, finally, to apply in cascade the emotion and speaker identity transformations to the reference emotion. The resulting model is of the emotional target speaker. A possible alternative design is to merge all our emotional data into a richer average emotion model.

The method can be used for cross-speaker transplantation, but also for intra-speaker interpolation [20] and was successfully used for creating interpolated emotional synthetic speech in [21].

Two different evaluations were carried out, a first one that compared the proposed emotion transplantation system with the traditional neutral synthetic voice to validate the transplantation method, and a second one that compared the neutral synthetic voice with an average emotion transplanted into the speaker (alternative design) in order to verify that the benefits of transplanting the correct emotion into the speaker are higher than just modifying the neutral speech to sound less machine-like. Four emotions (anger, happiness, sadness and surprise) learnt from the Spanish Emotional Voices corpus were transplanted into 3 male speakers and 3 female speakers. It was a binary preference test (the options were transplanted correct emotion vs. neutral voice or the transplanted average emotion vs. neutral voice) which included a MOS Speech Quality (SQ) and Emotional Strength (ES) for both samples). The results are shown on Table 1 in Appendix 6.6. Both transplanted emotional models and transplanted averaged are preferred over neutral models of the target speakers for every emotion, but the preference is 12% higher for the transplanted non-averaged models, while SQ is only 0.2 points lower and ES is 0.5 points higher.

# 6 Conclusions

This deliverable has described evaluation experiments and emotion transplantation studies conducted in task T3.5. The research was divided into four main parts whose main findings are summarised below.

We developed a potential new vocoding scheme based on computing the glottal excitation at the synthesis stage, directly from acoustic parameters, using a DNN. The method was used in two experiments. In the first one, synthesis of a normal speaking style was studied and the DNN-based method achieved synthesis quality that was equal to that of a high-quality benchmark system. In the second experiment, HMM-based synthesis of three styles (breathy, normal, Lombard) was studied together with adaptation. Results show that the DNN-based method resulted in a statistically significant quality improvement compared to a previously developed vocoder but a significant improvement was observed only for the Lombard speech. Results were slightly surprising because we discovered that the DNN is sometimes unable to predict abrupt transients in time-domain glottal excitations near instants of glottal closure and therefore we expected to get a quality improvement in the synthesis of breathy speech rather than in the synthesis of Lombard speech. It is, though, worth emphasising that these two experiments are our first investigations into DNN-based mapping from acoustic features to glottal waveforms and until now we have not analysed in detail, for example, the choice of those acoustic features.

In addition to the DNN-based vocoder, synthesis evaluations were conducted on laughter and on the impact of periodic and aperiodic components of the excitation. As expected, the former turned out to be difficult, and we found that none of the HMM-based synthesisers compared was able approach the quality of copy-synthesis. Experiments also showed that vocoders using simple excitations were rated as better than those based on more advanced excitation modelling. For better synthesis of laughter, increased robustness is needed, especially in more advanced vocoding methods. In studying the impact of periodic and aperiodic components of the excitation, a generalised vocoder was built that enables investigating the perceptual importance of different factors that can be used to avoid buzziness in HMM-based speech synthesis. Evaluation results indicate that in combining periodic excitation waveforms with a periodic noise sequences, the impact of spectral weighting is large while the role of temporal noise envelope is small.

Two methods for emotion transplantation, based on cross-speaker extrapolation and cross-speaker model adaptation, were developed. In the first one, the extrapolation of emotional acoustic patterns was defined to incorporate emotional content into new or previously-neutral synthetic voices. A perceptual test was conducted, where the speech quality, the emotional strength, emotional identification rates and speaker identity rates were evaluated. The acoustic emotional models of four emotions (anger, surprise, sadness and fear) were trained from an emotional female voice and extrapolated to a new synthetic neutral female voice. The emotional patterns over each speech component (spectra, log F0, aperiodicity bands and durations) were considered in the acoustic emotional model. With the proposed algorithm, acoustic emotional patterns are partially extrapolated to a target speaker without losing the target speaker identity. The strength of the emotion extrapolation can be modified successfully by varying the extrapolation factor. However, the strength of the extrapolation was found to have a negative impact on the resulting speech quality, especially in the extrapolation of the emotional patterns of the spectral component.

In the second part of our work on emotion transplantation, the aim was to learn the paralinguistic nuances of any particular emotion in order to transplant them into a new target speaker for whom only traditional, neutral read speech recordings are available. This is done by means of chaining a pair of adaptation functions, one that characterises the target speaker identity and another that defines the paralinguistic characteristics of the desired emotion. Finally, two perceptual evaluations were carried out. For these perceptual evaluations, four emotions (anger, happiness, sadness and surprise) from a Spanish emotional database and six target speakers (three male and three female) were considered. A first evaluation compared in terms of naturalness, speech quality and emotional strength the proposed transplantation method with traditional neutral read speech synthesis. This first test showed that there is a very clear preference (an average of 87% preference between all the emotions) for the emotional synthesiser, reaching as high as 96% for happiness, and a perceived increase in emotional strength of an average of 1.2 points in the MOS scale at a cost of only 0.4 points in speech quality. The second test compared an average

emotion transplantation with the neutral speech, and showed that just by adding an undefined 'emotional colour' to the voice we are able to improve the perceived naturalness of the synthetic speech up to an average of 75% preference at a cost of only 0.2 points in speech quality, although the average increase in perceived emotional strength only reaches 0.7 points.

# References


[1] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku. HMM-based Finnish text-to-speech system utilizing glottal inverse filtering. In *Proc. Interspeech*, pages 1881–1884, 2008.

[2] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku. HMM-based speech synthesis utilizing glottal inverse filtering. *IEEE Trans. on Audio, Speech, and Lang. Proc.*, 19(1):153–165, 2011.

[3] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku. Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis. In *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, pages 4564–4567, 2011.

[4] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda. The HMM-based speech synthesis system (HTS) version 2.0. In *Sixth ISCA Workshop on Speech Synthesis*, pages 294–299, 2007.

[5] T. Raitio, A. Suni, M. Vainio, and P. Alku. Synthesis and perception of breathy, normal, and lombard speech in the presence of noise. *Computer Speech & Language*, 28(2):648–664, 2014.

[6] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. *IEEE Trans. on Audio, Speech, and Lang. Proc.*, 17(1):66–83, 2009.

[7] T. Raitio, A. Suni, M. Vainio, and P. Alku. Comparing glottal-flow-excited statistical parametric speech synthesis methods. In *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, pages 7830–7834, 2013.

[8] S. Sundaram and S. Narayanan. Automatic acoustic synthesis of human-like laughter. *J. Acoust. Soc. Am.*, 121(1):527–535, 2007.

[9] E. Lasarcyk and J. Trouvain. Imitating conversational laughter with an articulatory speech synthesis. In *Proc. of Interdisciplinary Workshop on the Phonetics of Laughter*, pages 43–48, Saarbrücken, Germany, 2007.

[10] T. Sathya Adithya, K. Sudheer Kumar, and B. Yegnanarayana. Synthesis of laughter by modifying excitation characteristics. *J. Acoust. Soc. Am.*, 133(5):3072–3082, 2013.

[11] J. Urbain, H. Cakmak, and T. Dutoit. Evaluation of hmm-based laughter synthesis. In *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, pages 7835–7839, Vancouver, Canada, 2013.

[12] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. *Speech Commun.*, 27(3–4):187–207, 1999.

[13] H. Kawahara, Jo Estill, and O. Fujimura. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. In *2nd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, 2001.

[14] T. Drugman and T. Dutoit. The deterministic plus stochastic model of the residual signal and its applications. *IEEE Trans. on Audio, Speech, and Lang. Proc.*, 20(3):968–981, Mar. 2012.

[15] R. Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[16] Bajibabu Bollepalli, Jéromê Urbain, Tuomo Raitio, Joakim Gustafson, and Hüseyin Cakmak. A comparative evaluation of vocoding techniques for HMM-based laughter synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014. accepted.

[17] T. Yoshimura, K. Tokuda, T. Masuko, and T. Kitamura. Mixed-excitation for HMM-based speech synthesis. *Eurospeech*, pages 2259–2262, 2001.

[18] Y. Stylianou. Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification. *PhD thesis, Ecole Nationale Superieure des Telecommunications*, 1996.

[19] Y. Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Trans. Speech and Audio Processing*, 9(1):21–29, 2001.

[20] J. Lorenzo-Trueba, R. Barra-Chicote, J. Yamagishi, O. Watts, and J.M. Montero. Evaluation of a transplantation algorithm for expressive speech synthesis. In *proceedings of Workshop en Tecnologas Accesibles, IV Congreso Espaol de Informtica CEDI2013*, 2013.

[21] S. Lebai Lutfi, F. Fernndez-Martnez, J. Lorenzo-Trueba, R. Barra-Chicote, and J. M. Montero. I feel you: The design and evaluation of a domotic affect-sensitive spoken conversational agent. *Sensors*, 13(8):10519–10538, 2013.

# Appendix: Submitted conference and journal papers

## 6.1 DNN-based voice source modelling: Comparison to the conventional GlottHMM

# VOICE SOURCE MODELLING USING DEEP NEURAL NETWORKS FOR STATISTICAL PARAMETRIC SPEECH SYNTHESIS

*Tuomo Raitio*[*], *Heng Lu*[†], *John Kane*[‡], *Antti Suni*[§], *Martti Vainio*[§], *Simon King*[†], *Paavo Alku*[*]

[*] Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
[†] Centre for Speech Technology Research, University of Edinburgh, UK
[‡] Phonetics and Speech Laboratory, Trinity College Dublin, Ireland
[§] Institute of Behavioural Sciences, University of Helsinki, Finland

## ABSTRACT

This paper presents a voice source modelling method employing a deep neural network (DNN) to map from acoustic features to the time-domain glottal flow waveform. First, acoustic features and the glottal flow signal are estimated from each frame of the speech database. Pitch-synchronous glottal flow time-domain waveforms are extracted, interpolated to a constant duration, and stored in a codebook. Then, a DNN is trained to map from acoustic features to these duration-normalised glottal waveforms. At synthesis time, acoustic features are generated from a statistical parametric model, and from these, the trained DNN predicts the glottal flow waveform. Illustrations are provided to demonstrate that the proposed method successfully synthesizes the glottal flow waveform and enables easy modification of the waveform by adjusting the input values to the DNN. In a subjective listening test, the proposed method was rated as equal to a high-quality method employing a stored glottal flow waveform.

***Index Terms***— Deep neural network, DNN, voice source modelling, glottal flow, statistical parametric speech synthesis

## 1. INTRODUCTION

Statistical parametric speech synthesis, often known as hidden Markov model (HMM) speech synthesis [1, 2], is a flexible framework for synthesising speech. It has several attractive properties, such as the ability to vary speaking style and speaker characteristics, small memory footprint, and robustness. However, HMM-based speech synthesis suffers from lower speech quality than the unit selection approach [3] and this is thought to stem mainly from three factors: a) over-simplified vocoder techniques, b) acoustic modelling inaccuracy, and c) over-smoothing of the generated speech parameters [2]. This paper addresses the problem of over-simplified vocoders by introducing a new voice source modelling method using a deep neural network (DNN).

One of the key factors in improving the quality of statistical speech synthesis has been the development of better excitation modelling techniques. The earliest vocoders used a train of impulses [4] located at the glottal closure instants to model voiced excitation. The quality of this impulse-train-excited speech is poor with a buzzy sound quality due to the zero-phase character of the excitation. Several improvements, such as mixed excitation [5] and two-band excitation [6], have been introduced to alleviate this effect by mixing periodic excitation with aperiodic noise. Mixed excitation is used in, e.g., STRAIGHT [7, 8], which is one of the most widely used vocoders in HMM-based speech synthesis. Voiced excitation has also been improved by using a closed-loop training approach [9,10] or parametric models of the glottal flow [11,12].

The natural excitation of voiced speech, the glottal flow, is difficult to represent as a compressed parametric vector suitable for statistical parametric modelling. Therefore, sampling approaches that utilize the excitation waveform *per se* have been proposed that capture the detailed characteristics of the signal. This idea is not new (see e.g. [13–15]), but the development of statistical parametric synthesis has given rise to several novel excitation methods based on natural speech samples. For example, in [16, 17], a glottal flow pulse estimated from natural speech (using glottal inverse filtering) is manipulated in order to construct a more natural excitation signal. In [18–21], principal component analysis (PCA) is applied to pitch-synchronous residual/glottal flow signals to represent the excitation waveform. In [22, 23], a pitch-synchronous residual/glottal flow codebook is constructed, from which appropriate pulses are selected for synthesis.

Yet, sampling in the voice source domain exhibits some challenges similar to those in the unit selection approach [21, 23], i.e., finding the best sequence of units that well matches the given target specification and concatenate imperceptibly together. Purely sampling-based approaches are, like unit selection, inherently inflexible and limited by the available samples in the database: this limits the ability of the system to

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287678 (Simple[4]All), the Academy of Finland, and EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

change voice quality in a continuous manner, for example.

To overcome the above problems of using stored samples without attempting to construct a fully parametric model of glottal pulses (which has proved very challenging), we introduce a novel voice source modelling technique that can be considered as a compromise between waveform sampling and parametric modelling. The method is based on predicting the pitch-synchronous glottal flow directly in the time-domain by using a DNN. The DNN is used to map the modelled speech parameters to the actual excitation waveform, which can then be used directly for synthesis in combination with predicted vocal tract features. The proposed method has the flexibility of a parametric model because it is able to generate variation in the voice source waveform in response to changes in the speech features. It also exhibits some of the advantages of stored sample-based methods in that the predicted waveforms contain more detail than parametric models.

The rest of the paper is organized as follows. First, DNNs in the context of this work are briefly introduced in Section 2, after which the proposed DNN-based voice source modelling technique is described in Section 3. Experiments using the new method are described in Section 4, concentrating on DNN architecture and training, and on the use of the proposed method in copy-synthesis, voice source modification, and HMM-based synthesis. Finally, the new method and its potential applications are discussed in Section 5.

## 2. DEEP NEURAL NETWORKS

A DNN [24] is a feed-forward, artificial neural network that has at least two layers of hidden units between input and output layers. In this work, a DNN is used to build a mapping from extracted acoustic speech features to corresponding glottal flow pulses. This is a regression problem, where we are predicting continuously-valued outputs, so we chose a linear activation function for the output (regression) layer with sigmoid activation function units for the hidden layers. The latter is defined as

$$v_i = f(\sum_j W_{ij} x_j + b_i) \tag{1}$$

where $f(x) = 1/(1+\exp(-x))$ is the sigmoid logistic function, $W_{ij}$ and $b_i$ are weights and biases, and $x_j$ and $v_i$ are the input and output of the DNN, respectively. For the linear layer, the activation function is simply

$$v_i = \sum_j W_{ij} x_j + b_i \tag{2}$$

Restricted Boltzmann machine (RBM) pre-training can be used for preventing over-fitting to the data. RBM pre-training aims at unsupervised learning of the distributions of the input features. Since the input acoustic features are real valued in this work, a Gaussian-Bernoulli RBM [24] is employed for the visible (input) layer.

After optional pre-training, the DNN is trained ("fine-tuned") by back-propagating derivatives of a cost function that measures the discrepancy between the target outputs and the actual outputs. In this work, mean squared error (MSE) is used as the cost function. The error function is

$$error = \sum_j (\boldsymbol{v}_j - \hat{\boldsymbol{v}}_j)^2 \tag{3}$$

where $\hat{\boldsymbol{v}}_j$ is the regression target for DNN training.

## 3. DNN-BASED VOICE SOURCE MODELLING

Recently, for both automatic speech recognition [24] and speech synthesis [25], DNNs have shown improvements over conventional HMMs with Gaussian mixture models (HMM-GMMs). In our work, a DNN is used in conjunction with an HMM-GMM and the proposed approach is illustrated in Figure 1. First, frame-wise acoustic features are extracted from a database. In the feature extraction, iterative adaptive inverse filtering (IAIF) [26] is used to decompose the speech signal into a vocal tract filter and a voice source signal. The extracted speech parameters include the vocal tract linear prediction (LP) filter that is converted to a line spectral frequency (LSF) representation, and parameters describing the properties of the voice source, i.e., fundamental frequency (F0), frame energy, harmonic-to-noise ratio (HNR) of five frequency bands, and voice source LP spectrum converted to LSF. The extracted features, depicted in Table 1, are then used to train an HSMM-based synthesizer, as in [17].

The IAIF method produces an estimate of the voice source signal from which individual glottal flow pulses are extracted. To do this, glottal closure instants (GCIs) are detected from the differentiated glottal flow signal using a simple peak picking algorithm. This enables the extraction of two-pitch-period, GCI-centred glottal flow pulses, delimited by two other GCIs. The pulse segments are interpolated to a constant duration of 25 ms (400 samples at 16 kHz sampling rate), windowed with the Hanning window, normalized in energy, and stored in a codebook. The fixed duration of the pulses is chosen as a compromise between minimizing the amount of data stored and limiting loss of spectral information.

Given the set of glottal pulses and corresponding vectors of 47 acoustic parameters (Table 1), a mapping is established by training the DNN. RBM pre-training is used to alleviate over-fitting, after which back-propagation is applied. For synthesis, both vocal tract and voice source parameters are generated from context-dependent HMMs, as in [17]. Instead of using the source speech parameters to select a sequence of stored pulse waveforms drawn from the codebook, we use the complete set of 47 acoustic parameters as input to the DNN, which outputs glottal flow derivative waveforms. The generated glottal flow pulses are interpolated to a duration corresponding to the required F0, scaled in energy, mixed with noise according to the HNR measure, and overlap-added to

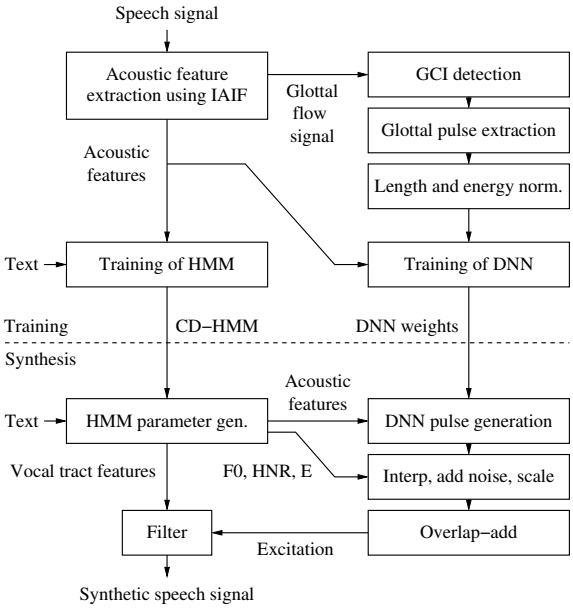



**Fig. 1**. Illustration of the proposed HMM-based speech synthesis using DNN-based voice source modelling.

generate the excitation for synthesis. Alternatively, the DNN pulses can be used as a target for selecting the closest matching stored glottal flow waveforms from the codebook (similar to [23]). The vocal tract filter already generated by the HMM is then used to filter the excitation signal, producing speech.

## 4. EXPERIMENTS

### 4.1. Experimental setup

Two Finnish speech databases, male *MV* and female *Heini*, recorded for the purpose of speech synthesis, were used in the experiments. The male voice comprises 600 sentences (approx. 1 h of speech) and the female database comprises 500 sentences. Both voices were sampled at 16 kHz.

The GlottHMM vocoder [17, 23] was used for extracting the acoustic features and the glottal flow signal using IAIF. Glottal flow pulse codebooks were constructed for both databases in order to train the DNN-based voice source model. The codebooks contained all 203,172 or 203,768 pulses for the male or female speakers, respectively. Additionally, smaller codebooks were constructed for both speakers from 20 sentences of speech material, in order to implement the alternative method in which the DNN output is used to select a natural pulse from the codebook; these codebooks consisted only of 7,495 and 8,131 pulses in order to minimize computational cost at synthesis time. The standard HTS 2.1 method [27] was used for training the HMM-based system.

**Table 1**. Acoustic features used for training the HMM-based synthesis and the DNN-based voice source model.

| Feature | Number of parameters |
|---|---|
| Energy | 1 |
| Fundamental frequency | 1 |
| Harmonic-to-noise ratio | 5 |
| Voice source spectrum | 10 |
| Vocal tract spectrum | 30 |

### 4.2. DNN training

The DNN architecture as described in Section 2 is used. The input is the 47-dimensional vector composed of the extracted acoustic speech features listed in Table 1 and the target output is a 400 sample duration normalised glottal flow pulse.

In order to determine the optimal number of layers and hidden units for DNN, six different systems (A–F) were trained by varying the number of hidden layers (from 1 to 3) and the number of units per layer (from 800 to 1200). Unsupervised RBM pre-training was tried for one configuration. 200,000 training examples were used for training with 3,000 examples for cross-validation. The training and development errors for each system are presented in Table 2. The results show that system F with 2 hidden layers and 1000 units per hidden layer gave best results, with RBM pre-training slightly improving performance (compare system F to system B).

### 4.3. Voice source modelling and modification

Copy-synthesis for unseen speech data (i.e., not in the training or validation sets) using the proposed method is illustrated in Figure 2, which shows the original (differentiated) excitation estimated by IAIF from natural speech and the synthetic DNN-based excitation generated from the extracted parameters. The DNN-based excitation has been mixed with noise according to the HNR measure. In informal listening, the proposed voice source modelling method produces natural sounding copy-synthesis, either by directly using the DNN generated pulses or by using them as a target to select pulses from the smaller codebook.

A potential advantage of predicting pulses with the DNN is the ability to continuously adjust the glottal flow waveform in response to the input acoustic features. Figure 3 demonstrates this ability: frame energy, F0, and HNR are varied individually while other parameters are left unchanged, and pulses are generated from the trained DNN. The pulse waveform displays a continuous and consistent change in response

| | Hidden layers | Units per layer | Pre-training | Train error | Dev set error |
|---|---|---|---|---|---|
| A | 1 | 1000 | No | 0.4109 | 0.4990 |
| B | 2 | 1000 | No | 0.3980 | 0.4875 |
| C | 3 | 1000 | No | 0.3999 | 0.4891 |
| D | 2 | 800 | No | 0.4037 | 0.4925 |
| E | 2 | 1200 | No | 0.4134 | 0.5015 |
| F | 2 | 1000 | Yes | 0.3935 | 0.4846 |

**Table 2**. Training and development mean squared error (MSE) for various DNN configurations.

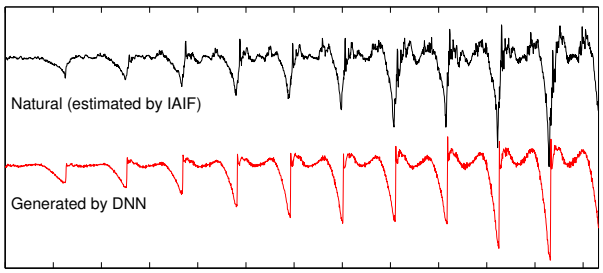



**Fig. 2**. Demonstration of the DNN-based excitation generation by copy-synthesis of a Finnish male speech segment [vie]. The upper signal (black) represents the estimated differentiated glottal flow obtained by IAIF. The lower signal (red) represents the excitation generated by DNN according to the acoustic features, with noise mixed in according to the HNR.

to the varied speech parameter. For example, with low input energy, the glottal pulse shows a less prominent peak at the GCI whilst with high input energy the pulse has a very sharp discontinuity at the GCI. Similarly natural behaviour is observed also with F0 and HNR. This opens up possibilities for more flexible voice source modification.

### 4.4. Subjective evaluation of HMM synthesis

In order to demonstrate the capability and assess the quality of the proposed method, an online subjective evaluation was carried out. Three different methods were chosen for comparison: 1) Conventional GlottHMM synthesis [17] using a single natural glottal flow pulse, of which spectrum is matched according to the voice source LSF, 2) DNN-based voice source modelling, and 3) DNN-based voice source model used as a target cost for selecting natural glottal flow pulses from a small codebook. The latest single pulse GlottHMM was selected for comparison since it has been found to be a reliable method for producing high quality synthetic speech, and better than STRAIGHT with male speech [17].

A comparison category rating (CCR) test was used, in which pairs of stimuli are presented to participants, whose task is to indicate the difference between the two samples on a comparison mean opinion score (CMOS) scale, which is a discrete seven-point scale ranging from much worse ($-3$) to much better (3). All three combinations of the systems (1–2, 1–3, 2–3) were evaluated. 50 utterances were synthesized from held-out data from both speakers and for each of the three systems (300 stimuli in total). In order to reduce the workload on participants, 10 sentences from both speakers were randomly selected for each participant and presented to them in each of the three system combinations. Thus each participant rated a total of 60 stimuli pairs. Also the ordering of the pairs of stimuli was randomized. 26 people (15 Finnish and 11 non-Finnish) participated in the evaluation. The CCR test responses are summarized by calculating the mean score for each evaluated method, which yields the order

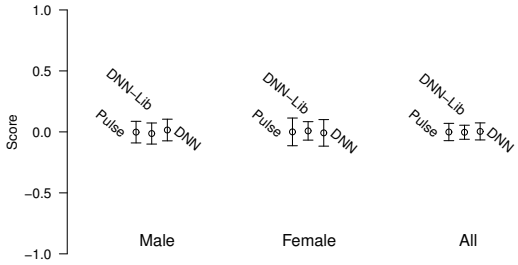



**Fig. 4**. Results of the subjective evaluation showing that the DNN-based methods are rated as equal to the baseline system.

of preference and distances between all the methods (i.e., the amount of preference relative to each other). The results of the CCR test, plotted in Figure 4, are encouraging in showing that both new DNN-based methods are rated as equal to the high-quality baseline synthesis system.

## 5. CONCLUSIONS

This paper presented a voice source modelling method based on predicting the time domain glottal flow waveform using a DNN. In the experiments presented in this paper, the proposed DNN-based method is shown to successfully generate acoustic feature-dependent glottal flow waveforms and to produce high-quality HMM-based speech synthesis, comparable to the state-of-the-art methods. In addition to accurate voice source modelling, the method offers possibilities for automatic or manual voice source modification.

## REFERENCES

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. Eurospeech*, Sep. 1999, pp. 2374–2350.

[2] Heiga Zen, Keiichi Tokuda, and Alan W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[3] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, 1996, pp. 373–376.

[4] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation in HMM-based speech synthesis system," in *Proc. Eurospeech*, 1997, pp. 2523–2526.

[6] S. J. Kim and M. Hahn, "Two-band excitation for HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, 2007.

[7] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3–4, pp. 187–207, 1999.

### 6.2 DNN-based voice source modelling: Evaluation with varying vocal effort

# Deep neural network based trainable voice source model for synthesis of speech with varying vocal effort

*Tuomo Raitio*[1], *Antti Suni*[2], *Lauri Juvela*[1], *Martti Vainio*[2], *Paavo Alku*[1]

[1]Department of Signal Processing and Acoustics, Aalto University, Finland
[2]Institute of Behavioural Sciences, University of Helsinki, Finland
firstname.lastname@aalto.fi, firstname.lastname@helsinki.fi

## Abstract

This paper studies a deep neural network (DNN) based voice source modelling method in the synthesis of speech with varying vocal effort. The new trainable voice source model learns a mapping between the acoustic features and the time-domain pitch-synchronous glottal flow waveform using a DNN. The voice source model is trained with various speech material from breathy, normal, and Lombard speech. In synthesis, a normal voice is first adapted to a desired style, and using the flexible DNN-based voice source model, a style-specific excitation waveform is automatically generated based on the adapted acoustic features. The proposed voice source model is compared to a robust and high-quality excitation modelling method based on manually selected mean glottal flow pulses for each vocal effort level and using a spectral matching filter to correctly match the voice source spectrum to a desired style. Subjective evaluations show that the proposed DNN-based method is rated comparable to the baseline method, but avoids the manual selection of the pulses and is computationally faster than a system using a spectral matching filter.
**Index Terms**: Speech synthesis, deep neural network, DNN, voice source modelling, vocal effort, glottal flow

## 1. Introduction

Statistical parametric speech synthesis, also known as hidden Markov model (HMM) based speech synthesis [1, 2], is a popular framework for synthesising speech and a good alternative for the unit selection approach [3]. It has several benefits such as the ability to vary speaking style and speaker characteristics [4–8], small memory footprint [9, 10], and robustness [11]. However, statistical speech synthesis suffers from lower segmental speech quality compared to the unit selection systems that concatenate natural speech waveforms [3]. This degradation is thought to stem mainly from three factors: a) oversimplified vocoder techniques that are incapable of representing natural speech waveforms in detail b) acoustic modelling inaccuracy, and c) over-smoothing of the generated speech parameters [2]. This paper addresses the first factor by introducing a flexible voice source model that uses a deep neural network (DNN), with the aim of better modelling variations in the voice source signal and interaction between the source and the filter.

The modelling of the excitation signal in HMM-based speech synthesis has greatly improved since the first vocoders that used a simple impulse train excitation [12]. The quality of such simple excitation is poor due to the unnatural zero-phase character of the excitation. Mixed excitation [13] and two-band excitation [14] has greatly improved the quality by mixing periodic excitation with aperiodic noise. This mixed excitation scheme is also used in the most prevalent vocoder in speech synthesis, STRAIGHT [15, 16]. Also closed-loop training [17, 18] and parametric models of the glottal flow [19, 20] have been proposed for improving the speech quality.

Since the context-dependent characteristics of the glottal flow waveform are difficult to represent using a simple parametric voice source signal, several approaches have utilised the excitation waveform *per se* in order to preserve the natural characteristics in the waveform. The idea is not new (see e.g. [21–23]), but the development of statistical speech synthesis and vocoders have given new applications for the approach. Recently, natural glottal flow pulses or residual waveforms have been used in several vocoding approaches [24–31].

Reproducing different speaking styles has long been the strength of statistical speech synthesis. Through adaptation and similar techniques, a continuous degree of varying style can be reproduced [4–8, 32, 33]. However, only few studies have explicitly investigated the modelling of the changes in the excitation waveform in response to changes in speaking style. In consequence, while mostly changing the pitch and overall spectrum, the changes in the voice characteristics are rather limited compared to natural speech. In contrast, the experiments in [33, 34] have shown that by using an appropriate glottal flow pulse for synthesising a specific style, improvements in the perceived impression of the style are achieved. However, the current approaches need human intervention, such as manually extracting and selecting the style-specific excitation waveforms.

The aim of this work is to present and extend the work on the DNN-based voice source modelling method, preliminary presented in [35], and apply it to the reproduction of various vocal effort levels similar to the study in [33]. The new DNN-based voice source modelling method is based on learning a mapping between the acoustic features and the time-domain glottal flow waveform using DNN. Thus, in synthesis, the excitation waveform can be directly generated from the acoustic features. Subjective evaluations are performed to find out if the new simpler and automatic DNN-based method can reproduce the same quality and impression of vocal effort as the previously published method without manual intervention.

## 2. DNN-based voice source modelling

The proposed DNN-based voice source modelling method and its use in synthesis of various speaking styles is illustrated in Figure 1. In the training part, acoustic features are first extracted from a speech database at 5-ms intervals. As the aim is to reproduce different speaking styles, the speech database should contain both normal and style-specific speech, labelled accordingly. The feature extraction uses iterative adaptive in-

verse filtering (IAIF) [36] in order to decompose speech signals into the vocal tract filter and the voice source signal. This enables the further parametrisation of the voice source characteristics and the segmentation of the glottal flow waveforms. Speech features described in Table 1 are extracted, i.e., the fundamental frequency (F0), frame energy, harmonic-to-noise ratio (HNR) of five frequency bands, voice source linear prediction (LP) spectrum converted to line spectral frequencies (LSF), and vocal tract LP spectrum converted to LSF. The acoustic features of normal style are used for training an HMM-based voice, after which it can be adapted to different speaking styles.

The output voice source signal by the IAIF algorithm is used for extracting pitch-synchronous glottal flow pulse segments. First, glottal closure instants (GCIs) are detected from the differentiated glottal flow signal using peak picking at fundamental period intervals, and two-pitch-period, GCI-centred glottal flow waveform segments are extracted. The pulse segments are interpolated to a constant duration of 25 ms (400 samples at 16 kHz sampling rate), windowed with the Hanning window, and normalised in energy. The pulses are stored in a codebook and linked with the corresponding acoustics features of the frame. The duration of the pulses is selected as a compromise between minimising the amount of data stored and limiting the loss of spectral information in the pulses. A mapping between the acoustic features and the glottal flow waveform segments is established by training a DNN. Random initialisation of the DNN weights is used, after which back-propagation is applied. In order to train a flexible voice source model, speech parameters from all speaking styles were used for the DNN training.

The normal voice is adapted as in [8] to a desired style using the style-specific data and an interpolation/extrapolation coefficient, which defines the amount of adaptation from the normal voice to the desired style. After the adaptation of the voice, style-specific acoustic features are generated from context-dependent HMMs (CD-HMM) according to text input as in [25]. The acoustic features are used as input to DNN, which outputs the context and style-specific glottal flow waveforms. The generated glottal flow waveforms are interpolated to a desired length according to F0, scaled in energy, and mixed with noise according to the HNR measure as in [31]. The individual two-pitch-period waveforms are overlap-added in order to create a continuous excitation, which is filtered with the vocal tract filter generated from HMMs to create speech.

# 3. Experiments

## 3.1. Speech material

Two speech corpora, a male and a female speaker [33], were used in the experiments. For both speakers, three different vocal effort levels were utilised: breathy, normal, and Lombard. The normal style consists of 1450 sentences, comprising approximately two hours of speech for both speakers. Lombard speech was elicited by playing babble noise with 80 dB SPL

Table 1: *Acoustic features used for training the HMM-based voice and the DNN-based voice source model.*

| Feature | Number of parameters |
|---|---|
| Energy | 1 |
| Fundamental frequency | 1 |
| Harmonic-to-noise ratio | 5 |
| Voice source spectrum | 10 |
| Vocal tract spectrum | 30 |

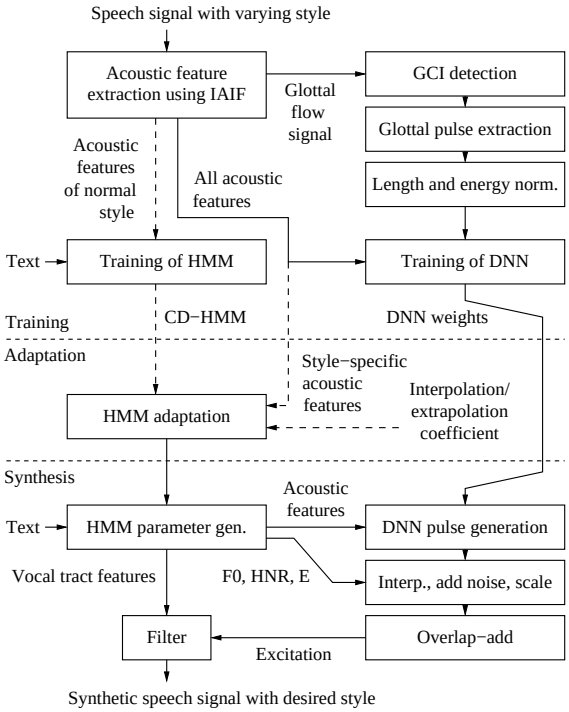



Figure 1: *Illustration of the proposed DNN-based voice source modelling method for synthesis of varying speaking styles.*

to the speaker's ears through headphones while recording, and feeding back the speaker's own voice through headphones, corresponding to a level of speaking in a normal room without headphones. The Lombard style consists of 300 sentences. The breathy speaking style was elicited by increasing the level of the speaker's feedback through headphones as well as instructing the subjects to speak softly without whispering. 200 sentences were read in the breathy style. The recording and processing of the speech data are described in more detail in [33].

## 3.2. Training of deep neural networks

A DNN [37] is a feed-forward, artificial neural network that has at least two layers of hidden units between input and output layers. Recently, DNNs have been successfully used for both automatic speech recognition [37] and speech synthesis [38], and DNNs have shown improvements over conventional HMM-based systems. In this work, a DNN is used in conjunction with an HMM-based approach for mapping between the acoustic features and the time-domain glottal flow waveform. The input for the DNN is the 47-dimensional acoustic feature vector, consisting of the features described in Table 1, and the output is the 400 sample duration normalised glottal flow waveform. For the hidden and output layers, sigmoid and linear activation functions are used, respectively. The DNN is trained by back-propagating derivatives of the mean squared error (MSE) cost function that measures the discrepancy between the target and actual outputs.

Previously in our research on DNN-based voice source modelling [35], a rather large network of 1000 neurons per layer with three two hidden layer was proposed for learning the mapping between the acoustic features and the glottal flow waveform. In our recent studies, smaller DNN architectures have been shown to learn the mapping more efficiently, while also

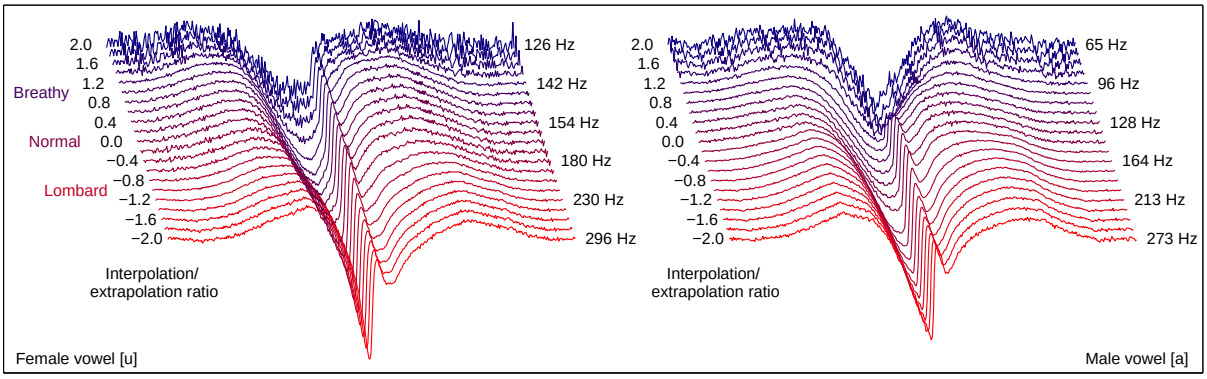



Figure 2: *Demonstration of the DNN-based excitation modelling by interpolating and extrapolating different HMM-based speaking styles from original breathy (*$1.0$*), normal (*$0.0$*), and Lombard speech (*$-1.0$*), and generating the DNN-based pulses corresponding to the generated speech parameters of various degrees of the styles. The resulting pulses (without interpolation in time, scaling in magnitude or adding noise) are shown for female vowel [u] and male vowel [a].*

taking less time to train. In this work, a 2-hidden-layer DNN is used with 100 and 200 neurons in the first and the second hidden layers, respectively. In a reasonable time, it achieved much lower errors than the DNN architecture proposed in [35].

In addition, restricted Boltzmann machine (RBM) pre-training, used in [35], turned out to achieve fast initial reduction in training error, but the error curves saturated also very rapidly and did not achieve even nearly as low errors as random weight initialisation. This seems to indicate that the RBM pre-training helped in learning the main characteristics of the glottal flow waveform, but reduced the flexibility of the model to learn the multitude of variations in the glottal waveform shape. Thus, in this work, random initialisation of the DNN weights is used.

Since the 400-sample-length glottal flow waveform is rather high-dimensional, an approach using principal component analysis (PCA) was also experimented with. The glottal flow waveforms in the training database were decomposed into 40 principal components (PCs) and the corresponding weights, and a mapping between the 47 acoustic features and the 40 PCs were then established using a DNN. The results were similar to the sample-based approach, but glottal flow waveform was inconsistent when using unseen or noisy input data. Thus, the time-domain glottal flow waveform is used in this work.

Due to occasional small errors in the GCI estimation, and due to the averaging effect of the DNN training, the GCI peaks of the generated pulses are slightly smoother than those in the waveform inverse filtered from natural speech. In order to compensate this constant difference in spectral domain, a fixed pre-emphasis is applied at synthesis stage. The amount of pre-emphasis is estimated by comparing the spectra of the voice source signals over all styles synthesised with the DNN-based method and conventional GlottHMM synthesis using natural glottal flow pulse and a source spectral matching scheme [25]. Best match between the two spectra was achieved with first-order differentiator with $\alpha = 0.387$.

### 3.3. Handling data sparsity

Robustness to data sparsity is a crucial property of a generative model, and especially in speech synthesis, data sparsity is a common problem. It is often not possible to include all possible input cases in the training material, and thus it is important that a model can interpolate or extrapolate an appropriate output from input parameters that are not included in the training set.

In order to demonstrate the ability of the proposed DNN-based voice source model to create natural glottal flow waveform despite data sparsity, two training sets were constructed with the other one missing a part of the input parameter values. A data set of 280,651 input vectors and output pulse waveforms were used to train a baseline DNN using the male speech data. Since energy of the speech frame is highly dependent on the speaking style, and the glottal pulse waveform shows considerable changes in relation to changes in energy (see [35]), it was chosen as a feature to be altered in this experiment. The energy in the original training set ranged from $-23.3$ dB to 41.0 dB. A modified training set was constructed by removing all data points with energy values from 0 dB to 15 dB. After discarding the specific data, the number of training samples in the modified set was 227,777, removing around 19% of the total samples and corresponding exemplars of glottal flow waveforms. Both DNNs were trained similarly and the errors of the generated glottal flow waveforms were measured using a test set with i) all data, ii) in-domain data, iii) out-of-domain data. The mean, maximum, and minimum relative change in errors ($E$) are shown in Table 2. The results show that the overall error is only slightly increased when moving from the in-domain data (0.73%) to the out-of-domain data (2.07%), indicating that the model can rather successfully interpolate/extrapolate the output.

### 3.4. Voice building

The HMM training and adaptation procedures were identical to the experiments done in [33]. The training of the normal voices followed the standard HTS method [39]. Speech features described in Table 1 were extracted using the GlottHMM vocoder [25] and delta, and delta-delta features were added. Semi hidden Markov models were used as acoustic models, and features were trained in individual streams except the vocal tract LSFs and energy were trained together.

Table 2: *Mean, maximum, and minimum change in the error* $E$ *over the generated glottal flow waveforms when using a training data with induced data sparsity in comparison to using all data.*

| Test data | $\Delta$mean($E$) | $\Delta$max($E$) | $\Delta$min($E$) |
|---|---|---|---|
| All data | 1.17 % | 1.37 % | $-7.86$ % |
| In-domain data | 0.73 % | 1.37 % | $-7.86$ % |
| Out-of-domain data | 2.07 % | $-1.89$ % | 36.18 % |

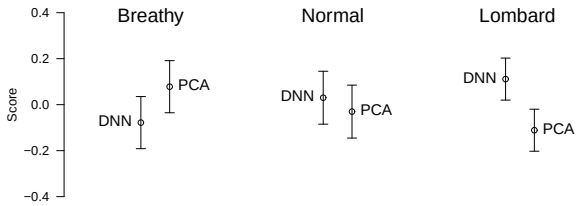



Figure 3: *Results of the quality test.*

In order to create the low and high vocal effort voices (breathy and Lombard), the normal voice models were adapted with constrained structural maximum a posteriori linear regression combined with maximum a posteriori (CSMAPLR + MAP) adaptation technique [8]. The speaker-dependent voice source model DNNs were trained using all speech material including breathy, normal, and Lombard speech.

For demonstrating the interpolation and extrapolation characteristics of the DNN-based voice source modelling, both voices were adapted to various degrees of vocal effort from very breathy to very Lombard, and corresponding speech parameters were generated. Normal training voices being at point 0.0 and adaptation samples at points 1.0 (breathy) and at $-1.0$ (Lombard), adapted voices were created between 2.0 (very breathy) and $-2.0$ (very Lombard) with a step size of 0.2. The parameters of each voice were then fed to the DNNs to generate style-specific glottal flow waveforms. Generated waveforms for female vowel [u] and male vowel [a] are shown in Figure 2.

### 3.5. Subjective evaluation

In order to evaluate the performance of the proposed method, subjective evaluations were conducted using three vocal effort levels. The final voices used in the subjective evaluation were created at points 1.0 (breathy), 0.0 (normal), and $-1.0$ (Lombard) for both speakers. A high-quality mean glottal flow pulse excitation scheme was selected for a reference baseline system, which has been successfully used in synthesising speech with varying vocal effort [33]. The baseline system uses a style-specific mean glottal flow pulse for each of the three styles [33] (corresponding to the PCA-based excitation in [29]), and a spectral matching scheme [25,33], where a pole-zero filter is used to filter the excitation signal in order to apply the desired spectral properties defined by the generated voice source spectrum.

Two types of tests were conducted to compare the proposed and the baseline systems. First, a comparison category rating (CCR) test was conducted to evaluate the speech quality. In a CCR test, listener hears two different samples and rates the quality difference between them on the 7-point comparison mean opinion score scale ranging from much worse ($-3$) to much better (3). A total of 14 native Finnish listeners evaluated 60 sample pairs each, and the preference of the methods was evaluated by averaging the listener scores for each method.

Secondly, a similarity test was conducted in order to assess the speaker and style similarity between the two methods. In the similarity test, listener is presented with two speech samples synthesised by the two methods, and a natural reference sample corresponding to the speaker and style of the synthetic samples. The task of the listener is to choose which of the two samples is more similar to the reference in terms of speaker and style, or no preference between the samples. A total of 14 native Finnish listeners evaluated 60 sample pairs each.

The mean scores of the quality test are shown for each vocal effort level with 95% confidence intervals in Figure 3. Only

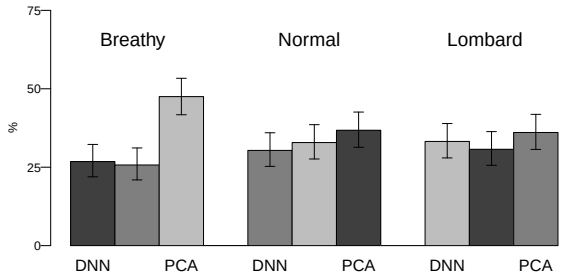



Figure 4: *Results of the similarity test.*

with Lombard speech, the difference between the two methods is statistically significant with the proposed method being rated higher in quality. Figure 3 presents the results of the similarity test, showing the proportion of answers (with 95% confidence intervals) for each method and for each vocal effort level. Only in the case of breathy speech, the results are statistically significant, where the baseline method is rated more similar.

## 4. Discussion and conclusions

The experiments show that the proposed DNN-based voice source modelling method is capable of successfully reproducing different degrees of vocal effort, and that it improves the synthesis quality with Lombard speech in comparison to the baseline method. Although the proposed method was able to generate breathier waveforms than the baseline system, and although the resulting breathy voice was perceptually softer based on informal listener reports, the similarity of the breathy voice was slightly decreased due to the absence of the spectral matching, as is used in [25,33]. In comparison to the baseline method, the proposed method avoids manual intervention needed for the voice style variation, and enables continuous style variation within an utterance, which is required for plausible expressive speech synthesis. Moreover, the generation of pulses from a DNN is computationally less expensive than filtering the excitation signal with a pole-zero spectral matching filter.

The study shows that the approximate shape of the glottal flow waveform can be successfully modelled by the proposed approach in order to generate various speaking styles. However, the proposed method does not seem to greatly improve the segmental quality of speech compared to using a pre-selected glottal flow pulses. This indicates that even though the DNN-based modelling approach is capable of generating the gross shape of the glottal flow pulse, it introduces an averaging effect that removes detailed variations of the pulse needed to achieve quality close to natural speech.

The existence of interaction between the source and filter is well known (see e.g. [40]), but it is hardly utilised in speech technology or in speech synthesis. In this work, the glottal flow waveform is predicted based on features including the vocal tract spectrum, but it seems that modelling the source and filter interaction by the proposed method is not adequate for greatly improving the segmental speech quality. Future directions of the study will be concentrated on the more accurate modelling of the source-filter interaction using the DNN-based approach.

## 5. Acknowledgements

This work has been supported by the EC-FP7 (2007–2013) n° 287678 (Simple$^4$All), and the Academy of Finland (256961).

# 6. References


[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. Eurospeech*, Sep. 1999, pp. 2374–2350.

[2] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[3] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, 1996, pp. 373–376.

[4] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, 2001, pp. 805–808.

[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation in HMM-based speech synthesis system," in *Proc. Eurospeech*, 1997, pp. 2523–2526.

[6] K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," in *Proc. ICSLP*, 2002, pp. 1269–1272.

[7] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 9, pp. 1406–1413, 2007.

[8] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 17, no. 1, pp. 66–83, 2009.

[9] S.-J. Kim, J.-J. Kim, and M.-S. Hahn, "HMM-based Korean speech synthesis system for hand-held devices," *IEEE Trans. Consum. Electron.*, vol. 52, no. 4, pp. 1384–1390, 2006.

[10] A. Gutkin, X. Gonzalvo, S. Breuer, and P. Taylor, "Quantized HMMs for low footprint text-to-speech synthesis," in *Proc. Interspeech*, 2010, pp. 837–840.

[11] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 17, no. 6, pp. 1208–1230, 2009.

[12] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[13] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis," in *Proc. Eurospeech*, 2001, pp. 2259–2262.

[14] S. J. Kim and M. Hahn, "Two-band excitation for HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, 2007.

[15] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3–4, pp. 187–207, 1999.

[16] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *2nd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, 2001.

[17] R. Maia, T. Toda, H. Zen, Y. Nankaku, and K. Tokuda, "An excitation model for HMM-based speech synthesis based on residual modeling," in *6th ISCA Speech Synthesis Workshop*, Aug. 2007.

[18] R. Maia, H. Zen, and M. J. F. Gales, "Statistical parametric speech synthesis with joint estimation of acoustic and excitation model parameters," in *7th ISCA Speech Synthesis Workshop*, Sep. 2010, pp. 88–93.

[19] J. Cabral, S. Renalds, K. Richmond, and J. Yamagishi, "Towards an improved modeling of the glottal source in statistical parametric speech synthesis," in *Sixth ISCA Workshop on Speech Synthesis*, 2007, pp. 113–118.

[20] ——, "Glottal spectral separation for parametric speech synthesis," in *Proc. Interspeech*, 2008, pp. 1829–1832.

[21] J. Holmes, "The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer," *IEEE Trans. Audio and Electroacoustics*, vol. 21, no. 3, pp. 298–305, Jun. 1973.

[22] K. Matsui, S. D. Pearson, K. Hata, and T. Kamai, "Improving naturalness in text-to-speech synthesis using natural glottal source," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, vol. 2, Apr. 1991, pp. 769–772.

[23] G. Fries, "Hybrid time- and frequency-domain speech synthesis with extended glottal source generation," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, vol. 1, 1994, pp. 581–584.

[24] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "HMM-based Finnish text-to-speech system utilizing glottal inverse filtering," in *Proc. Interspeech*, 2008, pp. 1881–1884.

[25] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 19, no. 1, pp. 153–165, 2011.

[26] T. Drugman, G. Wilfart, A. Moinet, and T. Dutoit, "Using a pitch-synchronous residual codebook for hybrid HMM/frame selection speech synthesis," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, Apr. 2009, pp. 3793–3796.

[27] J. Sung, D. Hong, K. Oh, and N. Kim, "Excitation modeling based on waveform interpolation for HMM-based speech synthesis," in *Proc. Interspeech*, 2010, pp. 813–816.

[28] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 20, no. 3, pp. 968–981, Mar. 2012.

[29] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Comparing glottal-flow-excited statistical parametric speech synthesis methods," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, 2013, pp. 7830–7834.

[30] T. Drugman, G. Wilfart, and T. Dutoit, "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis," in *Proc. Interspeech*, 2009, pp. 1779–1782.

[31] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, 2011, pp. 4564–4567.

[32] B. Picart, T. Drugman, and T. Dutoit, "Analysis and hmm-based synthesis of hypo and hyperarticulated speech," *Computer Speech & Language*, vol. 28, no. 2, pp. 687–707, 2014.

[33] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Synthesis and perception of breathy, normal, and lombard speech in the presence of noise," *Computer Speech & Language*, vol. 28, no. 2, pp. 648–664, 2014.

[34] T. Raitio, J. Kane, T. Drugman, and C. Gobl, "Hmm-based synthesis of creaky voice," in *Proc. Interspeech*, 2013, pp. 2316–2320.

[35] T. Raitio, H. Lu, J. Kane, A. Suni, M. Vainio, S. King, and P. Alku, "Voice source modelling using deep neural networks for statistical parametric speech synthesis," in *22nd European Signal Processing Conference (EUSIPCO)*, 2014, submitted.

[36] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Commun.*, vol. 11, no. 2-3, pp. 109–118, 1992.

[37] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Sig. Proc. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[38] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, 2013, pp. 7962–7966.

[39] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Sixth ISCA Workshop on Speech Synthesis*, 2007, pp. 294–299.

[40] I. R. Titze, "Nonlinear source-filter coupling in phonation: Theory," *J. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 2733–2749, 2008.

### 6.3 Vocoder evaluation for HMM-based synthesis of laughter

# A COMPARATIVE EVALUATION OF VOCODING TECHNIQUES FOR HMM-BASED LAUGHTER SYNTHESIS

*Bajibabu Bollepalli*[1]*, Jérôme Urbain*[2]*, Tuomo Raitio*[3]*, Joakim Gustafson*[1]*, Hüseyin Çakmak*[2]

[1]Department of Speech, Music and Hearing, KTH, Stockholm, Sweden
[2]TCTS Lab – University of Mons, Belgium
[3]Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland

## ABSTRACT

This paper presents an experimental comparison of various leading vocoders for the application of HMM-based laughter synthesis. Four vocoders, commonly used in HMM-based speech synthesis, are used in copy-synthesis and HMM-based synthesis of both male and female laughter. Subjective evaluations are conducted to assess the performance of the vocoders. The results show that all vocoders perform relatively well in copy-synthesis. In HMM-based laughter synthesis using original phonetic transcriptions, all synthesized laughter voices were significantly lower in quality than copy-synthesis, indicating a challenging task and room for improvements. Interestingly, two vocoders using rather simple and robust excitation modeling performed the best, indicating that robustness in speech parameter extraction and simple parameter representation in statistical modeling are key factors in successful laughter synthesis.

***Index Terms***— Laughter synthesis, vocoder, mel-cepstrum, STRAIGHT, DSM, GlottHMM, HTS, HMM

## 1. INTRODUCTION

Text-to-speech (TTS) synthesis systems have already reached high degree of intelligibility and naturalness, and they can be readily used in reading aloud a given text. However, applications such as human-machine interaction and speech-to-speech translation require that the synthetic speech includes more expressiveness and conversational characteristics. To bring expressiveness into speech synthesis systems, it is not sufficient to only concentrate on improving the verbal signals alone, since non-verbal signals also play an important role in expressing emotions and moods in human communication [1].

Laughter is one such non-verbal signal playing a key role in our daily conversations. It conveys information about emotions and fulfills important social functions, such as back-channeling. Integrating laughter into a speech synthesis system can bring the synthesis closer to natural human conversation [2]. Hence, the research on analysis, detection, and synthesis of laughter signals has seen a significant increase in the last decade. In this paper, we focus on acoustic laughter synthesis, and explore the role of vocoder techniques in statistical parametric laughter synthesis.

The paper is organized as follows. Section 2 gives the background of work done in laughter processing and laughter synthesis in particular. Section 3 describes the different vocoders compared in this work. Section 4 focuses on the perceptual evaluation experiment carried out to compare the vocoders in their capabilities to produce natural laughter. The results of these experiments are discussed in Section 5. Finally, Section 6 presents the conclusions of this work.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements n° 270780 (ILHAIRE) and n° 287678 (Simple[4]All). H. Çakmak receives a Ph.D. grant from the Fonds de la Recherche pour l'Industrie et l'Agriculture (F.R.I.A.), Belgium.

## 2. BACKGROUND

In the last decade, a considerable amount of research has been done on the analysis and detection of laughter (see e.g. [3]), whereas only a few studies have been conducted for synthesis. The characteristics of laughter and speech are slightly different. Formant frequencies in laughter have been reported to correspond to those of central vowels in speech, but acoustic features like fundamental frequency ($F_0$) has been shown to have higher variability in laughter than in speech [4]. Importantly, the proportion of fricatives in laughter has been reported to be about 40–50% [5], which is much higher than in speech. Despite the differences, the same speech processing algorithms have been applied for laughter analysis as for speech analysis.

As the acoustic behavior of laughter is different from speech, it is relatively easy to discriminate laughter from speech. Classification usually depends upon various machine learning methods, such as Gaussian mixture models (GMMs), support vector machines (SVMs), multi-layer perceptrons (MLPs), or hidden Markov models (HMMs), which all use traditional acoustic features (MFCCs, PLP, $F_0$, energy, etc.). Equal error rates (EER) vary between 2% and 15% depending on the data and classification method used [6, 7, 8].

On the other hand, acoustic laughter synthesis is an almost unexplored domain. In [9], Sundaram and Narayanan modeled the temporal behaviour of laughter using the principle of a damped simple harmonic motion of a mass-spring model. Laughs synthesized with this method were perceived as non-natural by naive listeners (average naturalness score of 1.71 on a 5-point Likert scale [10]. ranging from 1 (very poor) to 5 (excellent)). Lasarcyk and Trouvain [11] compared two laughter synthesis approaches: articulatory synthesis resulting from a 3D modeling of the vocal organs and diphone concatenation (obtained from a speech database). The 3D modeling led to the best results, but laughs could still not compete with natural human laughs in terms of naturalness. Recently two other methods have been proposed. Sathya et al. [12] synthesized voiced laughter bouts by controlling several excitation parameters of laughter vowels: pitch period, strength of excitation, amount of frication, number of laughter syllables, intensity ratio between the first and the last syllables, duration of fricative and vowel in each syllable. The synthesized laughs reached relatively high scores in perceived quality and acceptability, with values around 3 on a scale ranging from 1 to 5. However, it must be noted that no human laugh was included in the evaluation, which might have had a positive influ-

ence on the scores obtained by the synthesized laughs (as there is no "perfect" reference to compare with in the evaluation). Also, the method only enables the synthesis of voiced bouts (there is no control over unvoiced laughter parts). Finally, Urbain et al. [13] used HMMs to synthesize laughs from phonetic transcriptions, similar to the traditional methods used in statistical parametric speech synthesis. Models were trained using the HMM-based speech synthesis system (HTS) [14] on a range of phonetic clusters encountered in 64 laughs from one person. Subjective evaluation resulted in an average naturalness score of 2.6 out of 5 for the synthesized laughs.

From this brief review of the literature, it is clear that the research on HMM-based laughter synthesis is scarce – there exists only one study on HMM-based laughter synthesis using a single vocoder. In this work, we report the role of four state-of-the-art vocoders commonly used in statistical parametric speech synthesis for the application of HMM-based laughter synthesis.

## 3. VOCODERS

The following vocoders were chosen for comparison: 1) Impulse train excited mel-cepstrum based vocoder, 2) STRAIGHT [15, 16] using mixed excitation, 3) Deterministic plus stochastic model (DSM) [17], and 4) GlottHMM vocoder [18]. All the vocoders use the source-filter principle for synthesis, and thus there are two components that mostly differ among the systems: the type of spectral envelope extraction and representation, and the method for modeling and generating the excitation signal. The vocoders are depicted in Table 1 and described in more detail in the following sections.

### 3.1. Impulse train excited mel-cepstral vocoder

The impulse train excited mel-cepstrum based vocoder (denoted in this work as MCEP) describes speech with only two acoustic features: $F_0$ and speech spectrum. The speech spectrum is estimated using the algorithm described in [19]. Mel-cepstral coefficients are commonly used as the spectral representation of speech as they provide a good approximation of the preceptually relevant speech spectrum. By changing the values of $\alpha$ (frequency warping) and $\gamma$ (factor defining generalization between LP and cepstrum), various types of coefficients for spectral representation can be obtained [19]. Here, we use $\alpha = 0.42$ and $\gamma = 0$ which correspond to simple mel-cepstral coefficients. Both $F_0$ and mel-cepstrum are estimated using the pitch function in speech signal processing toolkit (SPTK) [20], which uses the RAPT method [21]. Speech is synthesized by exciting the mel-generalized log spectral approximation (MGLSA) filter [22] with either simple impulse train for voiced speech or white noise for unvoiced speech. This simple excitation method has an effect that the synthesized signal often sounds buzzy.

| System | Parameters | Excitation |
|---|---|---|
| MCEP | mcep: 35 + $F_0$: 1 | Impulse + noise |
| STRAIGHT | mcep: 35 + $F_0$: 1<br>band aperiodicity: 21 | Mixed excitation<br>+ noise |
| DSM | mcep: 35 + $F_0$: 1 | DSM + noise |
| GlottHMM | $F_0$: 1 + Energy: 1 +<br>HNR: 5 + source LSF: 10<br>+ vocal tract LSF: 30 | Stored glottal<br>flow pulse +<br>noise |

**Table 1**. Vocoders in test and their parameters and excitation type.

### 3.2. STRAIGHT

STRAIGHT [15, 16] was proposed mainly for the high quality analysis, synthesis, and modification of speech signals. However, more often STRAIGHT is used as a reference for comparing between different vocoders in HMM-based speech synthesis, since it is the most widely used vocoder, is robust and can produce synthetic speech of good quality [23]. STRAIGHT decomposes the speech signal into three components: 1) spectral features extracted using pitch-adaptive spectral smoothing and represented as mel-cepstrum, 2) band-aperiodicity features which represent the ratios between periodic and aperiodic components of 21 sub-bands, and 3) $F_0$ extracted using instantaneous-frequency-based pitch estimation. In synthesis, STRAIGHT uses mixed excitation [24] in which impulse and noise excitations are mixed according to the band-aperiodicity parameters in voiced speech. The excitation of unvoiced speech is white Gaussian noise. Overlap-add is used to construct the excitation, which is then used to excite a mel log spectrum approximation (MLSA) filter [25] corresponding to the STRAIGHT mel-cepstral coefficients.

### 3.3. Deterministic plus stochastic model (DSM)

The deterministic plus stochastic model (DSM) of the residual signal [26] first estimates the speech spectrum, and uses the inverse of the filter to reveal the speech residual. Glottal closure instant (GCI) detection is used to extract individual GCI-centered residual waveforms, which are further resampled to fixed duration. The residual waveforms are then decomposed into the deterministic and stochastic parts in frequency domain, separated by the *maximum voiced frequency* $F_m$ fixed at 4 kHz. The deterministic part is computed as the first principal component of a codebook of residual frames centered on glottal closure instants and having a duration of two pitch periods. The stochastic part consists of a white Gaussian noise filtered with the linear prediction (LP) model of the average high-pass filtered residual signal, and time-modulated according to the average Hilbert envelope of the stochastic part of the residual. White Gaussian noise is used as excitation for unvoiced speech. The DSM excitation is then passed through the MGLSA filter. The DSM vocoder has been shown to reduce buzziness and to achieve comparable synthesis quality as that of STRAIGHT [26]. DSM vocoder was also used in the previous HMM-based laughter synthesis work [13]. In this paper, STRAIGHT is used to extract $F_0$ and mel-cepstrum for the DSM analysis, but the extraction of voice source features and synthesis is performed using the DSM vocoder.

### 3.4. GlottHMM

The GlottHMM vocoder uses glottal inverse filtering (GIF) in order to separate the speech signal into the vocal tract filter contribution and the voice source signal. Iterative adaptive inverse filtering (IAIF) [27] is used for the GIF, inside which LP is used for the estimation of the spectrum. IAIF is based on repetitively estimating and canceling the vocal tract filter and voice source spectral contribution from the speech signal. The output of the IAIF are the LP coefficients, which are converted to line spectral frequencies (LSF) [28] in order to achieve a better parameter representation for the statistical modeling, and the voice source signal that is further parameterized into various features. First, pitch is estimated from the voice source signal using autocorrelation method. Harmonic-to-noise ratio (HNR) of five frequency bands is estimated by comparing the upper and lower smoothed spectral envelopes constructed from the harmonic peaks and the interharmonic valleys, respectively. In addition, the voice source spectrum is estimated with LP and converted to LSFs.

In synthesis, a pre-stored natural glottal flow pulse is used for creating the excitation. First, the pulse is interpolated to achieve a desired duration according to $F_0$, scaled in energy, and mixed with noise according to the HNR measures. The spectrum of the excitation is then matched to the voice source LP spectrum, after which the excitation is fed to the vocal tract filter to create speech.

## 4. EVALUATION

A subjective evaluation was carried out to compare the performance of the 4 vocoders in synthesizing natural laughs. For each vocoder, two types of samples were used: a) copy-synthesis, which consists of extracting the parameters from a laugh signal and re-synthesizing the same laugh from the extracted parameters; b) HMM-based synthesis, where HMM-based system is trained from a laughter database and laughs are then synthesized using the models and the original phonetic transcriptions of a laughter. Copy-synthesis can be seen as the theoretically best synthesis that can be obtained with a particular vocoder, while HMM-based synthesis shows the current performance that can be achieved when synthesizing new laughs. Human laughs were also included in the evaluation for reference.

Our initial hypotheses were the following:

- H1: Human laughs are more natural than copy-synthesis and HMM laughs.
- H2: Copy-synthesis laughs are more natural than HMM laughs, as they omit the modeling stage.
- H3: All vocoders are equivalent for laughter synthesis.

The third hypothesis concerns the comparison of the vocoders among themselves, which is the main objective of this work. The way this hypothesis is formulated illustrates the fact that we do not have a priori expectations that one vocoder would be better suited for laughter than other vocoders.

### 4.1. Data

For the purpose of this work, two voices from the AVLaughterCycle database [29] were selected: a female voice (subject 5, 54 laughs) and a male voice (subject 6, the same voice as in previous work [13], 64 laughs). As in [13], phonetic clusters were formed by grouping acoustically close phones found in the narrow phonetic annotations of the laughs [30]. This resulted in 10 phonetic clusters used for synthesis: 3 for consonants (nasals, fricatives and plosives), 4 for vowels (ə, a, ɪ and o), and 3 additional clusters were formed with typical laughter sounds: grunts, cackles, and nareal fricative (noisy airflow expelled through the nostrils). Inhalation and exhalation phones are distinguished and form separate clusters. Hence there are 20 clusters in total when considering both inhalation and exhalation clusters. For each voice, the phonetic clusters that did not have at least 11 occurrences were assigned to a garbage class.

For each voice and each of the considered vocoders and extracted parameters (see Table 1), HMM-based systems were trained using the standard HTS procedure [14, 31] using all the available laughs. For the test, five laughs lasting at least 3.5 seconds were randomly selected for each voice. For each vocoder, these laughs were synthesized from their phonetic transcriptions (HMM synthesis) as well as re-synthesized directly from their extracted parameters (copy-synthesis). The 5 original laughs were also included in the evaluation. This makes a total of 5 (original laughs) + 5 $\times$ 2 (HMM and copy-synthesis) $\times$ 4 (number of vocoders) = 45 laughs in the evaluation set for both voices.

### 4.2. Evaluation setup

A subjective evaluation was carried out using a web-based listening test, where listeners were asked to rate the quality of synthesized laughter signals on a 5-point Likert scale [10]. Participants were suggested to use headphones, and were then presented one laugh at a time. Participants could listen to the laugh as many times as they wanted and were asked to rate its naturalness on a 5-point Likert scale where only the highest (completely natural) and lowest (completely unnatural) options were labeled. The 45 laughter signals were presented in random order. 18 participants evaluated the male voice while 15 evaluated the female one. All listeners were between 25–35 years of age, and some of them were speech experts.

## 5. RESULTS

Figure 1 shows the means and 95% confidence intervals of the naturalness ratings for copy-synthesis (right) and HMM synthesis (left) of the male (upper) and female (lower) voices. The pairwise $p$-values (using the Bonferroni correction) between vocoders are shown in Table 2 for copy-synthesis and in Table 3 for HMM synthesis.

As expected (H1), original human laughs were perceived as more natural than all other laughs (copy-synthesis and HMM). In addition, H2 was also confirmed: for each vocoder, the naturalness achieved with copy-synthesis was significantly higher than with HMM synthesis. The most interesting is the comparison between the vocoders (H3). In copy-synthesis, GlottHMM was rated as less natural than all other vocoders (for both female and male), MCEP and DSM obtained similar naturalness scores, while STRAIGHT was slightly preferred for female laughs (but not for male laughs). This may indicate that STRAIGHT is potentially the most suitable vocoder for laughter synthesis with the female voice, while MCEP, DSM, and STRAIGHT are equivalently good for the male voice. This trend is generally confirmed when looking at HMM-based laughter synthesis (right plots), where it appears that MCEP obtained the best results for the female voice, followed by DSM, STRAIGHT and finally GlottHMM. For the male laughs, DSM achieved the best results, slightly over STRAIGHT and finally MCEP and GlottHMM, which were rated as similar. However, the only statistically significant differences with HMM synthesis were for the female voice with MCEP (significantly more natural than STRAIGHT and GlottHMM) and DSM (significantly better than GlottHMM).

These results indicate that MCEP and DSM are in general good choices for laughter synthesis. Both vocoders use simple parameter representation in statistical modeling: only $F_0$ and spectrum are

| Female | System | DSM | Glott | MCEP | STR | Nat |
|---|---|---|---|---|---|---|
| | DSM | – | **0.006** | 1 | 1 | **0** |
| | Glott | **0.006** | – | **0.04** | **0.002** | **0** |
| | MCEP | 1 | **0.04** | – | 1 | **0** |
| | STR | 1 | **0.002** | 1 | – | **0** |
| | Nat | **0** | **0** | **0** | **0** | – |
| Male | System | DSM | Glott | MCEP | STR | Nat |
| | DSM | – | **0.003** | 1 | 1 | **0** |
| | Glott | **0.003** | – | **0** | **0.002** | **0** |
| | MCEP | 1 | **0** | – | 1 | **0.27** |
| | STR | 1 | **0.002** | 1 | – | **0** |
| | Nat | **0** | **0** | **0.27** | **0** | – |

**Table 2**. Pairwise $p$-values between the vocoders copy-synthesis and natural laughs. Statistically significant results are marked in bold.

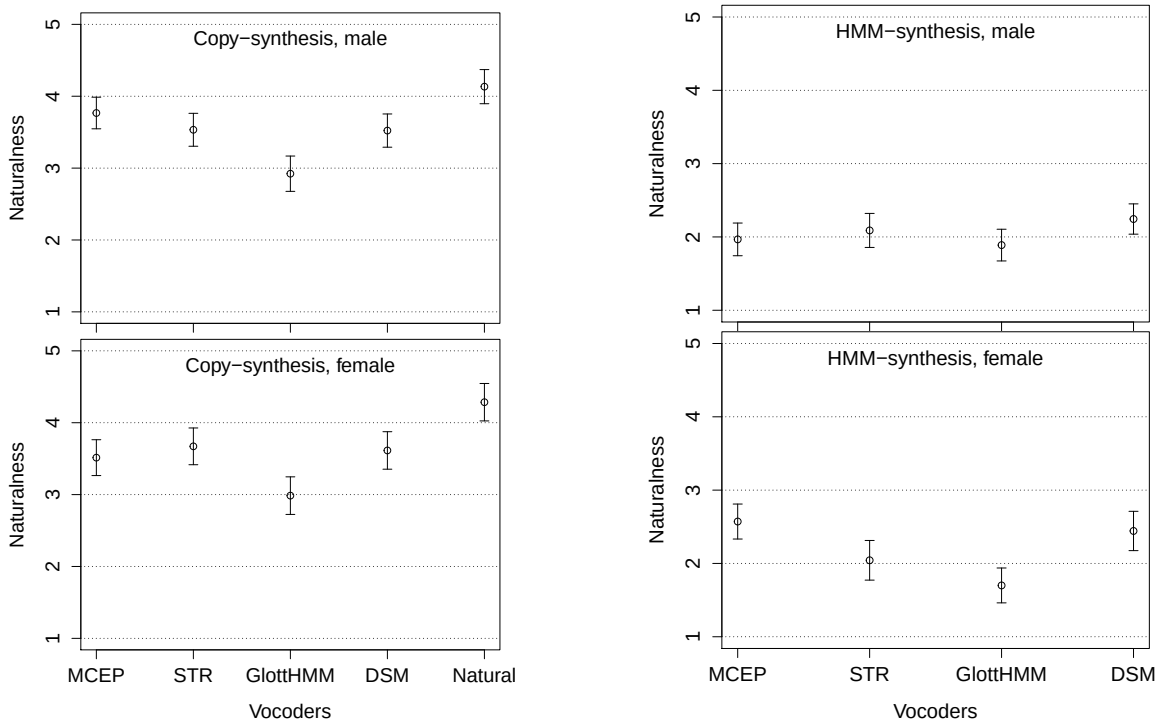



**Fig. 1**. Naturalness scores for copy-synthesis (left) and HMM synthesis (right) for the male (upper) and female (lower) speakers.

modeled and all other features are fixed. Accordingly, the synthesis procedure of these vocoders is very simple: the excitation generation depends only on the modeled $F_0$. In DSM, $F_m$, residual waveform, and noise time envelope are fixed and thus they cannot produce additional artefacts beyond possible errors in $F_0$ and spectrum. MCEP obtained the best naturalness scores for the female voice, although the known drawback of this method is its buzziness. This was likely not too disturbing as the the female voice used few voiced segments. The buzziness could, however, explain why male laughs synthesized with MCEP were perceived as less natural than female laughs, since the male laughs contained more and longer voiced segments.

STRAIGHT performed better in copy-synthesis with a female voice but cannot hold this advantage in HMM-based laughter synthesis, when statistical modeling is involved. This may well be due to the modeled aperiodicity parameters, which are difficult to estimate from the challenging laughter signals, consisting a lot of partly voiced sounds. Moreover, STRAIGHT pitch estimation is known to be unreliable with non-modal voices (see e.g. [32]), which is very often the case with laughter. Thus, the estimated aperiodicity parameters may have a lot of inconsistent variation, thus degrading the statistical modeling of the parameters. Therefore, in HMM synthesis, the mixed excitation may fail to produce an appropriate excitation.

GlottHMM also suffers occasionally from pitch estimation errors, especially if the voicing settings are not accurately set or speech material is challenging. At least the latter is true with laughter, in which the vocal folds do not reach a complete closure as in modal speech [33]. Pitch estimation errors are even more harmful for the GlottHMM vocoder than the other vocoders since the analysis of voiced and unvoiced sounds is treated completely in a different manner. Thus, voicing errors generate severe errors in the output parameters of GlottHMM. GlottHMM is also considerably more complex than the other systems, thus making the statistical modeling of all the parameters challenging with small amount of data.

Finally, the role of the training material was not studied in this experiment, but it is expected that it also has a significant effect, especially when dealing with challenging material such as laughter.

| Female | System | DSM | Glott | MCEP | STR |
|---|---|---|---|---|---|
| | DSM | – | **0.003** | 1 | 0.16 |
| | Glott | **0.003** | – | **0** | 0.34 |
| | MCEP | 1 | **0** | – | **0.02** |
| | STR | 0.16 | 0.34 | **0.02** | – |
| Male | System | DSM | Glott | MCEP | STR |
| | DSM | – | 0.14 | 0.46 | 1 |
| | Glott | 0.14 | – | 1 | 1 |
| | MCEP | 0.46 | 1 | – | 1 |
| | STR | 1 | 1 | 1 | – |

**Table 3**. Pairwise $p$-values between HMM synthesis of different vocoders. Statistically significant results are marked in bold.

## 6. SUMMARY AND CONCLUSIONS

This paper presented an experimental comparison of four vocoders for HMM-based laughter synthesis. The results show that all vocoders perform relatively well in copy-synthesis. However, in HMM-based laughter synthesis, all synthesized laughter voices were significantly lower in quality than in copy-synthesis. The evaluation results revealed that two vocoders using rather simple and robust excitation modeling performed the best, while two other vocoders using more complex analysis, parameter representation, and synthesis suffered from the statistical modeling. These findings suggest that the robustness of parameter extraction and representation is a key factor in laughter synthesis, and increased efforts should be directed on enhancing the robust estimation and representation of the acoustic parameters of laughter.

## 7. REFERENCES


[1] J. Robson and J. MackenzieBeck, "Hearing smiles-perceptual, acoustic and production aspects of labial spreading," in *Proc. of Inter. Conf. of the Phon. Sci. (ICPhS)*, San Francisco, USA, 1999, pp. 219–222.

[2] N. Campbell, "Conversational speech synthesis and the need for some laughter," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 14, no. 4, pp. 1171–1178, 2006.

[3] S. Petridis and M. Pantic, "Audiovisual discrimination between speech and laughter: Why and when visual information might help," *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 216–234, 2011.

[4] J.-A. Bachorowski, M. J. Smoski, and M. J. Owren, "The acoustic features of human laughter," *J. Acoust. Soc. Am.*, vol. 110, no. 3, pp. 1581–1597, 2001.

[5] J.-A. Bachorowski and M. J. Owren, "Not all laughs are alike: Voiced but not unvoiced laughter readily elicits positive affect," in *Psychological Science*, 2001, vol. 12, pp. 252–257.

[6] K. P. Truong and D. A. van Leeuwen, "Automatic discrimination between laughter and speech," *Speech Commun.*, vol. 49, pp. 144–158, 2007.

[7] M. T. Knox and N. Mirghafori, "Automatic laughter detection using neural networks," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 2973–2976.

[8] L. Kennedy and D. Ellis, "Laughter detection in meetings," in *NIST ICASSP 2004 Meeting Recognition Workshop*, Montreal, 2004, pp. 118–121.

[9] S. Sundaram and S. Narayanan, "Automatic acoustic synthesis of human-like laughter," *J. Acoust. Soc. Am.*, vol. 121, no. 1, pp. 527–535, 2007.

[10] R. Likert, "A technique for the measurement of attitudes," *Archives of psychology*, 1932.

[11] E. Lasarcyk and J. Trouvain, "Imitating conversational laughter with an articulatory speech synthesis," in *Proc. of Interdisciplinary Workshop on the Phonetics of Laughter*, Saarbrücken, Germany, 2007, pp. 43–48.

[12] T. Sathya Adithya, K. Sudheer Kumar, and B. Yegnanarayana, "Synthesis of laughter by modifying excitation characteristics," *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp. 3072–3082, 2013.

[13] J. Urbain, H. Cakmak, and T. Dutoit, "Evaluation of hmm-based laughter synthesis," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, Vancouver, Canada, 2013, pp. 7835–7839.

[14] [Online], "HMM-based speech synthesis system (HTS)," http://hts.sp.nitech.ac.jp/.

[15] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3–4, pp. 187–207, 1999.

[16] H. Kawahara, Jo Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *2nd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, 2001.

[17] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 20, no. 3, pp. 968–981, 2012.

[18] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "Hmm-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 19, no. 1, pp. 153–165, 2011.

[19] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis – A unified approach to speech spectral estimation," in *Proc. ICSLP*, 1994, vol. 94, pp. 18–22.

[20] [Online], "Speech signal processing toolkit (SPTK) v. 3.6," 2013.

[21] D. Talkin, "A robust algorithm for pitch tracking (rapt)," in *Speech Coding and Synthesis*, W. B. Klein and K. K. Palival, Eds. Elsevier, 1995.

[22] T. Kobayashi, S. Imai, and T. Fukuda, "Mel generalized log spectrum approximation (MGLSA) filter," *Journal of IEICE*, vol. J68-A, no. 6, pp. 610–611, 1985.

[23] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of nitech hmm-based speech synthesis system for the blizzard challenge 2005," in *IEICE Trans. Inf. and Syst.*, 2007, vol. E90-D, pp. 325–333.

[24] T. Yoshimura, K. Tokuda, T. Masuko, and T. Kitamura, "Mixed-excitation for HMM-based speech synthesis," *Proc. Eurospeech*, pp. 2259–2262, 2001.

[25] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, 1992, vol. 1, pp. 137–140.

[26] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 20, no. 3, pp. 968–981, 2012.

[27] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Commun.*, vol. 11, no. 2–3, pp. 109–118, 1992.

[28] F. K. Soong and B.-H. Juang, "Line spectrum pair (LSP) and speech data compression," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, Mar. 1984, vol. 9, pp. 37–40.

[29] J. Urbain, E. Bevacqua, T. Dutoit, A. Moinet, R. Niewiadomski, C. Pelachaud, B. Picart, J. Tilmanne, and J. Wagner, "The AVLaughterCycle database," in *Proc. of Seventh conference on Intl Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010, pp. 2996–3001.

[30] J. Urbain and T. Dutoit, "A phonetic analysis of natural laughter, for use in automatic laughter processing systems," in *Proc. of $4^{th}$ bi-annual Intl Conf. of the HUMAINE Association on Affective Computing and Intelligent Interaction (ACII2011)*, Memphis, Tennesse, 2011, pp. 397–406.

[31] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Sixth ISCA Workshop on Speech Synthesis*, 2007, pp. 294–299.

[32] T. Raitio, J. Kane, T. Drugman, and C. Gobl, "HMM-based synthesis of creaky voice," in *Proc. Interspeech*, 2013, pp. 2316–2320.

[33] Wallace Chafe, *The Importance of not being earnest. The feeling behind laughter and humor.*, vol. 3 of *Consciousness & Emotion Book Series*, John Benjamins Publishing Company, Amsterdam, The Nederlands, paperback 2009 edition, 2007.

## 6.4 Evaluation of the periodic and aperiodic components in excitation modelling

# EXCITATION MODELING FOR HMM-BASED SPEECH SYNTHESIS: BREAKING DOWN THE IMPACT OF PERIODIC AND APERIODIC COMPONENTS

*Thomas Drugman*[1]*, Tuomo Raitio*[2]

[1]TCTS Lab - University of Mons, Belgium
[2]Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland

**ABSTRACT**

HMM-based speech synthesis generally suffers from typical buzziness due to over-simplified excitation modeling of voiced speech. In order to alleviate this effect, several studies have proposed various new excitation models. No consensus has however been reached on what is the perceptual importance of the accurate modeling of the periodic and aperiodic components of voiced speech, and to what extent they separately contribute in improving naturalness. This paper considers a generalized mixed excitation modeling, common to various existing approaches, in which both periodic and aperiodic components coexist. At least three main factors may alter the quality of synthesis: periodic waveform, noise spectral weighting, and noise time envelope. Based on a large subjective evaluation, the goal of this paper is threefold: *i)* to evaluate the relative perceptual importance of each factor, *ii)* to investigate what is the most appropriate method to model the periodic and aperiodic components, and *iii)* to provide prospective clues for future work in excitation modeling.

***Index Terms***— HMM-based speech synthesis, excitation modeling, glottal flow, residual signal

## 1. INTRODUCTION

Statistical parametric speech synthesis based on Hidden Markov Models (HMMs) [1] emerged this last decade as a promising technique for the automatic generation of speech from text. This approach exhibits several advantages over concatenative speech synthesis approach [2]: flexibility to change the voice characteristics [3, 4, 5, 6], reduced memory footprint [7, 8], and enhanced robustness [9]. Nonetheless, although some progress has been achieved these last years, its main flaw is a degraded speech quality. This can be explained by two main factors: *i)* the synthesis relies on a parametric representation of the speech signal which results in a typical *buzziness*; *ii)* the synthesis relies on a statistical modeling of a given speech database, which results in a typical *muffledness* caused by oversmoothed generated trajectories.

This paper addresses the first issue and aims to enhance the naturalness of synthesized speech by improving the *excitation modeling*. In speech processing, the modeling of speech is generally based on the source-filter approach. In this framework, two options are possible according to what is considered to be the source and the filter. In the first case, the source is the glottal (air) flow as physiologically produced by the vocal folds, and the filter refers to the vocal tract response. Beyond the physiological motivation, this approach has the advantage to be more flexible, as proper modifications of the glottal contribution are expected to reflect changes in voice quality. Nonetheless, this approach requires to reliably and accurately separate these components from each other using glottal inverse filtering, which is a difficult inverse problem. In the second case, the filter corresponds to the overall spectral envelope of speech and the excitation is the residual signal obtained by filtering speech signal with the inverse of the estimated filter. The residual signal has the advantage to be easily obtained, however its amplitude spectrum is by definition flat and the information about the glottal spectral shaping is inextricably mixed in the filter component. As a consequence, its flexibility for speech modifications is more limited.

In all cases, separating the source and filter contribution is important as it can lead to their distinct characterization and modeling. Methods parameterizing the filter, such as the well-known linear prediction (LP) or mel-cepstrum like features [10], are widely used. On the contrary, methods modeling the excitation signal are still not well established and the accurate and perceptually relevant modeling of the excitation would benefit many speech speech processing areas.

The basic excitation model makes use of either a quasi-periodic pulse train for voiced speech, or white noise for unvoiced speech. This simplistic representation of voiced speech makes the resulting synthesis sound buzzy due to unnaturally strong higher harmonics. Various studies have focused on improving the excitation model by mixing periodic excitation with aperiodic noise, such as in the Mixed Excitation (ME) [11] approach. In ME, voiced excitation is composed of both periodic and aperiodic components of which relative magnitudes are controlled by band-pass voicing strengths. In a similar way, a ME consisting of a set of high-order state-dependent filters derived through a closed-loop procedure was proposed in [12]. In [13], a hybrid approach makes use of a codebook of pitch-synchronous residual frames which are selected at synthesis time according to the down-sampled version of the excitation. In [14, 15], the Deterministic plus Stochastic Model (DSM) of the residual signal is proposed. DSM excitation consists of two components: the deterministic waveform called eigenresidual, which is obtained by Principal Component Analysis (PCA) on a set of pitch-synchronous residual frames, and an aperiodic excitation delimited by maximum voiced frequency and modulated in time according to a speaker-specific time envelope.

In parallel, similar improvements using a glottal flow modeling have been introduced. The approach described in [16] incorporates the Liljencrants–Fant (LF) [17] model so as to reduce the buzziness and increase the flexibility. A natural glottal flow pulse estimated by glottal inverse filtering from a sustained vowel is modified according to voice source features and mixed with noise in the so-called GlottHMM approach presented in [18] and further refined in [19]. A synthesis approach using LF model was also introduced in [20]. In [21], a glottal source pulse library is extracted from natural speech and

T. Drugman is supported by FNRS. T. Raitio is supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287678. The authors would like to thank Vasilis Karaiskos for running the listening tests.

pulses are selected according to voice source features for synthesis. All these techniques (modeling either residual or glottal flow) have been shown to provide a higher naturalness in HMM-based speech synthesis, compared to the traditional pulse excitation.

Despite all the advances in excitation modeling, no consensus has been reached yet on the perceptual effect of each component in voice source modeling, and to what extent they separately contribute in improving naturalness. In the frame of HMM-based speech synthesis, this papers investigate the perceptual impact of the three main factors in excitation modeling: waveform used for periodic excitation, spectral weighting between the periodic and aperiodic components, and the envelope used for the time modulation of the noise. The goal of this paper is threefold: *i)* to evaluate the relative importance each component in modeling the excitation, *ii)* to investigate what is the most appropriate method to model these components, *iii)* to provide prospective clues for future work in excitation modeling.

The paper is structured as follows. Section 2 presents the general vocoding framework used in various existing approaches and describes the alternatives considered throughout this paper. Section 3 deals with the experimental protocol, providing details about the implementation of our HMM-based speech synthesizers and describing the subjective evaluation and its results. Section 4 finally discusses the implications of the study and concludes the paper.

## 2. GENERAL VOCODING FRAMEWORK

The great majority of excitation models rely on a similar mixed excitation model in which both periodic and aperiodic components coexist during the production of voiced sounds. The workflow of this generalized vocoder is displayed in Fig. 1. The periodic contribution of the excitation $e_p(t)$ is obtained from a specific waveform whose duration is adapted to the current $F_0$ value, and which is then filtered using some aperiodicity measurements. As for the aperiodic excitation component $e_a(t)$, it results from a white Gaussian noise that is spectrally modified using these same aperiodicity measurements and modulated in time using a given envelope. Note that all this process is achieved pitch-synchronously. The two components $e_p(t)$ and $e_a(t)$ are then summed up and the pitch-synchronous windowed frames are overlap-added. The resulting excitation contribution finally goes through the filter to give the speech signal. The three main factors impacting the performance of this generalized excitation model are now studied in the remainder of this paper: periodic waveform, noise spectral weighting and noise envelope.

### 2.1. Periodic Waveform

In the simplest source-filter vocoder, Dirac pulses at fundamental period intervals are used to create the voiced excitation. Usually improvements in excitation modeling are compared with either this

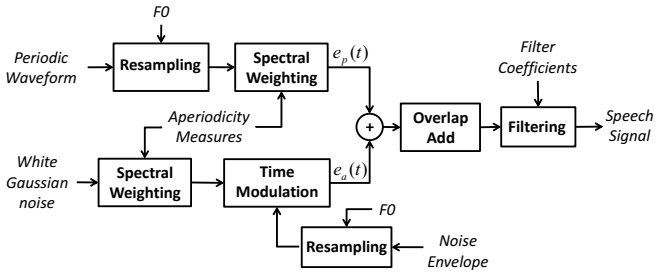



**Fig. 1**. Workflow of generalized vocoder using mixed excitation.

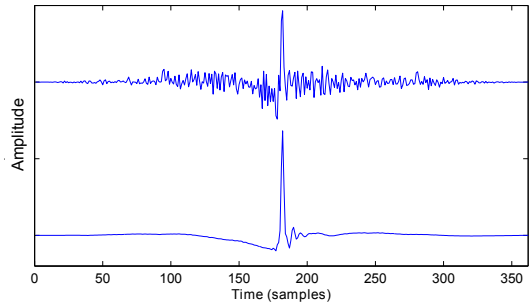



**Fig. 2**. Natural residual excitation frame (*upper signal*) and eigenresidual (*lower signal*) for speaker AWB.

simple model or the mixed excitation [11], which is used e.g. in the most commonly used vocoder STRAIGHT [22, 23]. Improvements over the simple excitation are rather easy to achieve either by using more natural periodic waveform or by mixing the periodic component with noise. However, the comparison between more complex methods (e.g. STRAIGHT) may be ambiguous, since evaluations are usually made between whole vocoder architectures using different parameterization methods, parameters representations, and HMM training. Also, the contributions of the periodic and aperiodic components are usually left undetermined.

Only few studies have evaluated the perceptual differences between various deterministic waveforms other impulse train. Experiments in [24] have shown that mean glottal flow pulse of a pulse library (similar to eigenresidual in [15]) was rated better in quality than excitation using pulse library and a pulse reconstructed from 12 PCA components. The latter comparison was also informally done in [15] with the same conclusion that adding more components does not improve the quality. The lower quality of the pulse library method was due to slight irregularities in the excitation due to imperfect pulse selection. Also in creaky voice synthesis, the excitation waveform has been shown to have relevant perceptual effect [25].

In this paper, we consider the reconstruction of the residual signal with three possible periodic waveforms: *i)* the Dirac impulse as used in the simplest vocoder; *ii)* a *natural* excitation residual frame; *iii)* speaker-dependent eigenresidual as proposed in [15]. Note that the choice of the natural residual frame was not arbitrary and resulted from the consideration of several criteria: *a)* having a low pitch to avoid as much as possible up-sampling to the target $F_0$ (as this will cause energy holes in high frequencies); *b)* its amplitude spectrum must be as flat as possible to avoid artefacts due to residual resonances; *c)* having a clear discontinuity at the GCI. The natural residual and eigenresidual for the male speaker considered in this paper are illustrated in Fig. 2.

### 2.2. Spectral Weighting

In order to reduce the buzziness caused by a too strong harmonicity, it has been shown to beneficial to adopt an approach in which both periodic and aperiodic components may coexist [11]. Two main techniques were proposed in the literature for this purpose. The first one relies on a multiband approach where, for each spectral band, the energy of periodic and aperiodic contributions is controlled by *aperiodicity measurements*. These measurements can be computed in various ways. In [11], they consist of correlation coefficients calculated in each band. In [19], they are derived from the strength of the cepstral fundamental period peak, while in [23, 21] they are determined based on the ratio between the upper and lower smoothed spectral

envelopes. The second technique for spectral weighting makes use of a maximum voiced frequency (usually noted $F_m$) which demarcates the boundary between the periodic component (which holds only in the low frequencies) and the aperiodic component (which holds only in the high frequencies). This idea originates from the Harmonic plus Noise Model (HNM) of speech [26], and was later integrated into several methods for excitation modeling in HMM-based speech synthesis [14], [20].

The perceptual effect of these difference methods has not been studied in the context of HMM-based speech synthesis, Thus, four options for spectral weighting are investigated in this paper: *i)* the aperiodic component is discarded and the excitation consists only of the periodic contribution; *ii)* use of a static maximum voiced frequency $F_m$ fixed to 4 kHz as is done in [27] and [15]; *iii)* use of dynamic $F_m$ value estimated using the algorithm described in [26]; *iv)* use of the HNR measurements proposed in [21].

### 2.3. Envelope for Noise Modulation

In addition to modeling the spectral characteristics of the noise (as described in Section 2.2), some studies have addressed its time properties. The motivation for this arises from the observation that the time distribution of the noise is not uniform and exhibits a synchronization with the glottal cycle. In [26], a pitch-synchronous parametric triangular envelope is proposed. In [28], authors compare the triangular to Hilbert energy envelopes in the frame of HNM and report a slight improvement. In [29], an alternative parametric representation of a triangular envelope is proposed. It is however worth mentioning that none of these works have been tested in the context of HMM-based speech synthesis, which requires slowly-varying parameter trajectories for a proper statistical modeling. Finally, a speaker-dependent noise waveform envelope was proposed in [15], which is extracted by averaging glottal closure instant (GCI synchronous Hilbert envelopes of the stochastic part of the excitation.

Three possible noise envelopes are further studied in this paper: *i)* uniform distribution; *ii)* the triangular window proposed in [26]; *iii)* the speaker-dependent Hilbert envelope proposed in the DSM approach [15]. An illustration of this latter waveform is shown in Fig. 3 for the female speaker considered in this study.

## 3. EXPERIMENTS

### 3.1. HMM-based Voice Building

In order to find out the perceptual effects for each of the studied excitation component, HMM-based voices were built and used in subjective listening tests. To prevent perceptual effects due to other factors than the ones in study, a single system architecture was used that is capable of producing all the different component combinations. The speech features used to train the HMMs are depicted in Table 1. In feature extraction, fundamental frequency ($F_0$) and HNR were extracted using the GlottHMM vocoder [19, 21] while SPTK 3.6 [30] was used to extract the speech spectrum. The spectrum was parameterized using a 30th order mel-generalised cepstral (MGC) analysis [31] with $\alpha = 0.42$ and $\gamma = -1/3$. MGCs were then converted to line spectral frequencies (LSF) for better parameter representation for HMM training. The maximum voiced frequency was estimated by the algorithm described in [26]. All other data such as the periodic waveform or the noise envelope have been extracted as explained in Section 2 by a GCI-synchronous analysis, where GCIs are detected using the SEDREAMS algorithm [32].

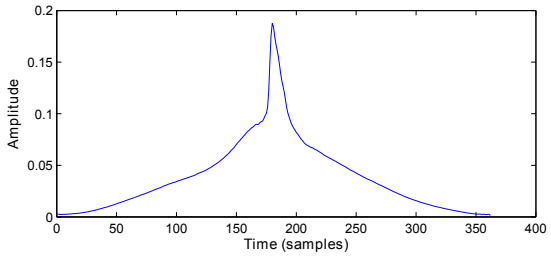



**Fig. 3**. Speaker-dependent Hilbert envelope for speaker SLT.

The HTS 2.1 HMM architecture [33] was used for training. All features were modeled in individual streams. Only $F_0$ and spectrum were used for the alignment. In synthesis, parameters were generated considering global variance [34] except for the spectrum. Excitation was generated using the vocoder described in Section 2 where the excitation waveform and noise modeling were varied according to the desired setup. Finally, the excitation was filtered with the mel-generalised log spectral approximation (MGLSA) filter [35].

Two databases recorded for the purpose of developing text-to-speech (TTS) synthesis were used to build voices for the experiments. These voices are Scottish English male AWB and US English female SLT from the ARCTIC database [36], which consist of 1,138 and 1,132 sentences, respectively. 1,000 sentences were used for training both voices and the rest was used for testing.

### 3.2. Subjective Evaluations

Subjective evaluation was performed in three separate steps in order to find out the effect of each component and also their possible interactions. The idea was to first select the best noise spectral weighting according to a subjective evaluation among 4 systems. Then, the best spectral weighting method according to the first evaluation is used to study the effect of the noise time envelope, in which 3 systems are evaluated. Finally, in the third test, both the best noise spectral weighting and the best time envelope are used in the study of the effect of the periodic waveform, in which 3 systems are compared.

Comparison Category Rating (CCR) test was used in order to determine the quality difference between the systems. In CCR test, listeners are presented with speech sample pairs from which listeners rate the difference of the two samples on the comparison mean opinion score (CMOS) scale, which is a discrete seven-point scale ranging from *"much worse"* ($-3$) to *"much better"* (3). All possible system combinations were evaluated (e.g. for three systems: 1–2, 1–3, 2–3) in both directions (e.g. 1–2 and 2–1). Thus, there were 6 comparisons per sample for 3-system test and 12 for the 4-system test. The CCR test responses were summarized by calculating the mean scores and 95% confidence intervals for each evaluated method. The mean yields the order of preference and distances between all the methods (i.e., the amount of preference relative to each other). A Wilcoxon signed-rank test was finally used (as the scores were rarely normally distributed) for further testing the significance between the means of each method pair. The systems used across the 3 CCR tests are summarized in Table 2 in concordance with the methods explained in Section 2.

**Table 1**. Speech features used for training the HMM system.

| Feature | N. of params. |
|---|---|
| Fundamental frequency | 1 |
| Maximum voiced frequency ($F_m$) | 1 |
| Harmonic-to-noise ratio | 5 |
| Mel-generalized cepstrum | 30 |

All test samples (137 for AWB and 132 for SLT) were synthesized for the three tests using each system ($4 + 3 + 3 = 10$ systems) and were included in the listening tests. Thus, a total of 2,690 ($10 \times (137 + 132)$) samples were synthesized. The loudness of the sentences were normalized according to ITU-T P.56. In order to reduce the workload on participants, 5 sentences from each speaker were randomly selected for each participant and presented to them in each test. Also ten null pairs (same samples in the pair) were included in order to test the consistency of the listeners. Thus each participant rated a total of 130, 70, and 70 stimuli pairs in the first, second, and third test, respectively.

Listening tests were performed in sound proof booths with high-quality headphones. All participants were university students and native speakers of English, and they were paid for the participation. 24, 21, and 24 listeners participated in the three tests, respectively. However, after inspection of the results, some participants were removed due to inconsistent results for the null pairs. Thus, results from 20 listeners in each test were finally used.

### 3.3. Results

In the first test (CCR1), the perceptual effect of the noise spectral weighting was studied by evaluating the 4 approaches presented in Table 2. The basic Dirac pulse was used as the periodic waveform in synthesis in order to emphasize the perceptual effect of the noise models. Constant time envelope was also used. The results are shown in Figure 4 (uppermost graph). Discrepancies are observed across male and female speakers. For male, HNR and DynFm show no statistically significant difference, but for the female voice, DynFm is rated better. FixFm is rated always worse than HNR and DynFm except for the female speaker. System without any noise (Imp) is always rated the worst. These results are also confirmed by the statistical Wilcoxon test. Since DynFm was rated better or equal than the rest of the systems, it is used in the rest of the experiments.

In the second test (CCR2), the effect of the noise time envelope was studied. The 3 systems considered in CCR2 are depicted in Table 2 (middle part) and the corresponding results are shown in Figure 4 (middle graph). The results show no statistically significant differences between the methods. Thus, the results indicate that the noise time envelope has no perceptual relevance, and the simplest one, constant time envelope, is used in the third experiment.

In the third test (CCR3), the effect of periodic waveform was studied by including the 3 systems in Table 2 (bottom part). The corresponding results are shown in Figure 4 (bottom graph). Results also diverge across male and female speakers. For male, the natural residual frame and the eigenresidual are rated equally good while the impulse excitation is rated worse than the natural residual. For the female speaker, impulse excitation and eigenresidual are rated equal while natural residual is rated worse that the two others. If scores are averaged, there are no statistically significant differences. These results are confirmed by the Wilcoxon signed-rank test.

**Table 2**. Systems in the three subjective evaluations (CCR1/2/3).

| CCR1 | Effect of noise spectral weighting |
|---|---|
| Imp | Impulse excitation without noise |
| FixFm | Impulse excitation + noise according to fixed $F_m$ |
| DynFm | Impulse excitation + noise according to dynamic $F_m$ |
| HNR | Impulse excitation + noise according to HNR |

| CCR2 | Effect of noise time envelope |
|---|---|
| Con | Imp. exc. + dyn. $F_m$ noise + constant time envelope |
| Tri | Imp. exc. + dyn. $F_m$ noise + triangular time envelope |
| DSM | Imp. exc. + dyn. $F_m$ noise + DSM time envelope |

| CCR3 | Effect of deterministic waveform |
|---|---|
| Imp | Impulse excitation + dyn. $F_m$ noise + const. time env. |
| Nat | Natural residual + dyn. $F_m$ noise + const. time env. |
| Eig | Eigenresidual + dyn. $F_m$ noise + const. time env. |

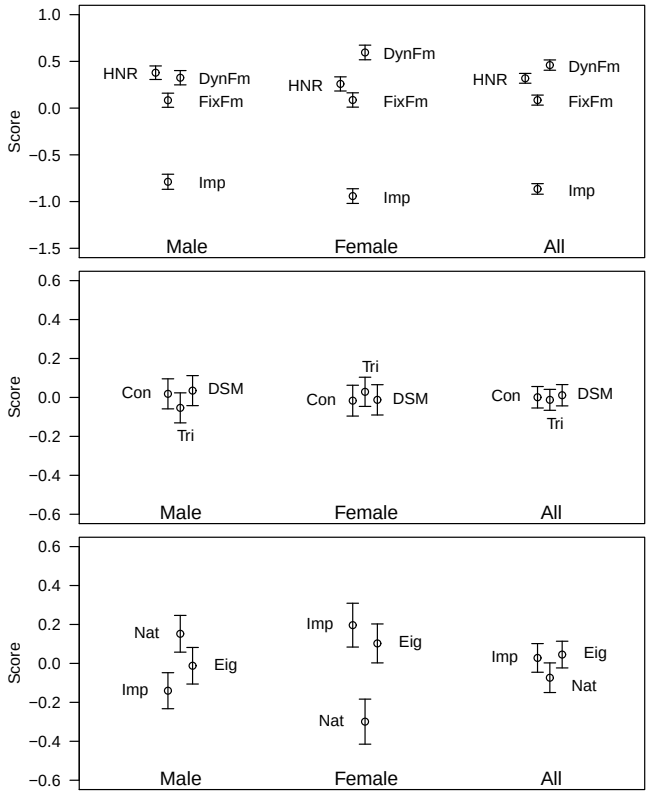



**Fig. 4**. Results of the subjective evaluation comparing noise spectral weighting (uppermost), noise time envelope (middle), and periodic waveform (bottom).

## 4. CONCLUSION

This paper addressed the problem of excitation modeling in order to improve the naturalness in HMM-based speech synthesis. Based on a generalized vocoder, three main factors influencing the quality of synthesis were studied: periodic waveform, noise spectral weighting, and noise time envelope. A subjective evaluation was performed in order to determine the perceptual importance of each factor. Our results clearly indicate that: *i)* the spectral weighting is an essential feature as it leads to the greatest perceptual differences; *ii)*; incorporating a noise model during the production of voiced sound is crucial. This can be efficiently achieved based on HNR measures or using a maximum voiced frequency; *iii)* the perceptual impact of the noise envelope seems to be negligible; *iv)* it is necessary to adapt the periodic waveform according the speaker's $F_0$ range as it will affect the excitation phase properties. These conclusions should be carefully considered when designing new excitation models. As a result, we believe that future research efforts should focus on new strategies to weight the energy of both periodic and aperiodic components in several spectral bands, as well as on a better understanding of the phase information in the periodic waveform.

## 5. REFERENCES


[1] Heiga Zen, Keiichi Tokuda, and Alan W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, 1996, pp. 373–376.

[3] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation in HMM-based speech synthesis system," in *Proc. Eurospeech*, 1997, pp. 2523–2526.

[4] K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," in *Proc. ICSLP*, 2002, pp. 1269–1272.

[5] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 9, pp. 1406–1413, 2007.

[6] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 17, no. 1, pp. 66–83, 2009.

[7] S.-J. Kim, J.-J. Kim, and M.-S. Hahn, "HMM-based Korean speech synthesis system for hand-held devices," *IEEE Trans. Consum. Electron.*, vol. 52, no. 4, pp. 1384–1390, 2006.

[8] A. Gutkin, X. Gonzalvo, S. Breuer, and P. Taylor, "Quantized HMMs for low footprint text-to-speech synthesis," in *Proc. Interspeech*, 2010, pp. 837–840.

[9] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 17, no. 6, pp. 1208–1230, 2009.

[10] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - A unified approach to speech spectral estimation," *Proc. ICSLP*, 1994.

[11] T. Yoshimura, K. Tokuda, T. Masuko, and T. Kitamura, "Mixed-excitation for HMM-based speech synthesis," *Eurospeech*, pp. 2259–2262, 2001.

[12] R. Maia, T. Toda, H. Zen, Y. Nankaku, and K. Tokuda, "An excitation model for HMM-based speech synthesis based on residual modeling," *ISCA SSW6*, 2007.

[13] T. Drugman, G. Wilfart, A. Moinet, and T. Dutoit, "Using a pitch-synchronous residual codebook for hybrid HMM/frame selection speech synthesis," *ICASSP*, pp. 3793–3796, 2009.

[14] T. Drugman, G. Wilfart, and T. Dutoit, "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis," *Interspeech*, 2009.

[15] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *IEEE Trans. on Audio Speech and Language Processing*, vol. 20, no. 3, pp. 968–981, 2012.

[16] J. Cabral, S. Renalds, K. Richmond, and J. Yamagishi, "Towards an improved modeling of the glottal source in statistical parametric speech synthesis," in *Sixth ISCA Workshop on Speech Synthesis*, 2007, pp. 113–118.

[17] G. Fant, "The LF-model revisited. Transformations and frequency domain analysis," *STL-QPSR*, vol. 36, no. 2–3, pp. 119–156, 1995.

[18] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "HMM-based Finnish text-to-speech system utilizing glottal inverse filtering," in *Proc. Interspeech*, 2008, pp. 1881–1884.

[19] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. on Audio Speech and Language Processing*, vol. 19, no. 1, pp. 153–165, 2011.

[20] P. Lanchantin, G. Degottex, and X. Rodet, "A HMM-based speech synthesis system using a new glottal source and vocal-tract separation method," *ICASSP*, pp. 4630–4633, 2010.

[21] T. Raitio, A. Suni, H. Pulakka, and M. Vainioand P. Alku, "Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis," *ICASSP*, pp. 4564–4567, 2011.

[22] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3–4, pp. 187–207, 1999.

[23] H. Kawahara, Jo Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *2nd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, 2001.

[24] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Comparing glottal-flow-excited statistical parametric speech synthesis methods," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, 2013, pp. 7830–7834.

[25] T. Raitio, J. Kane, T. Drugman, and C. Gobl, "HMM-based synthesis of creaky voice," in *Proc. Interspeech*, 2013, pp. 2316–2320.

[26] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, 2001.

[27] Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," *PhD thesis, Ecole Nationale Superieure des Telecommunications*, 1996.

[28] Y. Pantazis and Y. Stylianou, "Improving the modeling of the noise part in the harmonic plus noise model of speech," *ICASSP*, pp. 4609–4612, 2008.

[29] J. Cabral and J. Carson-Berndsen, "Towards a better representation of the envelope modulation of aspiration noise," *Proc. NOLISP*, pp. 67–74, 2013.

[30] [Online], "Speech signal processing toolkit (SPTK) v. 3.6," 2013.

[31] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis – A unified approach to speech spectral estimation," in *The 3rd International Conference on Spoken Language Processing (ICSLP)*, 1994, pp. 18–22.

[32] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: a quantitative review," *IEEE Trans. on Audio Speech and Language Processing*, vol. 20, no. 3, pp. 994–1006, 2012.

[33] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Sixth ISCA Workshop on Speech Synthesis*, 2007, pp. 294–299.

[34] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.

[35] T. Kobayashi, S. Imai, and T. Fukuda, "Mel generalized-log spectrum approximation (MGLSA) filter," *Journal of IEICE*, vol. J68-A, no. 6, pp. 610–611, 1985.

[36] [Online], "CMU ARCTIC," 2013, http://festvox.org/cmu arctic/.

## 6.5 Emotion extrapolation

# Extrapolating Acoustic Emotional Patterns to New Speakers in HMM-based Speech Synthesis

Roberto Barra-Chicote, Juan Manuel Montero, *Member, IEEE,* Javier Macias-Guarasa, *Member, IEEE,* Junichi Yamagishi, *Member, IEEE,* Simon King *Senior Member, IEEE,*

**Abstract**

We propose a new method for the extrapolation of emotional acoustic patterns in order to incorporate emotional content into new or previously neutral synthetic voices. The method is demonstrated using acoustic emotion models of four emotions (*anger*, *surprise*, *sadness* and *fear*) which have been trained on emotional female speech data and extrapolated to a new synthetic neutral female voice.

Our analysis shows that the emotional patterns are partially extrapolated to the *target* speaker without losing the *target* speaker identity. The strength of the emotion extrapolation can be successfully varied using an extrapolation factor. However, the strength of the extrapolation has a negative impact on the resulting speech quality, especially when extrapolating the spectral component, which plays an important role in the realisation of *anger*. We propose a new metric (*Emotional Extrapolation Performance (EEP)*) to evaluate the goodness of the extrapolation to a target speaker. Good EEP scores have been obtained in the extrapolation of *fear*, *sadness* and *anger*. However, the acoustic emotional patterns of *surprise* can not be extrapolated with this method.

**Index Terms**

emotion extrapolation, emotional speech synthesis, parametric speech synthesis

## I. Introduction

In the context of acoustic emotional speech synthesis, the requirement to record professional actors who know how to portray the emotions that are to be synthesised, introduces an additional cost in the building of new emotional

R. Barra-Chicote* and Juan Manuel Montero are with Grupo de Tecnologia del Habla, Universidad Politecnica de Madrid, ETSI Telecomunicación, Ciudad Universitaria s/n, 28040 Madrid, Spain. TEL: +34-915-495-700, ext 4254 FAX: +34-913-367-323 E-mail: barra@die.upm.es juancho@die.upm.es. *Corresponding author.

Javier Macias-Guarasa is with Department of Electronics, University of Alcala, Ctra. de Madrid-Barcelona, Km. 33,600, 28805-Alcalá de Henares (Madrid), Spain.

J. Yamagishi and S. King, are with the Centre for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, EH8 9AB, United Kingdom. TEL: +44-131-650-4434 FAX: +44-131-650-6626 E-mail: jyamagis@inf.ed.ac.uk and Simon.King@ed.ac.uk.

The work leading to these results has received funding from the European Union under grant agreement number 287678. RB and JMM are supported by project INAPRA (DPI2010-21247-C02-02)

synthetic voices.

Therefore, it would be useful to build emotional speech models which learn the emotional patterns of a specific speaker and then extrapolate those patterns to conventional neutral voices. The extrapolation of acoustic emotional patterns to different speakers would allow the improvement of the expressiveness of synthetic voices built with existing corpora and more easily and cheaply produce text-to-speech (TTS) suitable for applications like storytelling, toys, virtual agents, etc.

Using the unit selection TTS method [1]–[5], high quality synthetic speech can be produced [6], especially for normal neutral reading styles. Because this method tries to minimise signal processing, highly natural synthetic speech can only be synthesised if the appropriate units are in the inventory. The minimisation of signal processing procedures is particularly important, since it not only degrades speech quality [7], [8], but also impairs emotion identification, emotional strength and speaker similarity [9].

Another inconvenience of unit selection speech synthesis is that the resulting voices are fixed and significant effort is needed to create multiple emotions and speakers. Some researchers have suggested to use rules to incorporate prosodic or phonological strategies into unit selection [10]–[12], found from small or blended emotional speech corpora, to modify the target F0 and duration contours [13]. However, the design of an appropriate target cost function is far from easy because the relationship between the components of the target cost and listeners' perceptions is unclear [14].

The other state-of-the-art TTS method is statistical parametric speech synthesis, which has notable advantages over unit selection in this area: since all acoustic parameters are modelled within a single framework, it is straightforward to transform or modify the speaking style or emotion by using interpolation of Hidden Markov Models (HMM) [15], multiple regression of emotion vectors [16] and/or HMM adaptation techniques [17]. However, the main drawback of statistical parametric speech synthesis is that the spectrum and prosody generated from HMMs tend to be over-smoothed and lacking the richness of detail present in natural spectral and prosodic patterns, because of the inherent averaging in the statistical approach; these details are crucial for properly conveying emotions. However, since the HMM-based approach requires less data than unit selection and also enables the generation of intermediate or exaggerated emotions, it is still an attractive proposition for modelling acoustic emotional patterns.

HMM adaptation techniques [18], [19] provide a powerful tool to create new synthetic voices with relative little data of the target speaker. These techniques have been successfully applied to the synthesis of emotional speech [17], [20], but they require emotional data from the target speaker, and listeners' identification of the natural emotional data must first be confirmed to ensure the perception of the emotion in the target speaker voice. This usually entails recording professional actors. Large, high quality neutral speech corpora are available for speech synthesis, but it could be very difficult to obtain additional recordings from the same speakers. As an alternative, it would be possible to interpolate emotional speech models [15] of an emotional speaker voice and a target speaker voice, but in this case, the similarity with the target speaker would decrease, as the target speaker would be identified as the emotional source speaker.

One of the main functions of emotions is adaptation [21]. Given a stable situation and a stable state (understood

as a neutral reference state), we can view emotional processes [22] as a deviation from that neutral state, used by the individual to try to adapt to a new situation generated by an incoming stimulus. Evaluation of these stimuli has an *offset* effect in the subjective experience and physiological support, that is previously modulated by an appraisal filter that depend on the previous individual experience and social and culture information [23]. Physiological changes (e.g., changes in voice), depend on these factors too, and this suggests that it is necessary for the listener to know the offset introduced by the speaker (e.g., due to previous experience) to completely identify his or her emotional state.

Based on previous HMM interpolation algorithms [15], we propose a new method for the extrapolation of acoustic emotional patterns to new target speakers (for whom we have no emotional speech training data), where the acoustic emotional patterns are learnt as deviations of emotional speech models of a source speaker from his or her neutral model.

## II. CORPORA

In this work, we used two corpora:

- The *Spanish Expressive Voices* (SEV) corpus [24], used to build emotional voices of a source speaker, from which acoustic emotional patterns will be learnt.
- The UPC_ESMA [25] corpus, used to build the neutral voice of a target speaker, to which we will apply the previously learnt emotional patterns.

The SEV corpus comprises speech and video recordings of an actor and an actress speaking in a neutral style and simulating six basic emotions: *happiness*, *sadness*, *anger*, *surprise*, *fear* and *disgust*. SEV presents a relatively large size for a corpus of this type (more than 100 minutes of speech per emotion). In this work only the speech data of the actress have been used (almost one hour per emotion). The SEV corpus covers speech data in several genres such as isolated word pronunciations, short and long sentences selected from the SES corpus [26], narrative texts, a political speech, short and long interviews, question answering situations, and short dialogues. The texts of all utterances are emotionally neutral. The database has been automatically labelled. The female speech data has been validated through perceptual tests, achieving an *Emotion Identification Rate* (EIR) as high as 90% [24]. Emotional voices based on this corpus using statistical parametric speech synthesis and unit selection synthesis have been successfully evaluated [9].

The UPC_ESMA corpus comprises a set of 776 recordings of a professional actress in neutral style. It covers short sentences (almost 30 minutes of speech), medium length paragraphs (almost 30 minutes of speech) and large literary paragraphs (almost 45 minutes of speech), all phonetically balanced. Voices based on this corpus using statistical parametric speech synthesis and unit selection synthesis have also been successfully evaluated [27].

## III. BUILDING VOICES

The synthetic voices were built using a statistical parametric speech synthesis technique, using the HTS Toolkit [28] adapted for Spanish [9]. Our Spanish system, using the UPC_ESMA corpus, exhibited very good performance in

a Spanish speech synthesis competition [27], [29]. Each emotional voice was built from scratch using speech data only of the target emotion.

The HMM-based speech synthesis system comprises three components: speech analysis, HMM training, and speech generation.

- In the speech analysis part, three kinds of parameters for the STRAIGHT [30] mel-cepstral vocoder with mixed excitation (the mel-cepstrum, $\log F0$ and a set of aperiodicity measures) are extracted as feature vectors to be modelled by the HMMs.
- In the HMM training part, context-dependent multi-stream left-to-right Multi-Space Distribution Hidden Semi-Markov Models – MSD-HSMMs [31] – are trained for each emotion using a maximum likelihood criterion.
- In the speech generation part, acoustic feature parameters are generated from the MSD-HSMMs using a parameter generation algorithm that considers the Global Variance (GV) of a trajectory to be generated [32]. Finally, an excitation signal is generated using mixed excitation (pulse plus band-filtered noise components) and Pitch-Synchronous Overlap and Add (PSOLA) [33]. This signal is used to excite a Mel-Logarithmic Spectrum Approximation (MLSA) filter corresponding to the STRAIGHT mel-cepstral coefficients, generating the synthetic speech waveform.

## IV. Emotion Modelling and Extrapolation

In our method, we start by using neutral and emotional speech from a given source speaker to learn the relevant emotional patterns. In a second step, these emotional patterns will be applied to the neutral speech model for a target speaker, so that no emotional speech data from this target speaker will be required.

In our case, we use the female speaker from the SEV corpus as the *source* speaker (*src*), and the female speaker from the UPC_ESMA corpus as the *target* speaker (*tgt*).

The emotion models, which will be extrapolated to the neutral voice of the *target* speaker, will be estimated from the neutral and emotional recordings of the *source* speaker.

The training material for the source speaker was limited to those emotions that were better identified in earlier tests, having at least a 70% emotion identification rate relative to natural speech [9]. These emotions were: *anger*, *surprise*, *sadness*, and *fear*.

Figure 1 summarizes the proposed emotion extrapolation method.

### A. Emotion Model Definition and the Proposed Method

HMMs for each emotion may have different clustering tree structures and therefore it is not straightforward to extrapolate at the model level. Therefore, the extrapolation of HMMs is done on-line at synthesis time, using *interpolation between observations*, as in [15] which is the simplest interpolation method described in [34].

Let $E_1, \cdots, E_N$ represent the $N$ emotions and let $E_0$ represent neutral speech. First, we convert a given text into a context-dependent phoneme label sequence. Then, by consulting the context clustering decision trees built for each state of each feature in the HMMs for neutral and each emotion of a source speaker, the context-dependent

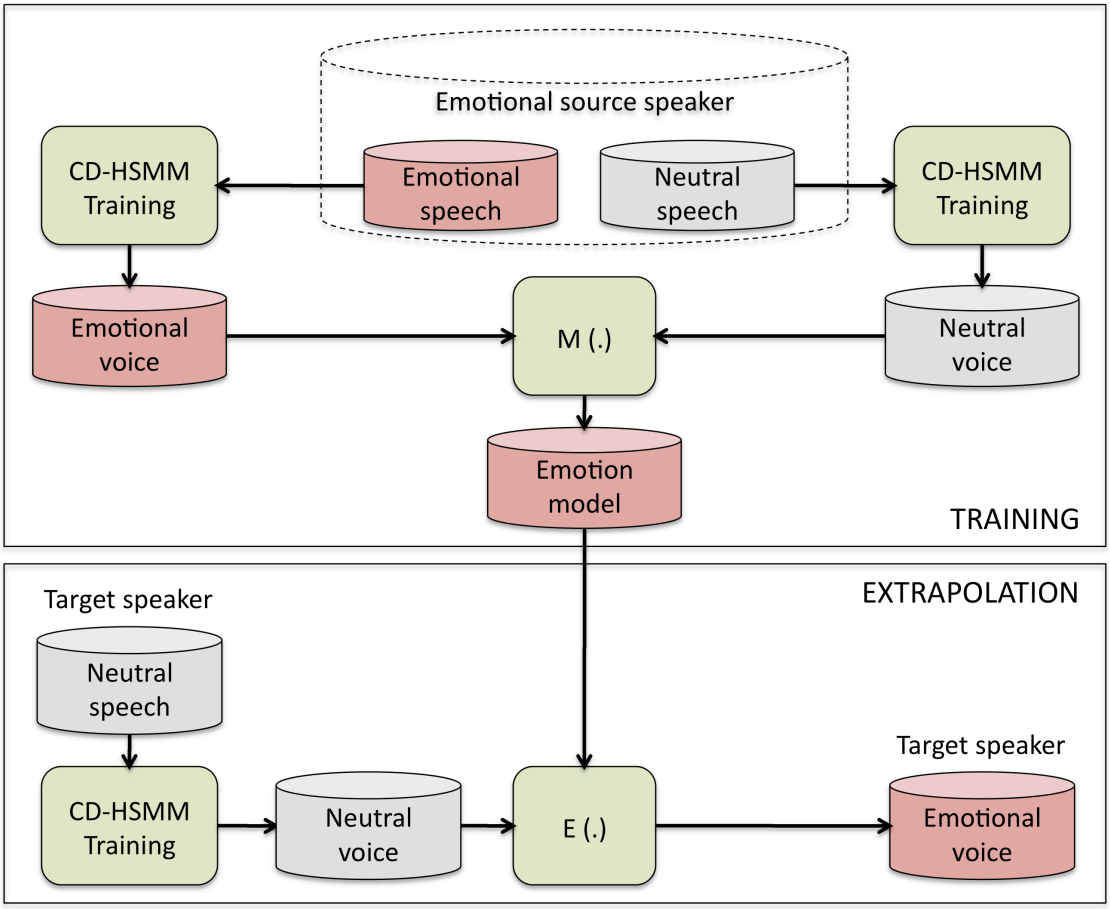



Fig. 1. Graphical representation of the emotion extrapolation method.

phoneme label sequences are converted into $N+1$ sentence-sized HMMs, one for each emotion and one for natural speech, $\lambda_0 \cdots \lambda_N$, having different state sequences. However note that they have the same total number of states $I$. Each state contains several Gaussian pdfs for each of the acoustic features and a single Gaussian pdf for duration. The Gaussian pdf for state $i$ in the sentence-sized HMM $\lambda_n$ for emotion $n$ is characterized by a mean vector $\boldsymbol{\mu}_{i_n}$ and a covariance matrix $\boldsymbol{\Sigma}_{i_n}$. The dimension of the mean vector may vary depending on the acoustic features. Then, we calculate differences and scaling of the mean and covariance $(\Delta\boldsymbol{\mu}_{i_n}, \Delta\boldsymbol{\Sigma}_{i_n})$ for each state $i$ between neutral $E_0$ and each emotion $E_n$.

$$\Delta\boldsymbol{\mu}_{i_n} = \boldsymbol{\mu}_{i_n} - \boldsymbol{\mu}_{i_0}, \tag{1}$$

$$\Delta\boldsymbol{\Sigma}_{i_n} = \boldsymbol{\Sigma}_{i_n}\boldsymbol{\Sigma}_{i_0}^{-1} \tag{2}$$

The key idea of the proposed method is simple – we assume that these differences above are relatively speaker-independent and thus may be applied to different speakers.

In a similar way to the source speaker, by consulting the context clustering decision trees built for each state of

each feature in the HMMs for the neutral speaking style $\widehat{E}_0$ of a target speaker, the context-dependent phoneme label sequences are converted into a sentence HMM $\widehat{\lambda}_0$. This sentence HMM also has the same total number of states $I$. Let the mean vector and covariance matrix of a Gaussian pdf for state $i$ in the $\widehat{\lambda}_0$ be $\widehat{\boldsymbol{\mu}}_{i_0}$ and $\widehat{\boldsymbol{\Sigma}}_{i_0}$, respectively. Using the differences above, we define a new mean vector and a new covariance matrix for that Gaussian pdf for the target emotion $E_n$ for the target speaker, denoted by $\widehat{\boldsymbol{\mu}}_{i_n}, \widehat{\boldsymbol{\Sigma}}_{i_n}$, as follows:

$$\widehat{\boldsymbol{\mu}}_{i_n} \equiv \widehat{\boldsymbol{\mu}}_{i_0} + k\Delta\boldsymbol{\mu}_{i_n}, \tag{3}$$

$$\widehat{\boldsymbol{\Sigma}}_{i_n} \equiv \widehat{\boldsymbol{\Sigma}}_{i_0} \cdot k^2\Delta\boldsymbol{\Sigma}_{i_n} \tag{4}$$

where $k$ is an extrapolation factor between neutral $\widehat{E}_0$ and the target emotion $\widehat{E}_n$. A graphical representation of the proposed method is shown in Figure 2. Ellipses represent the gaussian distributions for a hypothetical 2D speech component. Note that this can be applied to the Gaussians for all the acoustic features, including spectrum, $\log F0$, and duration in the same way. It is also possible to apply the same concept to the weights representing voice/unvoiced ratios.

## V. Analysis of the Emotion Extrapolation

The emotion extrapolation function has the potential to change:

- the perceived emotion in the target speaker voice;
- the perceived speech quality (SQ);
- the perceived similarity with the target speaker.

Prior to an exhaustive perceptual evaluation of the emotion extrapolation algorithm proposed in Section IV, we first performed an initial analysis of the feasibility of the emotion extrapolation method over each speech component, over intonation and rhythm jointly and over the whole speech model.

A small set of five sentences, different to the final evaluation set, was selected for this analysis. The emotion extrapolation was evaluated using utterances synthesised with:

- the emotionally transformed model of the target speaker,
- the neutral synthetic voice of the target speaker,
- the neutral synthetic voice of the source speaker,
- and the emotional synthetic voice of the source speaker.

### *A. Emotion Extrapolation of the Spectral Component*

First, we analysed the emotion extrapolation of only the spectral component. For this initial study, we decided to set the extrapolation factor $k_{spc}$ equal to 1, so that the same acoustic deviation between the emotional speech and the neutral speech of the source speaker is applied to the neutral speech of the target speaker (in V-D and VII we will address the effect of the variation of the extrapolation factor).

The perception of the resulting transformation can be summarised as:

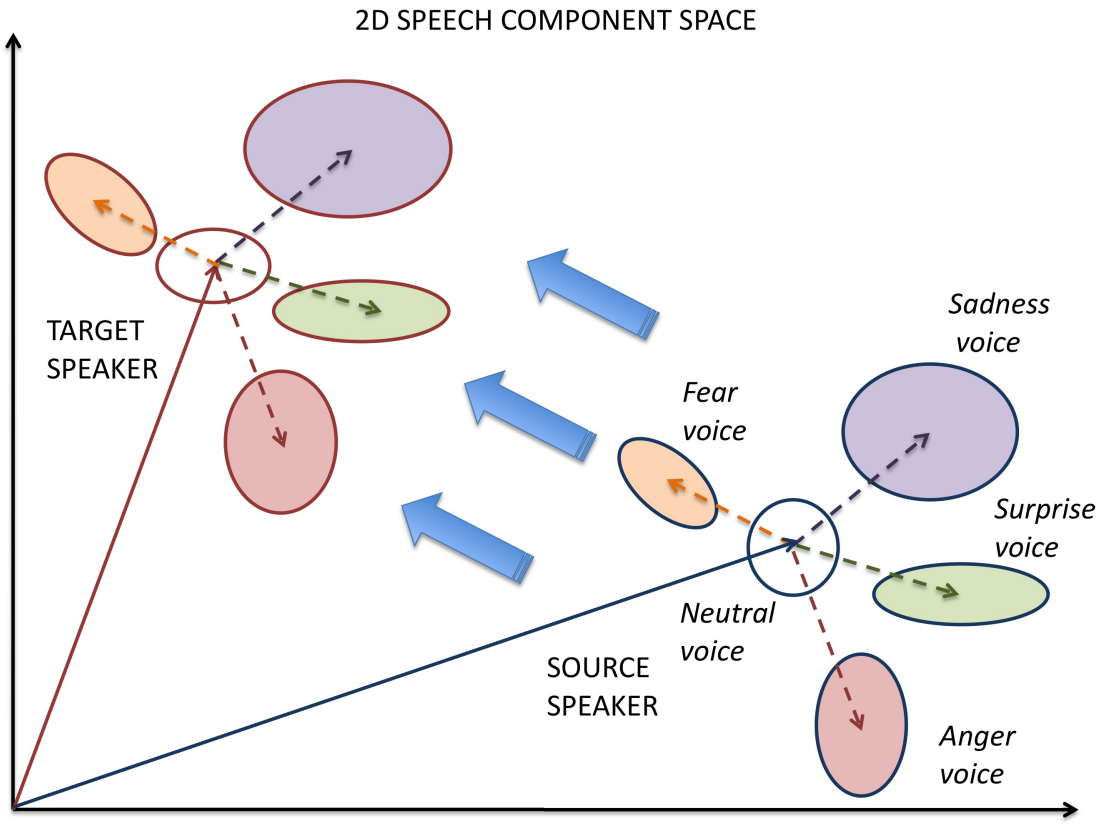



Fig. 2. Graphical example of extrapolation of the emotional space of a source speaker to a target speaker using an extrapolation factor $k$ equal to 1.

- Speech fragments that were transformed with reasonable quality (without instabilities or distortion), sound very close to the neutral speech.

  This might be expected, especially for "prosodic" emotions, like fear and surprise (as we deduced in [9]). However:

  - Some spectral coloration in the transformed spectra for fear could be perceived, a by-product of the extremely high pitch of this emotion.
  - Surprise extrapolation of the spectrum sounds clearly neutral. No emotional differences compared to the neutral target speech could be perceived.

  The extrapolation of anger, which was found to behave as a segmental emotion in [9], is able to convey the perceived effect of a "closed mouth" in the angry voice of the *source* speaker. However, this effect appears to be smoothed in the transformed voice.

  The extrapolation of sadness, an emotion with observed specific segmental and supra-segmental patterns [9], is

able to convey the "languor" of sadness. However the extrapolation of this emotion for the spectral component has consequences for the final perceived SQ.

- SQ is partially degraded for all emotions. Initially, we observed that the main degradation was due to the instability of the extrapolation for some phonemes. In order to avoid such instabilities in the transformation, we conducted an experiment where only the central state (of five states per phoneme) was transformed, arguing that transitional states between phones are less stable and stationary than the central states. No instabilities were found in this approach, but the resulting voice was again perceived as mostly neutral.

  Then, we analysed in which contexts the transformation was not stable, avoiding the transformation for those. We observed that most of the main instabilities occurred in:

  - Trills and taps including /R/ (like in "verdad") and /r/ (like in "quiere").
  - Transitional states from trills, taps, plosives (/p/, /t/, /k/) and aproximants (/b/, /d/, /g/) to initial transitional states of vowels.
  - Transitional states from vowels to initial transitional states of trills and taps (only for /R/ and /r/) and fricatives.

  Based on these initial findings, we conducted an additional experiment in which the transitional states of those problematic contexts were not transformed. Most of the instabilities disappeared, improving the overall SQ without apparently losing emotional content. However, in some cases it appeared necessary to avoid the transformation in the central states.

  Finally, we conducted an experiment in which only $\mu_{spc_i}$ was transformed. In this case, the aforementioned instabilities occurred, thus refuting the hypothesis that the transformation of ${\Sigma_{scp_i}}^2$ could be responsible for those instabilities.

### *B. Emotion Extrapolation of the Duration Component*

We did not observe degradation in the SQ when performing this transformation. Only a few artefacts appeared, due to excessively long unvoiced sounds, as a consequence of not transforming the voiced/unvoiced ratio, suggesting that prosodic components should be jointly extrapolated.

### *C. Emotion Extrapolation for F0 Component*

The emotion extrapolation of $\log F0$ caused minimal degradation in SQ, compared with the degradation introduced by the extrapolation of the spectral component.

We found that the extrapolation of the voiced/unvoiced ratio was not stable, so we conducted an informal experiment in which the voiced/unvoiced ratio of the synthesised speech of the *transformed* speaker was just copied from the emotional voice of the *source* speaker. The results showed an increase in SQ.

In addition, for the extrapolation of the $\log F0$ emotional component it has also been necessary to consider the voicing of each acoustic context in the source speaker voices and the target speaker voice. The extrapolation function can not be applied straightforwardly when an acoustic context of $\hat{\lambda}_0$ (sentence-sized HMM for neutral emotion of

the target speaker) or $\lambda_n$ (sentence-sized HMM for emotion $n$ of the source speaker) is an unvoiced context. In these cases, average parameters (estimated using the parameters of the previous voiced context dependent states for the considered utterance) were used instead of the parameters of the unvoiced acoustic context.

However, the emotional content is only partially perceived when the extrapolation is done only over $\log F0$ and voicing, suggesting that it is necessary to consider the extrapolation over the other speech components.

We also conducted an informal experiment where the emotion extrapolation was done only over the aperiodicity bands. This extrapolation was perceived almost as neutral with no SQ degradation.

Finally, the emotion extrapolation strategy was applied to duration, $\log F0$ (voicing was copied from the emotional voice of the *source* speaker) and the aperiodicity bands. The results showed that the SQ is similar to the SQ perceived when the extrapolation was applied separately, and that the emotion identification rate increased. However, the spectral "colour" of fear, the "closed mouth" effect of angry voice and the "languor" and "sob" of sadness are not perceived.

The absence of these patterns suggests that the emotion should also be extrapolated over the spectral component. Only for surprise (a mainly prosodic emotion [9]) we might consider not transforming the spectrum, trying to maintain similarity with the target speaker as much as possible.

### *D. Complete Extrapolation of Emotions*

Finally, we applied the emotion extrapolation function to all the speech components of the target speaker neutral voice. As expected, emotional patterns of fear, anger and sadness were perceived. Those patterns were perceived with a higher strength, compared to extrapolation over only one speech component.

The extrapolation factors of each speech component can modulate the strength of the extrapolated emotion. A value of $k_{[cmp]}$ less than $1.0$ would lead to a partial extrapolation of the emotion model for that specific speech component to the target speaker. A value of $k_{[cmp]}$ equal to $1.0$ would lead to the equivalent emotional deviation from the neutral voice of the *source* speaker, and a value of $k_{[cmp]}$ higher than $1.0$ will lead to an over-extrapolation of the emotion model. However, the impact of the extrapolation factor on SQ, the identification of the emotion and its emotional strength and the similarity to the target speaker would have to be measured by perceptual evaluation.

## VI. Design of the Perceptual Evaluation

### *A. Evaluation Metrics*

When emotional synthetic speech is evaluated, the key factors to consider are not only related to its overall quality (or naturalness), but also to the accuracy in the identification of the intended emotion and its emotional strength [9]. For the extrapolation of emotional speech patterns of one speaker to another speaker, identification of the target speaker should be evaluated as well as the previous metrics. In applications where the *target* speaker is not known to the users, the main goal would be to synthesise emotional speech as *another* speaker, trying to avoid similarity to the *source* speaker.

When we are also evaluating the extrapolation of emotional speech patterns from one speaker to another, the identification of the target speaker should be also evaluated. Furthermore, in applications where the *target* speaker is not known to the users, the main goal would be to synthesise emotional speech as produced by *other* speaker, trying to avoid the similarity with the *source* speaker.

In the perceptual evaluation carried out in this work, and taking into account all these considerations, we evaluated the following aspects of emotional synthetic speech:

1) Speech Quality (SQ): Listeners were required to evaluate the overall quality of the given emotional synthetic speech using a 5-point scale where 1 was labelled as "muy mala" (very bad), 2 as "mala" (bad), 3 as "aceptable" (acceptable), 4 as "buena" (good) and 5 as "muy buena" (very good).
2) Emotional Strength (ES): Listeners were required to assess the emotional strength of the given synthetic speech using a continuous slider. The endpoints of the slider were labelled "very weak" and "very strong".
3) Emotion Identification Rate (EIR): Listeners were required to identify the intended emotion in the given synthetic speech from a limited set of emotional categories: *anger*, *surprise*, *sadness*, *fear*, *neutral* or *another*.
4) Speaker Identification Rate (SIR): The listeners were required to assess the similarity to the *source* speaker, the *target* speaker or *neither* of both speakers. A continuous slider was presented to the listeners. The endpoints of the slider were labelled as "Totalmente locutor A" (totally speaker A) and "Totalmente locutor B" (totally speaker B). The middle of the slider was ticked and labelled with "ninguno" (*neither* option). The identification of one speaker is considered when the slider is moved to the side of that speaker.

### *B. Experimental Design*

Our goal was to obtain insight into how emotional speech patterns of a *source* speaker can be extrapolated to a *target* speaker without losing similarity to the voice of the *target* speaker. We also wished to find out how the method affects the SQ of the synthetic speech and the perception of the ES. We evaluated neutral natural speech and synthetic speech of *source* speaker and *target* speaker in order to establish the intrinsic characteristics of each synthetic voice. We also evaluated emotional synthetic speech of the *source* speaker, in order to establish upper bounds on speech quality, emotion identification rate, emotional strength and speaker identification rate.

In order to evaluate the scope of the emotion extrapolation method, we defined four systems, each one based on a different extrapolation factor $k = \{0.5, 0.75, 1.0, 1.25\}$ (same $k$ applied to the emotion extrapolation over each speech component).

Since different effects on the evaluation metrics were observed for each emotion during the development of the extrapolation method (section V), an *adhoc* system configuration was defined for each emotion, depending on its nature [9]. *Adhoc* configuration for the extrapolation of each emotion is presented in Table I. As mentioned in section V-A, some instabilities can occur when the extrapolation method is applied in certain ways, especially for extrapolation factors $k$ higher than $1.0$. We observed that a considerable amount of those instabilities can be eliminated imply by copying the intensity contour (the zeroth cepstral coefficient) of the emotional speech of the source speaker. For *anger*, due to its segmental nature, its spectral component was principally extrapolated using

TABLE I

DEFINITION OF THE *adhoc* EXTRAPOLATION CONFIGURATION FOR EACH EMOTION.

| SPEECH COMPONENT | EXTRAPOLATED EMOTION | | | |
|---|---|---|---|---|
| | **A**nger | **Su**rprise | **S**adness | **F**ear |
| Intensity (Cepstrum 0) | copied | copied | copied | copied |
| Spectra | 1.25 | 0.0 | 0.75 | 0.5 |
| F0 (& Aperiodicity) | 1.0 | 1.0 | 1.0 | 1.0 |
| Duration | 1.0 | 1.0 | 1.0 | 1.0 |

$k_{[cmp]}$ set to $1.25$; however we do not detect any degradation in SQ and SIR when this emotion is extrapolated over prosodic components, so $k_{logF0}$ and $k_{dur}$ were set to $1.0$. For prosodic emotions, $k_{logF0}$ and $k_{dur}$ higher than $1.0$ results in a reduction of the emotion's naturalness and SQ, so we also set them to $1.0$. Because of the clear prosodic nature of *surprise* and because no difference was observed in the emotion perception, no emotion extrapolation was done over its spectral component. In case of *fear* (clearly prosodic), a certain "spectral colour" in the transformed spectra was perceived (an artefact of the extremely high pitch of this emotion), so $k_{[cmp]}$ was set to $0.5$, trying not to extrapolate excessively, in order to avoid a reduction in SQ rates. Finally, for *sadness* (an emotion with a mixed nature) $k_{[cmp]}$ was set to $0.75$, trying also to prevent low SQ rates.

### *C. Perceptual Tests and Subjects*

The ten systems we built and evaluated are described in Table II. In our experiments we define a *scheme* as the combination of a synthesis system and a given emotion. A total of 28 *schemes* had to be evaluated (6 systems (from E to J in Table II) combined with four emotions, plus neutral speech systems of *source* speaker and *target* speaker (systems C and D, respectively) and neutral natural examples of both speakers (system A and B, respectively).

The experimental design was based on a balanced latin-square matrix, similar to the experimental design used in the Blizzard Challenge [35]. Each *scheme* generated a set of speech for the 28 sentences that were not included in the voice training sentences. They were medium length sentences, between 6 to 11 words and with an average length of 8 words. The content of the test sentences was emotionally neutral to allow listeners to focus only on acoustic cues. The latin-square design allows an evaluation of all schemes and all synthesised sentences whilst controlling ordering effects by ensuring that each group of listeners hears the stimuli in a different order.

The perceptual evaluation was divided into two sections:

- In the first part, speech quality, intended emotion and emotional strength were all evaluated in the same trial (and in this same "visual" order in the web page). The evaluation of this part was conducted via a web browser interface. Note that listeners were explicitly required to make each judgement independently from the others. Before making a decision, each utterance could be played as many times as the listener wished, but they could never go back to re-evaluate previous utterances.

TABLE II

DEFINITION AND NAMES OF SPEECH SYNTHESIS SYSTEMS USED FOR THE PERCEPTUAL EVALUATION.

| SYSTEM | VOICE | SPEAKER | EXTRAP. FACTOR ($k$) |
|---|---|---|---|
| A | NEUTRAL | *source* | – |
| B | NATURAL SPEECH | *target* | – |
| C | NEUTRAL | *source* | – |
| D | SYNTHETIC SPEECH | *target* | – |
| E | | *source* | – |
| F | | *transformed* | 0.5 |
| G | EMOTIONAL | *transformed* | 0.75 |
| H | SYNTHETIC SPEECH | *transformed* | 1.0 |
| I | | *transformed* | 1.25 |
| J | | *transformed* | $adhoc$ |

- In the second part, speaker identification was evaluated via a web browser interface. Four reference files of each speaker were presented to the listeners, so they could hear the files as many times as needed, before making a decision, but they could never go back to re-evaluate previous utterances.

Twenty eight listeners, having a similar socio-cultural profile, participated in the evaluation, which was carried out individually in a single session per listener. All listeners were from the Madrid area and were between twenty and forty years old, and none of them had a speech-related research background nor had they previously heard any of the SEV speech recordings. The evaluation was conducted in a quiet environment using headphones.

The authors decided to avoid long sessions, thus limiting to 56 the number of stimuli to be presented to each listener (28 stimuli evaluated two times, between the first and the second part of the evaluation), so that the average length of each session was 31 minutes.

The evaluation using 28 listeners provided 112 evaluation responses (i.e. 28 per emotion) for each system, except for systems A to D in Table II. Systems A to D are for the neutral emotion and have 28 evaluation responses per system. One evaluation response includes the listener's rating for speech quality, the identified emotion, the emotional strength and the speaker similarity, for a single stimuli.

## VII. RESULTS

### A. Analysis of Neutral Voices

First, we analysed the perceptual results for neutral speech (systems A to D from Table II). Results are shown in Table III. As expected, the highest SQ scores are for natural speech (systems A and B).

The neutral synthetic voices of both speakers (systems C and D) obtained similar SQ scores. However, neutral

TABLE III
EVALUATION RESULTS OF NEUTRAL VOICES.

| SYSTEM | SQ | EIR | SIR | | |
|---|---|---|---|---|---|
| | | | source | neither | target |
| A | 5 (4.5) | 76% | 83% | | 17% |
| B | 4 (4.0) | 72% | 10% | 7% | 83% |
| C | 3 (3.2) | 86% | 69% | | 31% |
| D | 3 (3.4) | 69% | 7% | | 93% |

synthetic voice of the source speaker (system C) was perceived as more robotic and buzzy than the neutral synthetic voice of the target speaker (system D), because the first one was built with half of the data.

This might be associated by the listeners with a neutral speaking style: the synthetic voice of the source speaker (86% EIR) outperformed the neutral natural speech (76% EIR). This effect also affected the SIR, the similarity of the synthetic voice with the source speaker decreased from a 83% for natural speech to 69% for synthetic speech.

SQ scores of the neutral synthetic voice of the target speaker are similar to the SQ scores of natural speech. Listeners clearly perceived the speaker identity of the target speaker from its neutral synthetic voice (93%), even better than from natural speech. Artefacts introduced by the synthesis process may have acoustically separated the speakers' voices.

### *B. Speech Quality (SQ)*

Figure 3 presents a boxplot showing SQ results for the emotional speech synthesised using different extrapolation factors. In the boxplot, the median is represented by a solid bar across a box showing the quartiles with whiskers extending to 1.5 times the inter-quartile range and outliers beyond this being represented as circles. The mean is represented by a cross. Significant differences between extrapolation factors are shown in a table bottom in the same figure. Emotional speech of the source speaker obtains the higher SQ results, as expected. The extrapolation factor produces certain degradation in SQ. SQ decreases whichever major is the extrapolation factor. Only $k = 0.5$ extrapolation factor obtains no statistically significant different SQ values when compared to the SQ of the source speaker. The $adhoc$ extrapolation scheme slightly reduced this SQ degradation.

We analysed the SQ scores obtained with each extrapolation factor $k$ for each emotion:

- *Surprise* was the emotion that obtained a higher SQ reduction relative to the SQ of the source speaker. For all the emotions, the SQ scores between the emotional speech of the source speaker and $k = 0.5$ extrapolation factor are similar, except for *surprise*.
- *Anger* extrapolation over the spectral component using $k_{spec} = 1.25$ (used in the $k = 1.25$ and $k = adhoc$ extrapolation configurations) introduces artefacts that reduced SQ scores.

### C. Emotional Strength (ES)

The Emotional Strength (ES) score, elicited from the listeners using a slider, was treated as a continuous variable without categorical information. Since every listener may use his or her own scale, we normalised the scores on a per listener basis.

The boxplot presented in Figure 4 shows the normalised ES scores. Pairwise $t$-tests with Bonferroni step-down correction were conducted to determine whether there are significant differences between the normalised ES scores of each extrapolation scheme. The table in the lower part of the figure shows the significant differences between extrapolation schemes and emotional synthetic speech of the *source* speaker for $p = 0.05$.

As was expected, the ES scores of the emotional synthetic speech of the *source* speaker are higher than the ES scores of all the extrapolation schemes. Contrary to the SQ scores, we obtained higher ES scores whichever major is the extrapolation factor $k$. ES scores for the *adhoc* scheme have no significant differences with the schemes using extrapolation $k$ values between 0.75 to 1.25. ES scores considering each emotion separately present these tendencies.

### D. Emotion Identification Rates (EIR)

The EIR results are shown in Figure 5. Emotional synthetic speech of the *source* speaker is clearly identified (EIRs is over 50%). Different EIRs (depending on the emotion) were obtained:

- Using extrapolation schemes with $k$ higher than 0.5, *sadness* and *fear* synthetic speech of the *transformed* speaker are better identified (69% using $k$ equal to 1.0 for both emotions) than the emotional synthetic speech of the *source* speaker (62% and 48% respectively).
- Low EIRs confirm that the extrapolation method was not able to extrapolate the acoustic emotional patterns of *surprise* to the *target* speaker.
- *Anger* speech of the *transformed* speaker is less accurately identified, whatever the value of $k$. The constraints considered in the emotion extrapolation over the spectral component (keeping neutral the context described in section V-A), and the sensitivity to instabilities in the extrapolation over this component, affected the identification of *anger*. However, the *k=0.5* and *adhoc* schemes are reasonably well identified (47% and 53% relative to the *source* speaker).

### E. Speaker Identification Rates (SIR)

The SIR results are shown in Figure 6. The speaker identity of the emotional synthetic speech of the *source* speaker (57%) is better identified than the identity of the emotional synthetic speech of the *transformed* speaker. The emotional patterns that affect the acoustic parameters of the voice of the source speaker make it difficult to identify the *source* speaker, as its emotional speech is also identified as the *target* speaker (28%) or even as *neither* (16%).

This is also confirmed, by the decrease of the *source* and *target* speaker's SIR, whatever the value of the extrapolation factor $k$. In addition, the similarity with other speakers (*neither option*) increases for any value of the

extrapolation factor. Only at an extrapolation scheme $k$ equal to 1.0, could listeners not differentiate the speaker identity between the *source* speaker and *target* speaker.

The *adhoc* extrapolation scheme obtains a good compromise between a high *target* SIR (40%), a lower *source* SIR (23%) and a 37% identified as other speaker (*neither* option).

We analysed the SIR for each emotion and extrapolation scheme separately: results are shown in Figure 7. The speaker identification for each emotional voice of the *source* speaker was a difficult task, except for surprise (83% *source* SIR). The *anger* voice of the *source* speaker was even identified as the *target* speaker (48%). The *source* SIR for emotions with a clear supra-segmental nature (*fear* and *surprise*) is higher than the *target* SIR.

However, for all the emotions and extrapolation schemes used (*transformed* voices), the *source* SIR is lower than the *source* SIR for the emotional synthetic voices of the *source* speaker. The proposed method has extrapolated the emotional patterns without extrapolating the identity of the *source* speaker, and the transformed speaker is notably identified as the *target* speaker or at least as another speaker different from the *source* speaker.

### *F. Emotion Extrapolation Performance Measure*

In order to further evaluate the proposed method, we now define an emotion extrapolation performance measure (EEP-measure):

$$EEP \;=\; 3 \cdot \frac{\hat{SQ} \cdot EIR \cdot (SIR_{st} + SIR_n)}{\hat{SQ} + EIR + (SIR_{st} + SIR_n)} \tag{5}$$

where $\hat{SQ}$ is the SQ score normalised to the range 0 to 1, $SIR_{st}$ is the SIR of the corresponding speaker (*source* or *target*) and $SIR_n$ is the SIR for the *neither* option.

Figure 8 shows the EEP value for each emotional voice of the *source* speaker and the *transformed* speaker using every extrapolation scheme. Based on these EEP results, our main conclusions are:

- A poor performance of *surprise* voice for the *transformed* speaker.
- On the contrary, good EEP values in the case of *fear* and *sadness* demonstrate that the proposed method has successfully extrapolated the acoustic emotional patterns of these two emotions to a different speaker.
- The EEP values for the *anger* voice of the *transformed* speaker using the *adhoc* scheme or an extrapolation factor $k$ equal to 0.5, are reasonably good considering how the constraints applied to the extrapolation of the spectral component affected the SQ and EIR of this emotion. However, future further analysis will be required to control instabilities in order to improve the extrapolation performance of *anger*.

## VIII. CONCLUSION

A method for the extrapolation of emotional acoustic patterns has been defined to incorporate emotional content into new or previously-neutral synthetic voices. A perceptual test was conducted, where the speech quality, the emotional strength, emotional identification rates and speaker identity rates were evaluated.

The acoustic emotional models of four emotions (*anger*, *surprise*, *sadness* and *fear*) were trained from an emotional female voice and extrapolated to a new synthetic neutral female voice. The emotional patterns over

each speech component (spectra, $\log F0$, aperiodicity bands and durations) have been considered in the acoustic emotional model.

With the proposed algorithm, acoustic emotional patterns are partially extrapolated to a *target* speaker without losing the *target* speaker identity. The strength of the emotion extrapolation can be modified successfully by varying the extrapolation factor. However, the strength of the extrapolation has negative impact on the resulting speech quality, especially in the extrapolation of the emotional patterns of the spectral component.

We have proposed a new metric – *Emotional Extrapolation Performance* – to evaluate the goodness of the extrapolation to a target speaker. Good EEP values were obtained in the extrapolation of *fear*, *sadness* and *anger*. *Surprise* obtained poor EEP values and will be analysed in further research.

## REFERENCES

[1] A. W. Black and N. Cambpbell, "Optimising selection of units from speech database for concatenative synthesis," in *Proceedings EUROSPEECH-95*, Sep. 1995, pp. 581–584.

[2] A. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proceedings ICASSP-96*, May 1996, pp. 373–376.

[3] R. Donovan and P. Woodland, "A hidden Markov-model-based trainable speech synthesizer," *Computer Speech and Language*, vol. 13, no. 3, pp. 223–241, 1999.

[4] A. Syrdal, C. Wightman, A. Conkie, Y. Stylianou, M. Beutnagel, J. Schroeter, V. Storm, K. Lee, and M. Makashay, "Corpus-based techniques in the AT&T NEXTGEN synthesis system," in *Proceedings ICSLP 2000*, Oct. 2000, pp. 411–416.

[5] R. A. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

[6] V. Karaiskos, S. King, R. A. J. Clark, and C. Mayo, "The Blizzard Challenge 2008," in *Proceedings Blizzard Challenge Workshop 2008*, Brisbane, Australia, September 2008.

[7] D. S. G. Vine and R. Sahandi, "Synthesising of emotional speech by concatenating multiple pitch recorded speech units," in *ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, 2000. [Online]. Available: http://eprints.bournemouth.ac.uk/10936/

[8] D. Tihelka and J. Matoušek, "Revealing the most significant deterioration factors in single candidate synthetic speech," in *Specom 2005, proceedings of 10th International Conference SPEECH and COMPUTER*. Moscow: Moscow State Linguistic University, 2005, pp. 171–174.

[9] R. Barra-Chicote, J. Yamagishi, S. King, J. M. Montero, and J. Macias-Guarasa, "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech," *Speech Communication*, vol. 52, no. 5, pp. 394–404, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/B6V1C-4XY4GDS-1/2/7e701c2305a5ff0713d2c2e83af6e760

[10] W. Hamza, R. Bakis, E. Eide, M. Picheny, and J. Pitrelli, "The IBM expressive speech synthesis system," in *Proc. ICSLP 2004*, 2004.

[11] J. Pitrelli, R. Bakis, E. Eide, R. Fernandez, W. Hamza, and M. Picheny, "The IBM expressive text-to-speech synthesis system for American English," *IEEE Trans. on Speech Audio Process.*, vol. 14, no. 4, pp. 1099–1108, Jul. 2006.

[12] G. Hofer, K. Richmond, and R. Clark, "Informed blending of databases for emotional speech synthesis," in *Proc. Interspeech 2005*, 2005, pp. 501–504.

[13] M. Schröder, "Emotional speech synthesis: a review," in *Proceedings EUROSPEECH 2001*, 2001, pp. 561–564.

[14] V. Strom and S. King, "Investigating Festival's target cost function using perceptual experiments," in *Proceedings Interspeech 2008*, 2008, pp. 1873–1876.

[15] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing,," *IEICE Trans. Inf. & Syst.*, vol. E88-D, no. 11, pp. 2484–2491, Nov. 2005.

[16] T. Nose, J. Yamagishi, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis,," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 9, pp. 1406–1413, Sep. 2007.

[17] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style adaptation technique for speech synthesis using HSMM and suprasegmental features," *IEICE Trans. Inf. & Syst.*, vol. E89-D, no. 3, pp. 1092–1099, Mar. 2006.

[18] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 2, pp. 533–543, Feb. 2007.

[19] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Audio, Speech, & Language Processing*, vol. 17, no. 1, pp. 66–83, January 2009.

[20] T. Nose, M. Tachibana, and T. Kobayashi, "HMM-based style control for expressive speech synthesis with arbitrary speaker's voice using model adaptation,," *IEICE Trans. Inf. & Syst.*, vol. E92-D 3, no. 9, pp. 489–497, Mar. 2009.

[21] C. Darwin, *The Expression of Emotions in Man and Animals*. Londres: John Murray. Reprint Chicago: University of Chicago Press, 1965, 1872.

[22] S. Rachman, "Emotional processing," *Behaviour Research and Therapy*, vol. 18, no. 1, pp. 51 – 60, 1980. [Online]. Available: http://www.sciencedirect.com/science/article/B6V5W-45WYXRY-11D/2/5dcadef5bb239e16bdd887c9f854e77b

[23] B. Mesquita and N. H. Frijda, "Cultural Variations in Emotions - a Review," *Psychological Bulletin*, vol. 112, no. 2, pp. 179–204+, 1992.

[24] R. Barra-Chicote, J. Montero, J. Macias-Guarasa, S. Lufti, J. M. Lucas, F. Fernandez, L. D'haro, R. San-Segundo, J. Ferreiros, R. Cordoba, and J. Pardo, "Spanish Expressive Voices: Corpus for emotion research in Spanish," in *Proceedings of 6th international conference on Language Resources and Evaluation*, 2008.

[25] A. Bonafonte and A. Moreno, "Documentation of the upc_esma spanish database," *TALP Research Center, Universitat Politecnica de Catalunya, Barcelona*, pp. 2781–2784, 2008.

[26] J. Montero, J. Gutierrez-Arriola, S. Palazuelos, E. Enrquez, and J. Pardo, "Spanish emotional speech: From database to TTS," in *Proceedings of ICSLP*, Sep. 1998, pp. 923–925.

[27] I. Sainz, "Analisis de los resultados de la evaluacin Albayzin-TTS 2008," in *V Jornadas en Tecnologia del Habla*, Nov. 2008.

[28] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A. Black, and T. Nose, *The HMM-based speech synthesis system (HTS) Version 2.1*, 2008, http://hts.sp.nitech.ac.jp/.

[29] R. Barra-Chicote, J. Yamagishi, J. Montero, S. King, S. Lutfi, and J. Macias-Guarasa, "Generacion de una voz sintetica en Castellano basada en HSMM para la Evaluacion Albayzin 2008: conversion texto a voz," in *V Jornadas en Tecnologia del Habla*, Nov. 2008, pp. 115–118.

[30] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

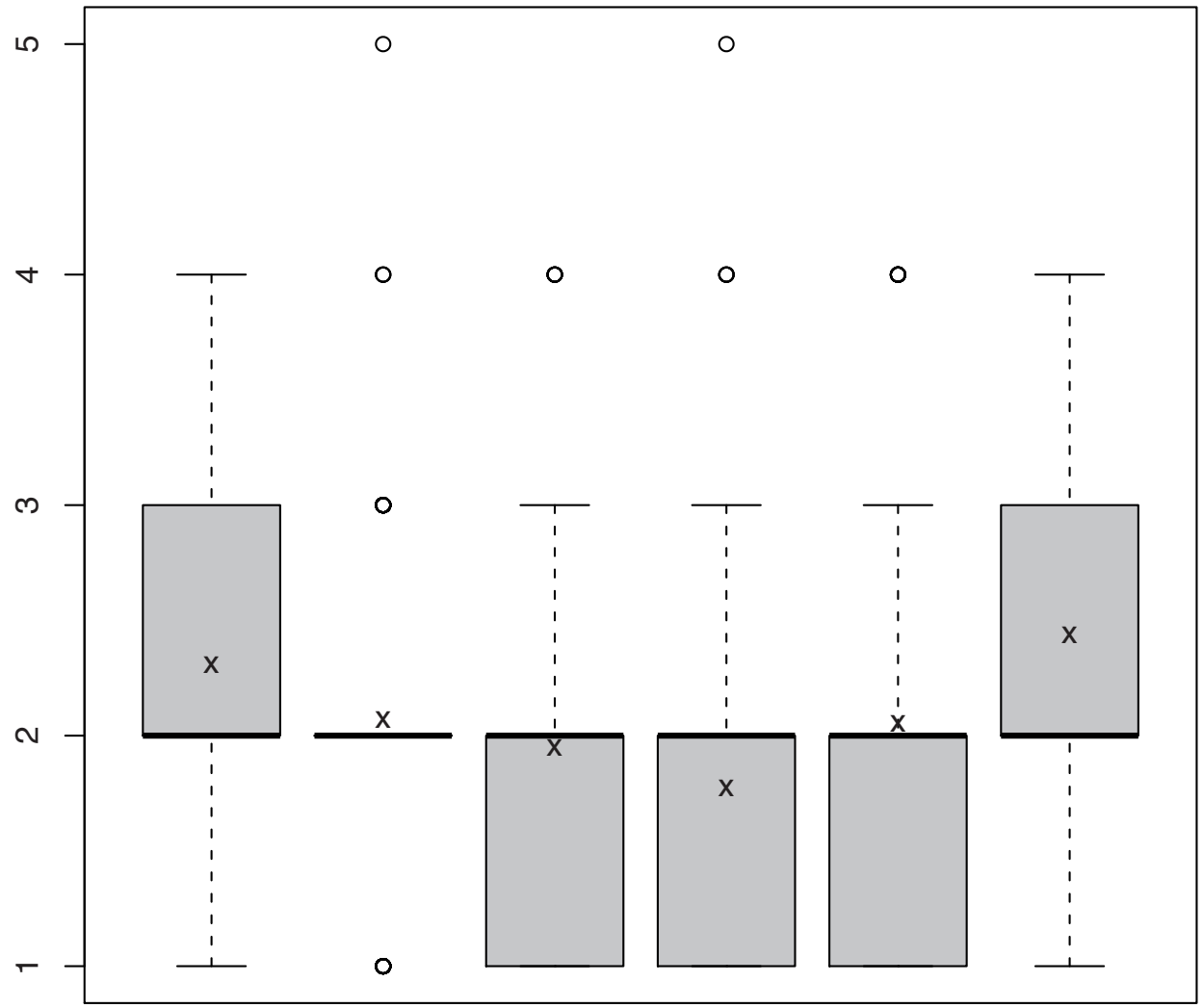
[31] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 825–834, May 2007.

[32] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 816–824, May 2007.

[33] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453–468, 1990.

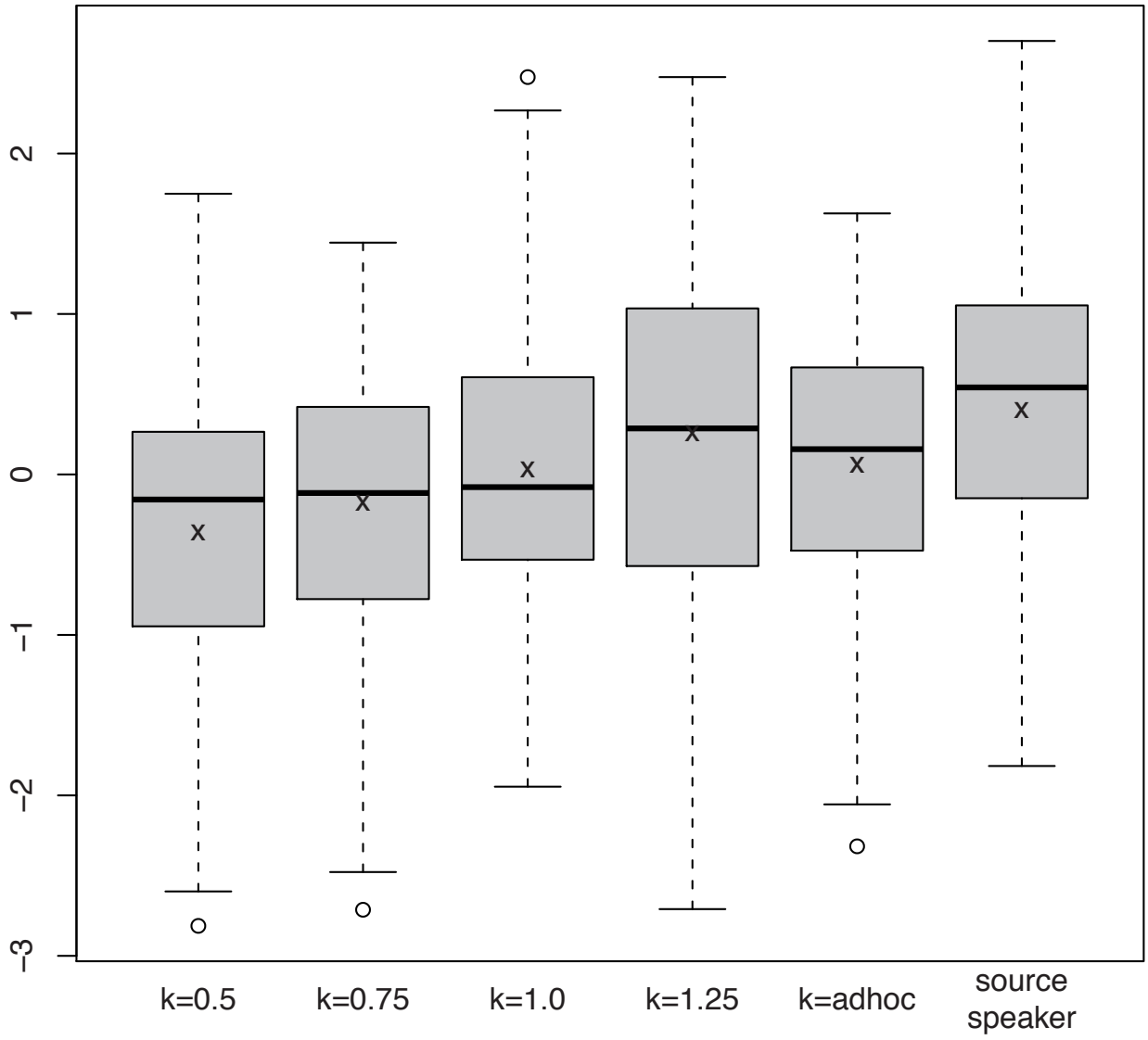
[34] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation for HMM-based speech synthesis system," *Acoustical Science and Technology*, vol. 21, no. 4, pp. 199–206, 2000.

[35] M. Fraser and S. King, "The Blizzard Challenge 2007," in *Proceedings BLZ3-2007 (in Proc. SSW6)*, Aug. 2007.




| | k=0.5 | k=0.75 | k=1.0 | k=1.25 | k=adhoc |
|---|---|---|---|---|---|
| k=0.75 | | | | | |
| k=1.00 | ■ | | | | |
| k=1.25 | ■ | ■ | | | |
| k=adhoc | | | | | |
| source speaker | | ■ | ■ | ■ | ■ |

Fig. 3. Boxplot showing speech quality (SQ) scores for emotional synthetic speech of the source speaker and the transformed speaker using different values of the extrapolation factor $k$. The table below the boxplot shows the results of pairwise Wilcoxon signed rank tests between extrapolation factors. ■ denotes a significant difference in Speech Quality (SQ) between a pair of extrapolation factors (significance level is $p = 0.05$).




| | k=0.5 | k=0.75 | k=1.0 | k=1.25 | k=adhoc |
|---|---|---|---|---|---|
| k=0.75 | | | | | |
| k=1.00 | ■ | | | | |
| k=1.25 | ■ | ■ | | | |
| k=adhoc | ■ | | | | |
| source speaker | ■ | ■ | ■ | | ■ |

Fig. 4. Boxplots showing normalised emotional strength (ES) scores for emotional synthetic speech of the source speaker and the transformed speaker using each extrapolation factor. The table below the boxplot shows results of pairwise t-tests between extrapolation factors. ■ denotes a significant difference in the normalised emotional strength (ES) between two synthesis schemes (significance level is $p = 0.05$).

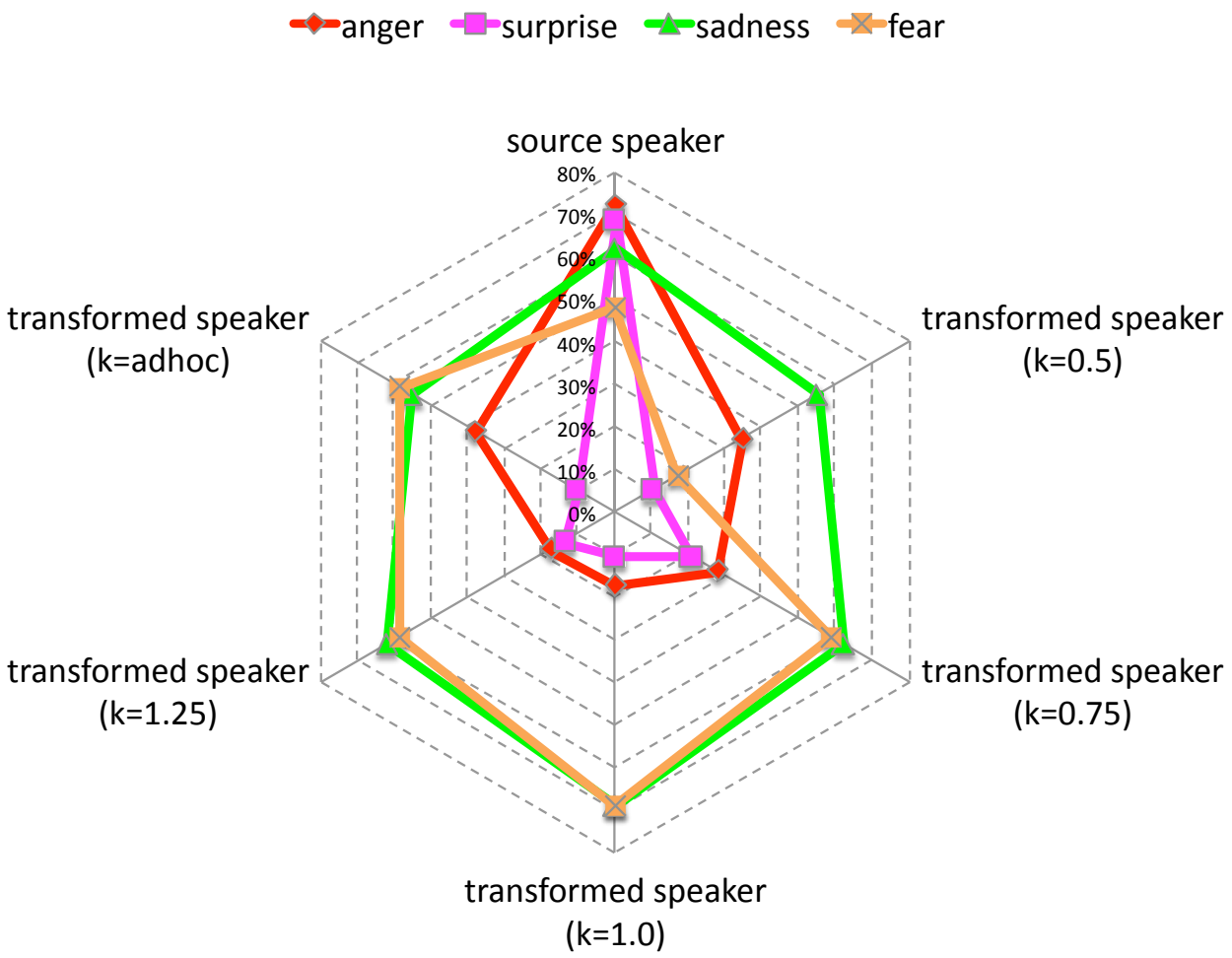



Fig. 5. EIRs for emotional synthetic speech of the source speaker and the transformed speaker using different values of the extrapolation factor $k$.

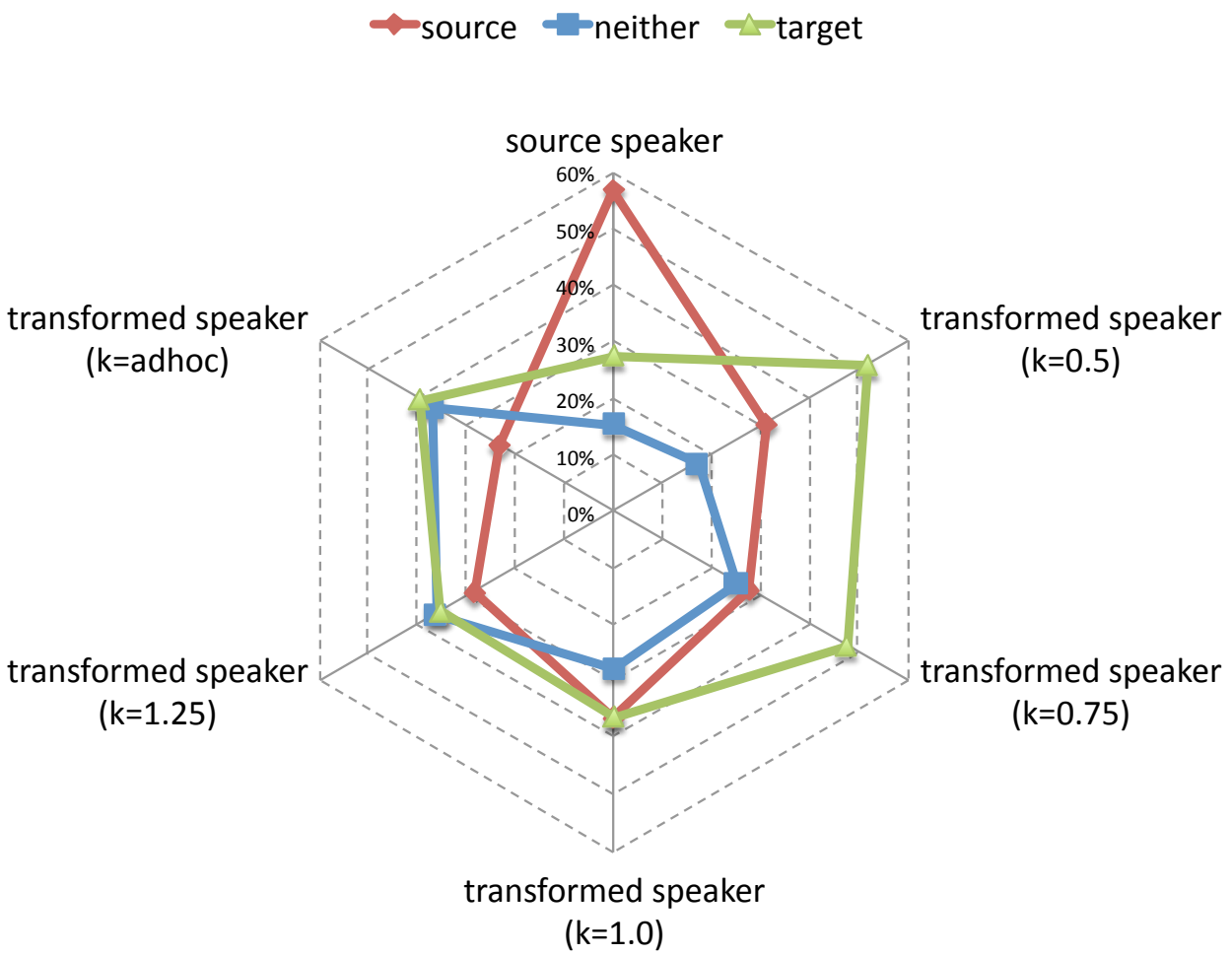



Fig. 6. SIRs for emotional synthetic speech of the source speaker and the transformed speaker using different values of the extrapolation factor $k$.

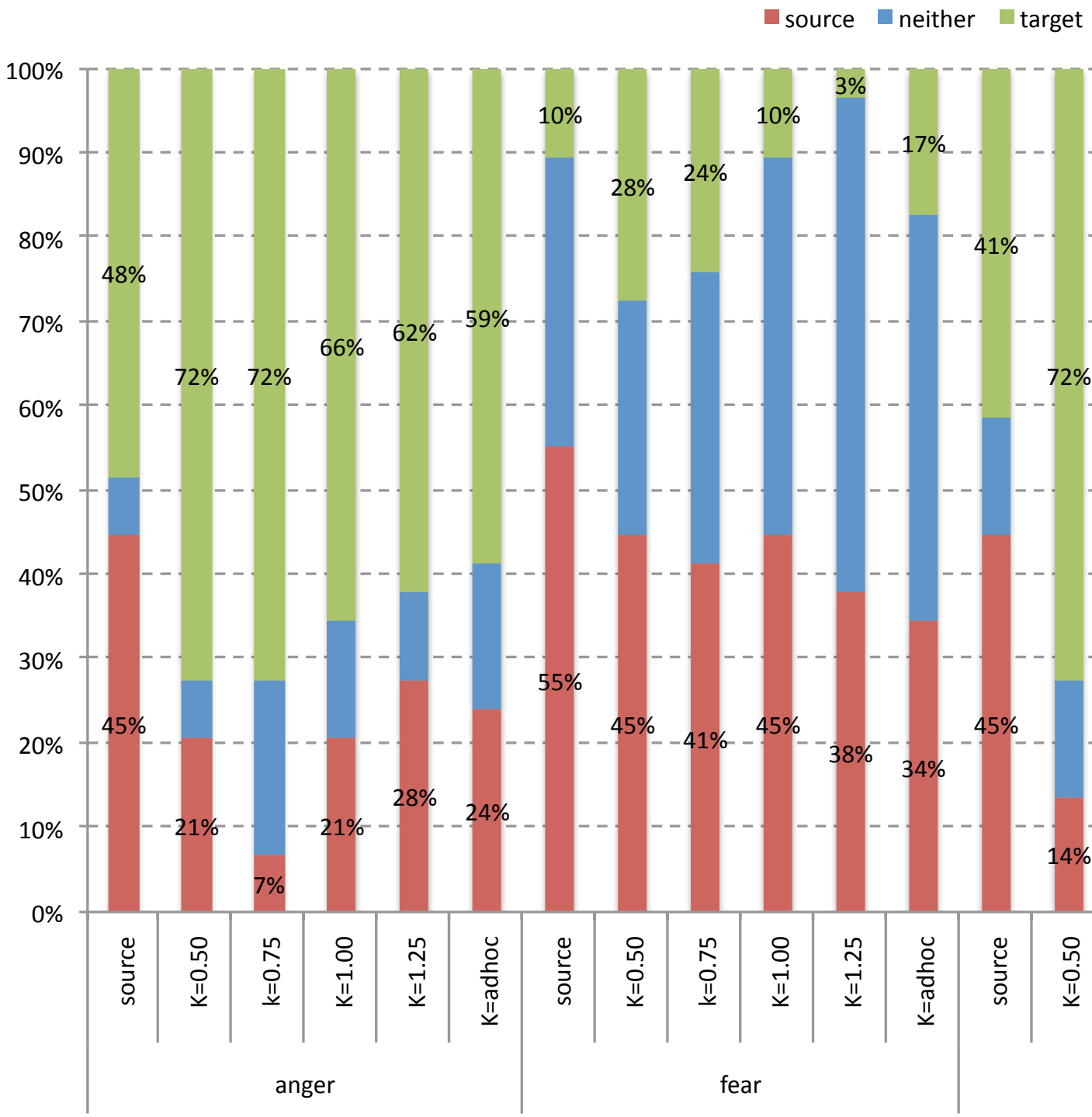



Fig. 7. SIRs for each emotional voice of the source speaker and the transformed speaker using different values of the extrapolation factor $k$.

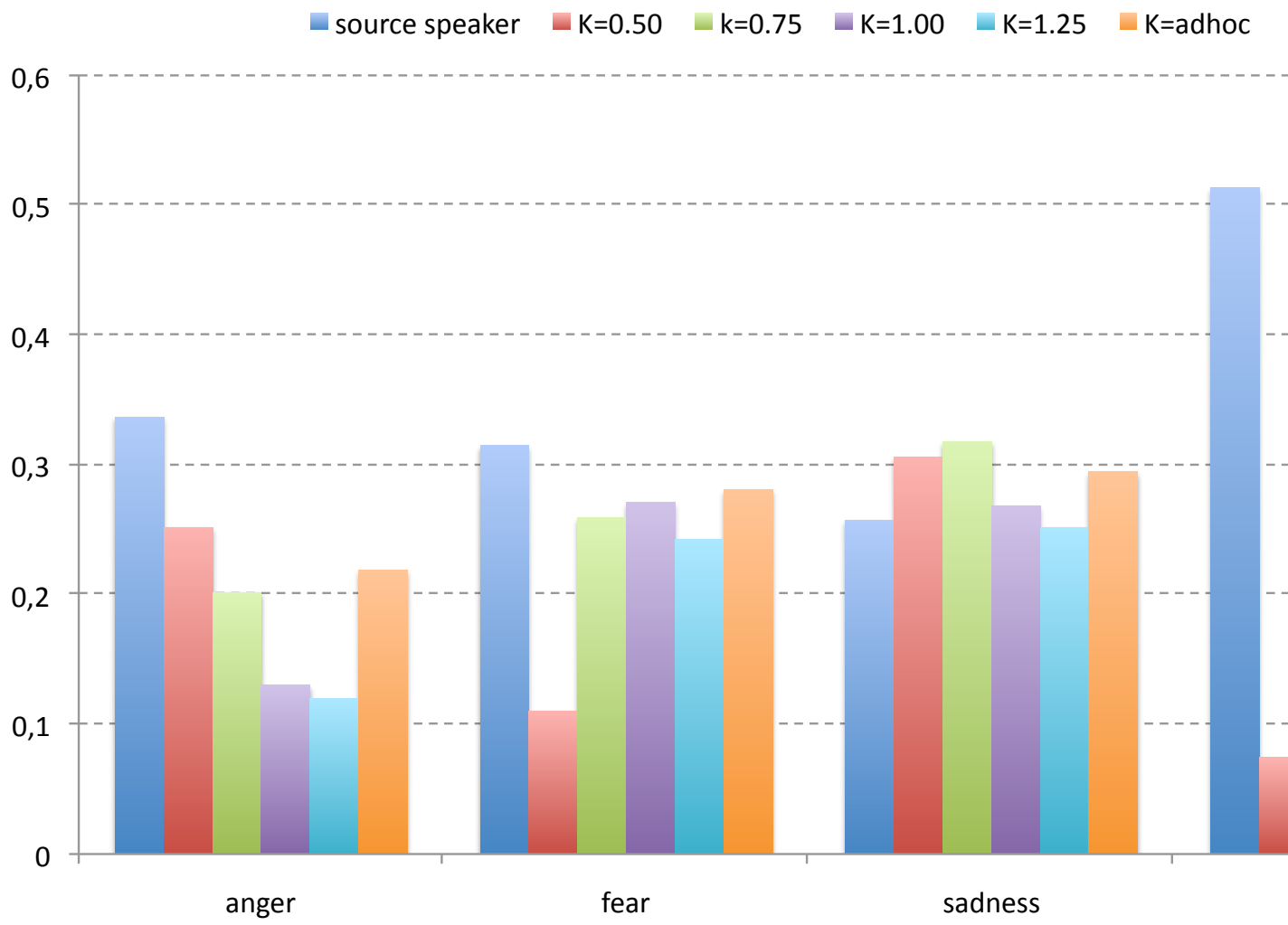



Fig. 8. EEP values for each emotional voice of the source speaker and the transformed speaker using different values of the extrapolation factor $k$.

## 6.6 Emotion transplantation

# Emotion Transplantation through Adaptation in HMM-based Speech Synthesis

Jaime Lorenzo-Trueba[a,*], Roberto Barra-Chicote[a], Junichi Yamagishi[b], Juan M. Montero[a]

[a] *Speech Technology Group, ETSI Telecomunicacion, Universidad Politecnica de Madrid, Spain*
[b] *CSTR, University of Edinburgh, United Kingdom*

**Abstract**

This paper proposes an emotion transplantation method capable of modifying a synthetic speech model through the use of CSMAPLR adaptation in order to incorporate emotional information learned from a different speaker model while maintaining the identity of the original speaker as much as possible. The proposed method relies on learning both emotional and speaker identity information by means of their adaptation function from an average voice model, and combining them into a single cascade transform capable of imbuing the desired emotion into the target speaker. This method is then applied to the task of transplanting four emotions (anger, happiness, sadness and surprise) into six target neutral speakers (3 male speakers and 3 female speakers) and evaluated in a pair of perceptual tests. The results of the evaluation show how the perceived naturalness for emotional text significantly favors the use of the proposed transplanted emotional speech synthesis when compared to traditional neutral speech synthesis, evidenced by a big increase in the perceived emotional strength of the synthesized utterances at a slight cost in speech quality.

*Keywords:* Expressive Speech Synthesis, Cascade Adaptation, Emotion Transplantation

## 1. Introduction

Current speech synthesis systems, whether we are talking about unit selection or HMM-based systems, can provide very good naturalness and intelligibility when synthesizing read speech regardless of the technology [1, 2] which is ideal for neutral speech interfaces that do not need to engage in a direct conversation with the user. On the other hand, applications such as dialog systems [3] or virtual characters, where simulating a more human-like behavior is necessary,

*Corresponding author
*Email address:* jaime.lorenzo@die.upm.es (Jaime Lorenzo-Trueba)

a neutral speech synthesis does not live up to the task. Imbuing the synthetic speech with expressive features (e.g. emotions, speaking styles...) is the role of expressive speech synthesis.

Due to the sheer amount of possible expressiveness, recording complete databases that cover all of them is unthinkable, making unit selection based systems fall behind in terms of scalability, although they are definitely capable of producing expressive speech [4, 5, 6, 7]. On the other hand, HMM-based systems, because of their parametric nature, can be easily adapted through speaker adaptation techniques and can be successfully used for this task, and have been proven to provide significant improvements in perceived speech quality [8].

One of the biggest problems of expressive speech synthesis is data acquisition. As human expressiveness is not a discrete space but a continuous one as the expressive strength and nuances vary greatly not only from person to person but from utterance to utterance for the same person. This problem can be focused on from different approaches: lexical analysis [9] for correctly classifying the available data and training more precise systems or acoustic analysis. For acoustic analysis several aspects have been considered such as expressiveness detection [10, 11, 12], expressiveness production [13, 14], expressive intensity control [15, 16] or expressiveness transplantation [17, 18].

The work present in this paper is enclosed mainly under the field of expressive speech synthesis, and aims to fix one of its main shortcomings: scalability. Human communication is so rich and so deep that it is impossible to imagine obtaining data for every combination of speaker and expressiveness, and that is why we want to propose a method capable of learning the paralinguistic information of emotional speech, control its emotional strength and transplant it to different speakers for whom we do not have any expressive information. We decided to focus on emotional speech as a particularization of expressive speech, but we can expect the transplantation method to be able to support different expressive domains.

A successful transplantation method that has been introduced lately [17, 18] is based on Cluster Adaptive Training (CAT) [19], a projective adaptation technique. As such it is only capable of producing speaker models based on linear combinations of the original training speaker models. The main advantage of this approach is that as the produced model is always a combination of pre-existing training models, the process is extremely robust, outputting very high quality speech [20]. On the other hand, the level of expressive strength or speaker similarity cannot be guaranteed as the transplantation reach is very constrained. Another known approach is the use of rule for doing simple feature transforms capable of providing expressive strength controllability and reasonably good recognition rates [21, 22], speech quality and speaker similarity degradation tends to be a problem.

Another approach to emotion transplantation is the use of rules to directly modify the synthesis models. This approach is theoretically capable of imbuing an emotion on any target speaker as long as we know the correct rules. In reality this approaches, while usually capable of providing emotional strength controllability and reasonably good recognition rates [21, 22], speech quality

and speaker similarity degradation tends to be a problem.

The proposed emotion transplantation method considers the best of both previously mentioned approaches: using adaptation to lessen speech quality degradation while using the adaptation functions as a pseudo-rules for modifying the speaker models. As a result we present a method capable of controlling expressive strength while reasonably maintaining speech quality and speaker identifiability when compared to non-transplanted expressive synthetic speech [23, 24]

The paper is organized as follows. In section 2 we introduce the neutral and emotional speech corpora we have used for training and evaluation purposes during the development of the proposed method. Section 3 introduces the transplantation method, where subsection 3.1 introduces the mathematical aspects of the used CSMAPLR adaptation and how it was expanded for our purposes, and subsection 3.2 explains in detail the procedure through which the emotion transplantation is applied presenting a pair of alternative implementations. Section 4 describes how the perceptual evaluations were carried out and analyzes the results. Finally in section 5 we present the conclusions to be drawn from this paper together with a brief summary of the main proposals.

## 2. Speech Corpora

For the development and evaluation of the proposed emotional speech transplantation method we employed both neutral and emotional databases. The emotional database (SEV [25]) has been evaluated previously for the Albayzin2012 speech synthesis evaluation, making it ideal for the introduced evaluation. The neutral data on the other hand is a combination of published databases (UVIGO-ESDA Database [26] and UPC-ESMA Database [27]) and a number of male and female speakers recorded in our laboratory environment.

**SEV Database** Emotional database consisting of a male and female speaker. Out of the available emotions only 4 of them were considered: anger, happiness, sadness and surprise also including the neutral voice as the reference. All the emotions were recorded for the same utterances favoring the learning of expressiveness cues. There is approximately 30 minutes of training speech for each emotion and speaker).

**UVIGO-ESDA Database** A database consisting of a single male Spanish speaker (UVD) in a neutral situation totaling approximately 2 hours of speech recorded in studio.

**UPC-ESMA Corpus** A database consisting of a single professional female speaker (UEM) totaling around 1.75 hours of neutral style speech, recorded in a noise-reduced room.

**Recorded Data** A number of male and female speakers were recorded in our acoustically-treated room, providing high quality and stable speech. Two male speakers (JLC and JEC) and two female speakers (NAS and EMA)

were used as the transplantation targets. Available data durations varies from 7 minutes for JEC to 30 minutes for JLC.

## 3. Emotion Transplantation

Emotion transplantation methodologies can be defined as the procedures that allow the modification of a synthetic speech model to incorporate emotional information learned from other speaker models while maintaining the identity of the original speaker as much as possible. By this definition it follows that transplantation is a field of study that aims to solve one of the biggest problems in expressive speech synthesis: scalability.

### *3.1. Adaptation-based Transplantation*

Adaptation is a powerful tool when considering emotional speech synthesis and more concretely emotion transplantation, as it allows us to exploit the versatility of HMM-based speech synthesis. In the task at hand we consider the adaptation task of generating a speaker model from an average voice model (AVM) and adaptation data for the desired target speaker [28].

Focusing on emotional speech adaptation, it has been proved that it is very important to consider not only the means of the HMM Gaussian Distributions but also the variances. This means that it is necessary for the adaptation algorithms to be more complex, or "constrained" as it is called. Ultimately, constrained structural maximum a posteriori linear regression (CSMAPLR) has been proposed and has been proven to be extremely successful for speaker adaptation, particularly when adapting from average voice models [29].

#### *3.1.1. CSMAPLR Adaptation*

CSMAPLR consists in applying the structural MAP criterion (SMAP) [30] to the CMLLR adaptation algorithm [31] and using the recursive MAP criterion [32] to estimate the transforms for simultaneously transforming the mean vectors and covariance matrices of the state output and duration distributions of the HSMM model.

There are three main reasons for using CSMAPLR as the adaptation technique. First of all is the aforementioned capability of not only adapting the mean vectors but also the covariance matrices. The second reason that differentiates CSMAPLR to the more traditional CMLLR adaptation, is that CSMAPLR makes use of the linguistic information of the regression tree by doing recursive MAP-based estimation of the transformation matrices from the root of the context decision tree to the lower nodes, effectively combining the advantages of SMAP and CMLLR. Finally, the fact that CSMAPLR relies on MAP-based estimations means that it is robust when using sparse adaptation data, which is frequently the case in the emotional speech synthesis task.

*3.1.2. Emotion Transplantation based on Cascade Adaptation*

The concept of cascade transforms has been used previously in automatic speech recognition to adapt the background models both to the target speaker and noise at the same time [33]. The transplantation method we present here is based on the same concept, but in this case we propose chaining transformations that model both the target emotion and the target speaker to produce emotional speech synthesis models.

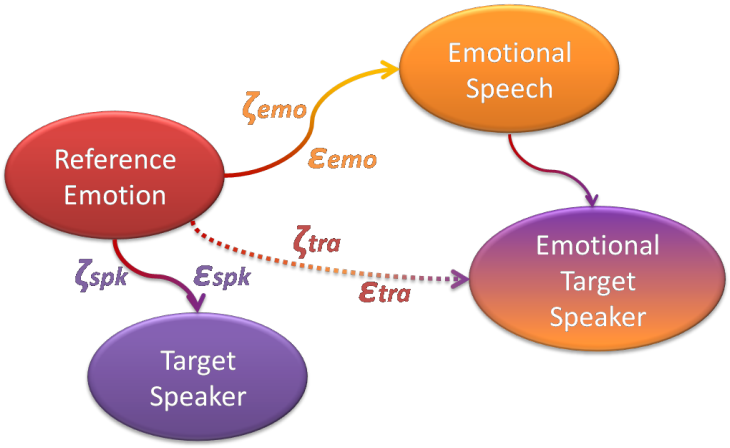



Figure 1: Schematic of the emotion transplantation method. The spheres represent the speaker models and the arrows the adaptation transformation functions. The dashed arrow represents the transplantation transformation equivalent to the proposed cascade method.

In figure 1 we can see the block diagram representation of the proposed transplantation through cascade adaptations method. If we define the CSMPALR transformation functions in terms of their rotation matrix $\zeta$ and bias vector $\epsilon$:

$$\bar{\mu}_{emo} = \zeta_{emo}\mu_N + \epsilon_{emo} \quad (1)$$

$$\bar{\Sigma}_{emo} = \zeta_{emo}\Sigma_N\zeta_{emo}^T \quad (2)$$

$$\bar{\mu}_{spk} = \zeta_{spk}\mu_N + \epsilon_{spk} \quad (3)$$

$$\bar{\Sigma}_{spk} = \zeta_{spk}\Sigma_N\zeta_{spk}^T \quad (4)$$

Where $\bar{\mu}_{emo/spk}$ and $\bar{\Sigma}_{emo/spk}$ are the mean vectors and covariance matrices of the emotional and target speaker models respectively. Then, the model resulting of applying first the emotion adaptation function and then the speaker identity becomes:

$$\bar{\mu}_{tra} = \zeta_{spk}\zeta_{emo}\mu_N + \zeta_{spk}\epsilon_{emo} + \epsilon_{spk} \quad (5)$$

$$\bar{\Sigma}_{tra} = \zeta_{spk}\zeta_{emo}\Sigma_N\zeta_{emo}^T\zeta_{spk}^T \quad (6)$$

The resulting speaker model will be able to produce emotional synthetic speech for the target speaker even if emotional training data for that particular speaker is not available.

### *3.2. Proposed Emotion Transplantation Method*

The proposed transplantation method can be summed up in three steps:

1. Adapt the reference emotion from the average voice model (Fig. 2.1).
2. Adapt the target speaker model and the target emotion from the reference emotion (Fig. 2.2).
3. Apply in cascade the emotion and speaker identity transformations to the reference emotion. The resulting model is the emotional target speaker (Fig. 2.3).

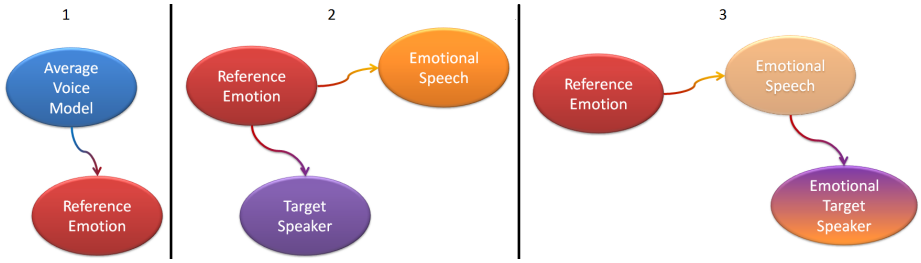



Figure 2: Step by step block diagram of the emotion transplantation method. The spheres represent the speaker models and the arrows the adaptation transformation functions.

The average voice model is obtained by applying Speaker Adaptive Training (SAT) [34] with as much training data as possible, which allows us to obtain a very context-rich background model to work with. A robust and complete AVM will be capable of producing better speech quality at synthesis time even with sparse emotional adaptation data. Also, sharing a background model for all the adaptation functions makes the cascade adaptation easier, because the context decision trees will be shared, making the adaptation functions immediately compatible between adapted models.

Adapting the reference emotion from the AVM is necessary because in the second step we have two objectives: on one hand we want to be able to learn the differences between the desired emotion and the reference emotion, effectively learning the nuances of the desired emotional speech. On the other hand we want to learn the difference between the target speaker speaking in the reference and said reference emotion, learning the nuances of the target speaker identity. If both adaptation transforms are not obtained from a single reference emotion, neutral in most cases, they will not learn the desired characteristics and the transplantation process would not be successful. Ideally, we want to have both data for the target emotion and reference emotion for the same speaker so the emotion adaptation function defines purely the target emotion, but if that is not available we can assume that using an average of different speakers will show the

relevant information of the emotions (either reference or target) while lessening the identities of the speakers.

Finally, we apply the cascade transforms for the emotion and the speaker identity as defined previously, obtaining the desired target emotional speaker. Neither the target speaker or the target emotion data had to be present in the AVM, and in the presented evaluation (section 4) we prove it to be successful with as little as 5 minutes of target speaker speech data or 30 minutes of emotional speech, which is why the proposed transplantation method is a good way of providing scalability in expressive speech synthesis. Nonetheless, we also provide some alternative approaches to the problem in case we want to simplify the transplantation process (alternative 1) or we do not have enough target emotional data to obtain reasonably good adapted models (alternative 2).

*3.2.1. Alternative 1: Including an emotions as a decision tree feature*

One possible alternative to the proposed emotion transplantation method is to include the emotion of each utterance as an additional feature to the training labels in the average model HMM training process, together with adding the respecting questions to the decision tree modeling. This would imbue the emotion characterization in the modeling process, producing more complex decision trees. By doing this, synthesizing the desired emotion for the target speaker only requires adapting to the target speaker from the average background model and including the emotion feature to the text label to be synthesized.

This alternative, much simpler than the proposed transplantation system would be expected to provide voices with speech quality similar to the average voice model and that of the neutral speech as there is no transplantation involved, but at the cost of speaker similarity and emotional strength. Moreover, this approach also removes the emotional strength control capabilities present in the proposed system unless there is a complex labeling of the emotional strength of all the utterances in the training database, a process that is very hard and costly and one of the main problems we avoid with out transplantation method.

A final limitation of this alternative is that the average voice model must be trained with all the emotions we want to synthesize in our task, with the utterances labeled accordingly. If we were to require the addition of a new emotion to the task the AVM would have to be trained once again so it is included in the decision tree. This is a process that can take a very long time, increasing with each additional emotion, which is also something a problem we wanted to solve with our proposed method.

*3.2.2. Alternative 2: Transplanting into an average emotion*

Another alternative is to join all our emotional data into an average emotion model [35]. This average, when transplanted into a neutral speaker model, can be expected to imbue an undefined emotion that removes the typical monotony in read speech models. This alternative can also be expected to provide higher quality speech when compared to transplanting a single emotion as the adaptation process for the average emotion can make use of much more data, thus giving more stable adaptation functions. This approach could be very useful

when the task does not require us to synthesize any particular emotion or if we do not have enough emotional data to obtain good emotion transplantation quality.

Naturally, this alternative also presents numerous shortcomings: if there is a significant bias towards positive or negative data in the average emotion model, transplanting the average emotion could be the same as transplanting an emotion, resulting on unnatural synthesized utterances for opposite emotions such as producing happy speech for a sad text. Another expected problem is that the naturalness should be lower than transplanting the correct emotion for the text to be synthesized.

## 4. Perceptual Evaluations

The goal of the perceptual evaluation was to verify if the expressiveness was transplanted successfully in terms of naturalness, speech quality and emotional strength. Naturalness measure was done by means of a preference tests, as they are very useful when we want to compare systems that are similar between them but with variation in some conditions [36], and are a reliable way to obtain statistically relevant preference measures that are more separable than the traditional MOS evaluations. Two different evaluations were carried out, a first one that compared the proposed emotion transplantation system with the traditional neutral synthetic voice to validate the transplantation method, and a second one that compared the neutral synthetic voice with an average emotion transplanted into the speaker (alternative 2 in section 3.2.2) in order to prove that the benefits of transplanting the correct emotion into the speaker are higher than just modifying the neutral speech to sound less machine-like.

Four emotions (anger, happiness, sadness and surprise) learned from the Spanish Emotional Voices corpus were transplanted into 3 male speakers and 3 female speakers, so for each testing session the total number of systems was 24. Following the latin-square [37] approach this meant that we needed 24 different utterances to be synthesized (or selected from the natural database) for all the systems, to be presented to the listeners in a random order without repeating. In the test the listener was presented by means of a web interface with two audio samples (transplanted correct emotion and neutral voice for the first version of the test, transplanted average emotion and neutral voice for the second), together with a transcription of the synthesized text and the intended emotion to be transmitted. The samples could be played as many times as desired by the listener, and the synthesized texts, not present in the training data, were written by ourselves to present clear emotional context that always corresponded with the transplanted emotion. Then, the listener was asked their preference on which of the samples was more adequate to transmit the desired emotion, ending with the traditional 5 point MOS evaluation for both speech quality (very bad to very good) and emotional strength (very low to very high). The evaluation for speech quality and emotional strength had to be answered for both samples regardless of the selected preference.

Table 1: Results for both preference tests

| **Preference** | Transplanted Emotion | Transplanted Average |
|---|---|---|
| Anger | 77% | 66% |
| Happiness | 96% | 87% |
| Sadness | 82% | 61% |
| Surprise | 95% | 84 % |

The results for the naturalness preference tests can be seen in table 1. The results for each of the six speakers are averaged into the row of every emotion, while the results for the first test that compares the transplanted correct emotion with the neutral voice are shown in the second column (Transplanted Emotion) and the results for the second test that compared the transplanted averaged emotion with the neutral voice are shown in the third column (Transplanted Average). The first result that can be drawn from the table is that for both tests there is a significant preference for the transplanted system when compared to the traditional neutral system according to the chi-squared significance test. Nonetheless, the results for transplanting the correct emotion are in average 10% higher than for the average emotion transplantation. In particular, positive emotions (happiness and surprise) show better results for both cases, reaching an extremely good 95-96% preference in the case of the first test. On the other hand negative emotions (anger and sadness) while still showing a very high preference for the proposed transplantation (an average of 79.5% ), they are not so good when transplanting an average emotion (63.5%). This means that while transplanting the target emotion gives a huge boost in naturalness to the synthesizer, coloring the voice with an undefined emotion helps the naturalness of positive text synthesis, it is not so helpful for synthesizing negative texts.

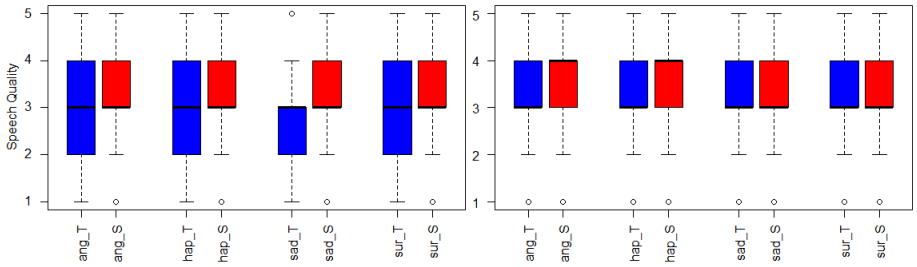



Figure 3: Boxplots for the speech quality results for the first and second test respectively. Red bars always represent the neutral system while blue bars represent the desired emotion transplantation results in the first boxplot and the average emotional model transplantation in the second case. Ang, hap, sad and sur mean anger, happiness, sadness and surprise respectively.

For the results of the MOS test for Speech quality we can look at the boxplots in figure 3. In these boxplots the results of all the different speakers have been averaged for all emotions, with the blue bars representing the neutral systems

and the red bars representing the transplanted system in each of the two tests. By looking at the plots it is clear that there is not a significant decrease in speech quality overall, although particularly for sadness in the transplanted emotion there is a slight decrease. In any case, we can also see that there is a greater variance in speech quality for the proposed system, which is due to possible inconsistencies in the spectrum transplantation to different speakers learned from the limited emotional speech data. On the other hand, for the average emotional model the results are very close to the original neutral synthetic speech, something that we can assume to be due to the increased robustness provided by the average emotional model.

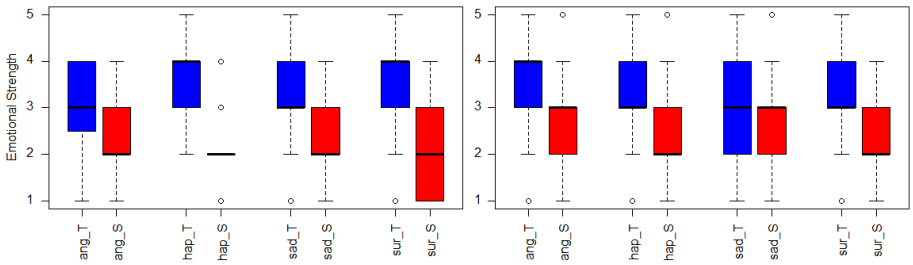



Figure 4: Boxplots for the emotional strength results for the first and second test respectively. Red bars always represent the neutral system while blue bars represent the desired emotion transplantation results in the first boxplot and the average emotional model transplantation in the second case. Ang, hap, sad and sur mean anger, happiness, sadness and surprise respectively.

The perceived emotional strength was also measured by means of a MOS test, whose results we can see in the boxplots in figure 4. The structure of these boxplots is the same as in the speech quality case. By looking at these results it is evident that there is a significant increase in the perceived emotional strength in most cases. Specially in the case of the first test, where we compared the proposed transplantation system with the neutral system, we can see how the perceived emotional strength is significantly superior in all cases excepting anger, where there are not definite results. On the other hand, the second test proves that an average emotional model is also capable of providing better emotional strength results although not so much in this case for the negative emotions, particularly sadness, as happened in the preference test.

Regarding the statistical significance of the results, for the preference test we applied the chi-squared criterion and for the speech quality and emotional strength MOS tests we applied the Wilkoxon Signed-Rank Test for a 95% confidence ratio. The results of applying the test can be seen in table 2, where we can see that all the results excepting the two closest ones (speech quality for Anger in the first phase of the evaluation and speech quality for surprise in the second phase) passed the verifications.

To sum up the results in the three categories for both tests, the proposed emotion transplantation system provides an average 87% preference rate when

Table 2: Results of the significance tests. An X means that that particular result is statistically significant and a blank means that it is not. The prefixes "Pref" stands for preference test, "SQ" for speech quality test and "ES" for emotional strength test. The suffixes "1" and "2" refer the first and second test respectively.

| **Emotion** | Pref-1 | SQ-1 | ES-1 | Pref-2 | SQ-2 | ES-2 |
|---|---|---|---|---|---|---|
| Anger | X | | X | X | X | X |
| Happiness | X | X | X | X | X | X |
| Sadness | X | X | X | X | X | X |
| Surprise | X | X | X | X | | X |

compared to the traditional neutral synthetic system while increasing the perceived emotional strength in an average of 1.2 points at the cost of 0.4 points in speech quality when using a 5 points MOS scale. At the same time, the contrast average emotion transplantation system provides an average 75% preference rate with an increase of 0.7 points in emotional strength at the cost of only 0.2 points in speech quality. All in all, we can say that the system is clearly capable of transplanting the emotional information learned from a source speaker into different target speakers regardless of gender with significant increases in perceived naturalness and emotional strength when compared to traditional systems at a slight cost in speech quality. This benefits are also validated by comparing with a simpler transplantation that just aims to give color to the voice in order to remove the excessive neutrality of traditional read speech synthesis, as the proposed emotion transplantation system clearly improves the results at a comparative decrease of only 0.2 points in speech quality.

## 5. Conclusions

We have proposed an emotion transplantation method capable of learning the paralinguistic nuances of any particular emotion in order to transplant them into a new target speaker for whom only traditional, neutral read speech recordings are available. This is done by means of chaining a pair of CSMAPLR adaptation functions, one that characterizes the target speaker identity and another that defines the paralinguistic characteristics of the desired emotion. Finally a pair of perceptual evaluations were carried out. For the perceptual evaluation, four emotions (anger, happiness, sadness and surprise) from an Spanish emotional database and six target speakers (three male and three female) were considered. a first evaluation compared in terms of naturalness, speech quality and emotional strength the proposed transplantation method with traditional neutral read speech synthesis. This first test showed that there is a very clear preference (an average of 87% preference between all the emotions) for the emotional synthesizer, reaching as high as 96% for happiness, and a perceived increase in emotional strength of an average of 1.2 points in the MOS scale at a cost of only 0.4 points in speech quality. The second test compared an average emotion transplantation with the neutral speech, and showed that just

by adding an undefined color to the voice is able to improve the perceived naturalness of the synthetic speech up to an average of 75% preference at a cost of only 0.2 points in speech quality, although the average increase in perceived emotional strength only reaches 0.7 points.

## 6. Acknowledgments

The work leading to these results has received funding from the European Union under grant agreement 287678. It has also been supported by TIMPANO(TIN2011-28169-C05-03), INAPRA (MICINN, DPI2010-21247-C02-02), and MA2VICMR (Comunidad Autonoma de Madrid, S2009/TIC-1542) projects. Jaime Lorenzo has been funded by Universidad Politecnica de Madrid under grant SBUPM-QTKTZHB. Authors also thank the other members of the Speech Technology Group and Simple4All project for the continuous and fruitful discussion on these topics.

## References

[1] R. Barra-Chicote, J. Yamagishi, S. King, J. M. Montero, J. Macias-Guarasa, Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech, Speech Communication 52 (5) (2010) 394–404.

[2] R. Barra-Chicote, Contributions to the analysis, design and evaluation of strategies for corpus-based emotional speech synthesis, Ph.D. thesis, ETSIT-UPM (2011).

[3] S. L. Lutfi, F. Fernández-Martínez, J. Lorenzo-Trueba, R. Barra-Chicote, J. M. Montero, I feel you: The design and evaluation of a domotic affect-sensitive spoken conversational agent, Sensors 13 (8) (2013) 10519–10538.

[4] D. Erro, E. Navas, I. Herndez, I. Saratxaga, Emotion conversion based on prosodic unit selection, Audio, Speech, and Language Processing, IEEE Transactions on 18 (5) (2010) 974–983.

[5] J. Adell, A. Bonafonte, D. Escudero-Mancebo, Modelling filled pauses prosody to synthesise disfluent speech, Proc. ISCA Speech Prosody, Chicago, USA.

[6] S. Andersson, K. Georgila, D. Traum, M. Aylett, R. A. Clark, Prediction and realisation of conversational characteristics by utilising spontaneous speech for unit selection, Speech Prosody.

[7] J. Adell, D. Escudero, A. Bonafonte, Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence, Speech Communication 54 (3) (2012) 459–476.

[8] J. Yamagishi, K. Onishi, T. Masuko, T. Kobayashi, Acoustic modeling of speaking styles and emotional expressions in hmm-based speech synthesis, IEICE TRANSACTIONS on Information and Systems 88 (3) (2005) 502–509.

[9] S. Andersson, J. Yamagishi, R. A. Clark, Synthesis and evaluation of conversational characteristics in hmm-based speech synthesis, Speech Communication 54 (2) (2012) 175–188.

[10] J. Lorenzo-Trueba, R. Barra-Chicote, T. Raitio, N. Obin, P. Alku, J. Yamagishi, J. M. Montero, Towards glottal source controllability in expressive speech synthesis, in: Interspeech 2012, 13th Annual Conference of the International Speech Communication Association. Portland, Oregon. September 9-13, 2012.

[11] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, G. Rigoll, Cross-corpus acoustic emotion recognition: variances and strategies, Affective Computing, IEEE Transactions on 1 (2) (2010) 119–131.

[12] M. El Ayadi, M. S. Kamel, F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognition 44 (3) (2011) 572–587.

[13] N. Obin, P. Lanchantin, A. Lacheret, X. Rodet, et al., Discrete/continuous modelling of speaking style in hmm-based speech synthesis: Design and evaluation, in: Interspeech, 2011.

[14] T. Raitio, A. Suni, M. Vainio, P. Alku, Synthesis and perception of breathy, normal, and lombard speech in the presence of noise, Computer Speech & Language.

[15] B. Picart, T. Drugman, T. Dutoit, Continuous control of the degree of articulation in hmm-based speech synthesis., in: INTERSPEECH, 2011, pp. 1797–1800.

[16] T. Nose, T. Kobayashi, An intuitive style control technique in hmm-based expressive speech synthesis using subjective style intensity and multiple-regression global variance model, Speech Communication.

[17] L. Chen, M. Gales, V. Wan, J. Latorre, M. Akamine, Exploring rich expressive information from audiobook data using cluster adaptive training, in: Interspeech 2012, 13th Annual Conference of the International Speech Communication Association. Portland, Oregon. September 9-13, 2012.

[18] J. Latorre, V. Wan, M. J. Gales, L. Chen, K. Chin, K. Knill, M. Akamine, Speech factorization for hmm-tts based on cluster adaptive training., in: INTERSPEECH, 2012.

[19] M. J. Gales, Cluster adaptive training of hidden markov models, Speech and Audio Processing, IEEE Transactions on 8 (4) (2000) 417–428.

[20] K. Yanagisawa, J. Latorre, V. Wan, M. J. Gales, S. King, Noise robustness in hmm-tts speaker adaptation, order 5 (2013) 10.

[21] E. Zovato, A. Pacchiotti, S. Quazza, S. Sandri, Towards emotional speech synthesis: A rule based approach, in: Fifth ISCA Workshop on Speech Synthesis, 2004.

[22] S. Takeda, Y. Kabuta, T. Inoue, M. Hatoko, Proposal of a japanese-speech-synthesis method with dimensional representation of emotions based on prosody as well as voice-quality conversion, International Journal of Affective Engineering 12 (2) (2013) 79–88.

[23] J. Lorenzo-Trueba, R. Barra-Chicote, J. Yamagishi, O. Watts, J. M. Montero, Towards speaking style transplantation in speech synthesis, in: 8th ISCA Speech Synthesis Workshop, 2013.

[24] J. Lorenzo-Trueba, R. Barra-Chicote, J. Yamagishi, O. Watts, J. Montero, Evaluation of a transplantation algorithm for expressive speech synthesis, in: proccedings of Workshop en Tecnologias Accesibles, IV Congreso Español de Informatica CEDI2013, 2013.

[25] R. Barra-Chicote, J. M. Montero, J. Macias-Guarasa, S. Lufti, J. M. Lucas, F. Fernandez, L. F. D'haro, R. San-Segundo, J. Ferreiros, R. Cordoba, J. M. Pardo, Spanish expressive voices: Corpus for emotion research in spanish, Proc. of LREC.

[26] C. M. E.T. Banga, Documentation of the uvigo-esda spanish database, Tech. rep., Grupo de Tecnoloxias Multimedia, Universidad de Vigo, Vigo, Espaa (2010).

[27] A. Bonafonte, A. Moreno, Documentation of the upc-esma spanish database, TALP Research Center, Universitat Politecnica de Catalunya, Barcelona (2008) 2781–2784.

[28] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, T. Kobayashi, A training method of average voice model for hmm-based speech synthesis, IEICE transactions on fundamentals of electronics, communications and computer sciences 86 (8) (2003) 1956–1963.

[29] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, J. Isogai, Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm, Audio, Speech, and Language Processing, IEEE Transactions on 17 (1) (2009) 66–83.

[30] K. Shinoda, C.-H. Lee, Structural map speaker adaptation using hierarchical priors, in: Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on, IEEE, 1997, pp. 381–388.