Deliverable D3.1

Automatic parameterization with WLP-based GIF-techniques

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	Automatic parameterization with WLP-based GIF-techniques
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 / Task T3.1
Editor	Tuomo Raitio (AALTO)
Editor address	tuomo.raitio@aalto.fi
Author(s), in alphabetical order	Paavo Alku, Mircea Giurgiu, Tuomo Raitio
EC Project Officer	Pierre Paul Sondag

Abstract

All statistical parametric speech synthesis methods make use of a vocoder during training to convert speech waveforms into compressed sets of parameters. In the synthesis phase, the vocoder is responsible for the inverse process, that is, the synthesis of speech from parameters obtained from statistical models. One of the objectives of SIMPLE4ALL is to search for new vocoders that parameterize speech with models that are closer to the real human speech production mechanism than those currently used in HMM-based synthesis, with the goal of producing more natural-sounding speech. In order to attain this goal, we are developing techniques that utilize glottal inverse filtering (GIF) in order to separate speech into glottal excitation and vocal tract parameters. This research has resulted in a vocoder, called GlottHMM, which is the main accomplishment reported here in D3.1. GlottHMM is implemented as a software package which includes methods for the automatic parameterization of speech using GIF-based techniques. In this package, it is possible to select either conventional linear prediction (LP) or weighted linear prediction (WLP) as the means of parameterizing the vocal tract. The deliverable also reports on results from several experiments in which GlottHMM was evaluated and tested jointly by the SIMPLE4ALL partners. These experiments indicate, for example, that GlottHMM succeeds in generating a clean intelligible synthetic voice from training data that had previously been compressed using a lossy method and which had been recorded in non-ideal conditions. In addition, the use of GlottHMM led to very good accuracy when evaluated in terms of recognition of emotions in a listening test.

Contents

1 Introduction

The vocoder is one of the key elements in statistical parametric speech synthesis and is widely considered to be one of the limiting factors in the quality of current systems. In the training phase, it is responsible for transforming the speech pressure signal into parameter streams which are then modeled statistically by HMM-like models. In the synthesis phase, the vocoder creates the time-domain waveform of synthetic speech starting from data from the statistical models. A major drawback in current statistical speech synthesis is its poorer signal quality in comparison to unit selection techniques: the choice of vocoding technique is one of the main factors affecting this. Therefore, SIMPLE4ALL includes work to develop new vocoding techniques and the main activity of this research takes place in WP3, particularly in Task T3.1 ("Source modeling"). Our work on this topic is motivated by one of the over-arching objectives of SIMPLE4ALL , namely the search for new models of speech that are closer to the real human speech production mechanism than the models currently used in HMM-based synthesis. The models we are studying are based on glottal inverse filtering (GIF), a computational method to separate the speech signal into the glottal excitation (i.e. the real source of voiced speech generated by the vocal folds) and the vocal tract. We expect that these new physiologically-oriented models are inherently able to capture the wide range of dynamics of human speech, and are particularly well-suited to capturing varied voice qualities.

Task T3.1 ("Source modeling") focuses on the development of new vocoding techniques based on GIF techniques where we have studied both traditional all-pole techniques (Linear prediction, LP) and their temporally-weighted variants (Weighted linear prediction, WLP) in modeling of the vocal tract. Deliverable D3.1 constitutes a package where these new vocoding techniques are implemented in C code and distributed together with a user manual. In the following, an overview of the developed automatic parameterization method is first given after which the contents of the program package are briefly described. In addition, evaluation results obtained are described, with reference to articles published jointly by SIMPLE4ALL partners.

2 The GlottHMM vocoder

The statistical parametric speech synthesis system GlottHMM is built on a basic framework of a hidden Markov model (HMM) based speech synthesis system but it uses a distinct type of vocoder for parameterizing and synthesizing speech. GlottHMM aims at the accurate modeling of the speech production mechanism by decomposing speech into the vocal tract filter and the voice source signal using glottal inverse filtering, and emphasizing the modeling of the voice source. GlottHMM has been in constant development since its first publication in 2008 [1].

2.1 Algorithm

In GlottHMM speech analysis, the speech signal is first high-pass filtered with a cut-off frequency of 70 Hz in order to remove possible low frequency fluctuations which may distort the glottal flow estimate. The speech signal is then windowed into two types of frames: a short frame is used to extract the vocal tract filter, voice source spectral envelope, and short time energy of speech. A longer frame is used for estimating several pitch periods of the glottal flow, which many of the voice source parameters require. The length of the shorter frame is 25 ms, whereas the length of the longer frame depends on the fundamental frequency (f_0) range of the speaker.

For glottal inverse filtering, iterative adaptive inverse filtering (IAIF) is used. The algorithm uses linear prediction (LP) for estimating the spectral envelope of speech. Alternatively, weighted linear prediction (WLP) can be used in order to reduce the degrading effect of the harmonics on the vocal tract spectrum estimate. The outputs of the IAIF algorithm are the estimated vocal tract LP filter and the estimated glottal flow signal. The vocal tract filter is converted into line spectral frequencies (LSFs), a parametric representation of LP information well-suited to be used in a statistical parametric speech synthesis system, providing stability and low spectral distortion.

The longer frame is processed with the same IAIF algorithm to estimate the glottal flow signal. Ideally, at least two complete pitch periods are included in the glottal flow estimate even at the lowest f_0 values. The glottal flow

Feature	Number of parameters	Type/Unit	Contribution
Vocal tract spectrum	20–30	LSF	Vocal tract filter
Energy	1	dB	Voice source
Fundamental frequency	1	$log f_0$	Voice source
Harmonic-to-noise ratio	5–10	dB/ERB	Voice source
Voice source spectrum	5–10	LSF	Voice source
Pulse library	1000–10000	Pulse waveform	Voice source

Table 2.1a: The GlotHMM vocoder parameters

signal is used for defining f_0, estimated with the autocorrelation method. The harmonic-to-noise ratio (HNR) of the signal indicates the degree of voicing, i.e., the relative amplitudes of the periodic vibratory glottal excitation and the aperiodic noise component of the excitation. The HNR is based on the ratio between the upper and lower smoothed spectral envelopes (defined by the harmonic peaks and interharmonic valleys, respectively) and averaged across five frequency bands according to the equivalent rectangular bandwidth (ERB) scale. Finally, the glottal closure instants (GCIs) are detected by a simple peak picking algorithm that searches for the negative excitation peaks of the glottal flow derivative at fundamental period intervals. For all the two-pitch period speech segments found, the modified IAIF algorithm is applied pitch-synchronously again in order to yield a better estimate of the glottal flow. The re-estimated two-period glottal flow derivative waveforms are windowed with a Hann window and a pulse library is constructed from the extracted waveforms. The speech features extracted by the vocoder are summarized in Table 2.1a.

There are several ways to utilize the glottal flow pulse library that is obtained at the end of parameterization. In the original implementation of GlottHMM, only a single pulse was used and modified for creating the voiced excitation. In a more complex version, a pulse from the library is selected for each time instant by minimizing the target cost of the voice source parameters and concatenation cost of the pulse waveforms, and finally the selected pulses are concatenated to create the excitation signal. In the latest experiments, an approach similar to the one in Drugman et al. (2012) was adopted by using only a single carefully-constructed pulse (the mean of the pulse library) as a basis for the voiced excitation.

Thus, in synthesis, the basis of voiced excitation can be either a single glottal flow pulse extracted from natural speech, a selection of pulses from the pulse library, or the mean of the pulse library. The excitation pulse(s) are interpolated according to f_0 and scaled according to the energy feature. In order to control the degree of voicing, the amount of noise in the excitation is matched by manipulating the phase and magnitude of the spectrum of each pulse according to the HNR within each ERB band. Furthermore, the spectral tilt of each pulse is modified according to the all-pole spectrum generated by the HMM. This is achieved by filtering the pulse train with an adaptive infinite impulse response (IIR) filter which flattens the spectrum of the pulse train and applies the desired spectrum. The unvoiced excitation is composed of white noise, whose gain is determined according to the energy measure generated by the HMM system. Formant enhancement can be applied to the vocal tract LSFs in order to reduce over-smoothing caused by the statistical modeling. Finally, LSFs are converted back to LP coefficients describing the vocal tract spectrum and used for filtering the combined excitation signal.

2.2 GlottHMM software package

The GlottHMM software package consists of the analysis and synthesis programs written in standard C. GlottHMM is primarily intended to be used as a vocoder in statistical parametric speech synthesis, but it can be used also for speech analysis and modification. The package also includes pulse library construction scripts written in Matlab. Default configuration files for parameterizing and synthesizing speech are also included. The whole software package is documented in an extensive manual both describing the use of the software from parameterization and

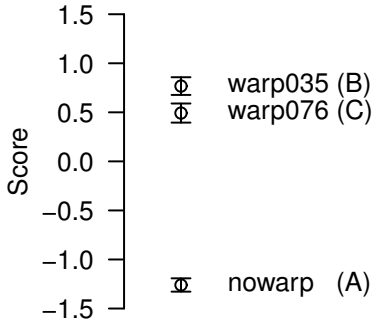

Figure 3.0a: Quality ratings of A) non-warped, B) slightly warped ($\alpha = 0.35$), and C) highly warped ($\alpha = 0.75$) systems.

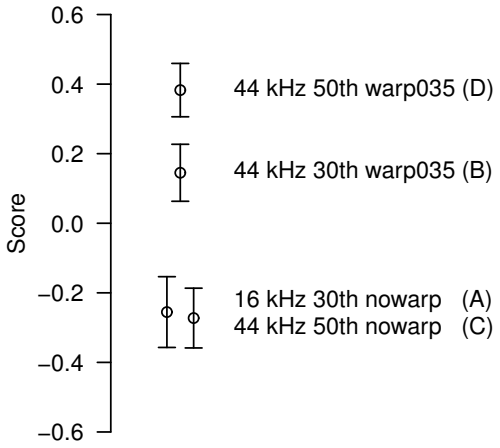

Figure 3.0b: Quality ratings of A) non-warped 16 kHz 30th order system, B) non-warped 44 kHz 50th order system, C) warped ($\alpha = 0.35$) 44 kHz 30th order system, and D) warped ($\alpha = 0.35$) 44 kHz 50th order system.

parameter training to synthesis and giving a technical description of the components of the package.

3 Collaborative evaluation of the GlottHMM vocoder

Our work on task T3.1 and in preparing deliverable D3.1 has been characterized by intensive collaboration with other SIMPLE4ALL participants. This collaboration has been particularly helpful in testing and evaluating the vocoder toolkit. Our collaboration with UH has addressed the role of the listening environment [2] in the evaluation of intelligibility of synthetic speech based on methods developed in T3.1. We compared conventional headphone tests with an evaluation system in which speech is listened to in highly realistic, yet controllable noise conditions which are created in the laboratory with a multichannel sound reproduction system. Our results indicate that the evaluation of intelligibility of synthetic speech is in general independent of whether one uses a mono, stereo or multichannel setup. In addition, our joint work with UH has addressed the effect of speech bandwidth on the quality of synthetic speech [3]. In [3], we compared synthetic speech, generated again with a tool developed in T3.1, by varying both the speech bandwidth (8 kHz vs. 22 kHz, corresponding to 16 kHz and 44 kHz sampling rates) and the all-pole vocal tract modeling technique utilized in GIF (ordinary LP vs. warped LP). Our results indicate that wideband synthetic speech, produced by using warped LP, was rated better than narrowband speech and wideband speech synthesized with vocal tract modeled with ordinary LP (see Figures 3.0a and 3.0b). Finally, UH and AALTO participated jointly in the annual international speech synthesis event, the Blizzard Challenge [4]. Even though the speech material of this year's Blizzard Challenge was more challenging than in previous years (see also deliverable D1.4) we were able to achieve a clean, intelligible voice with decent, above-average prosody characteristics and our submission was among the best HMM-based synthesizers (see Figures 3.0c, 3.0d, and 3.0e).

The vocoder described here in deliverable D3.1 has also been utilized in joint studies on expressive speech synthesis between UPM , AALTO and UEDIN [5]. Our joint study published in [5] gave encouraging results according to which the recognition accuracy of expressive speech could be improved with the help of GlottHMM. Finally, the vocoder has been tested and utilized from several perspectives by UTCN . The research in UTCN focused on: (a) the analysis of the effect of different configuration settings of the GlottHMM vocoder on the extracted glottal parameters (particularly on pitch extraction as a main factor in the synthesis stage), (b) the influence of the size of

pulse library on the quality of synthetic speech, and (c) the creation of a new Romanian voice with the GlottHMM technique using a small amount of speech data from an audiobook automatically annotated with the alignTK tool created by UTCN . These experiments indicate, among other things, that the pitch estimation of GlottHMM may deteriorate especially if noisy speech is processed with a frame length that is less than 30ms. Tests conducted by varying the size of the pulse library indicate that increasing the number of pulses in the library does not in itself result in better speech quality, because many other factors also contribute to the signal quality.

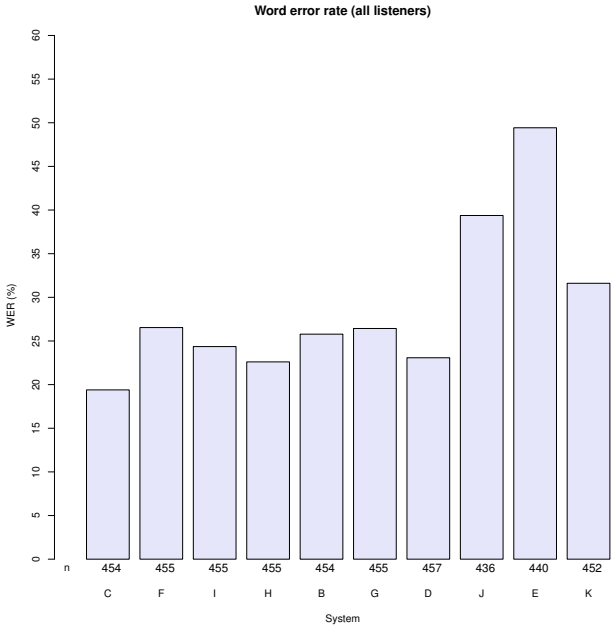

Figure 3.0c: Intelligibility results from the Blizzard Challenge 2012 (D – GlottHMM).

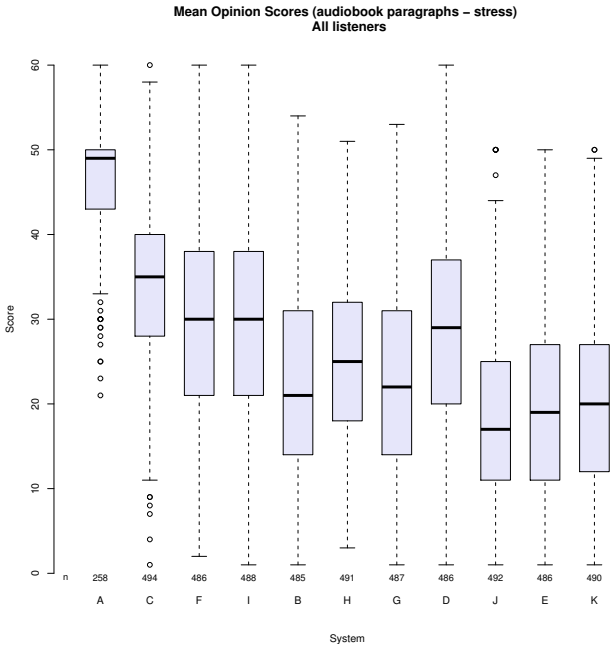

Figure 3.0d: Stress assignment results from the Blizzard Challenge 2012 (D – GlottHMM).

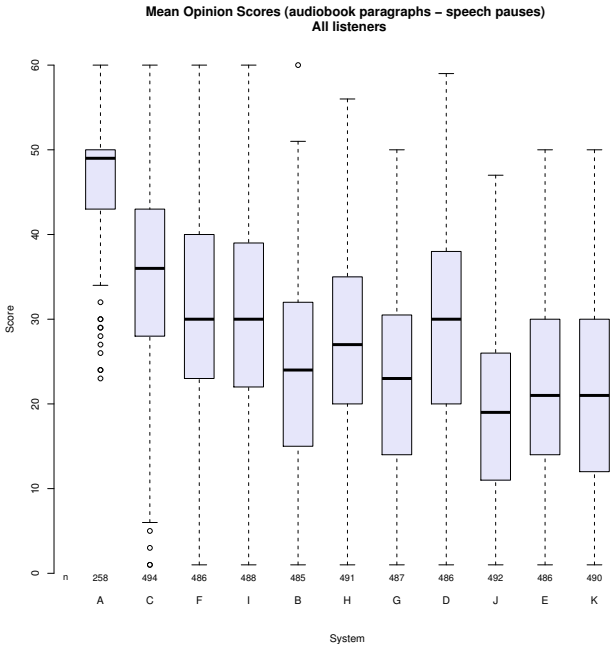

Figure 3.0e: Pause results from the Blizzard Challenge 2012 (D – GlottHMM).

Most recently, a study (not yet published) of the synthesis of speech across a wide vocal effort continuum and its perception in the presence of noise was conducted. Three types of speech were recorded and studied along the continuum: breathy, normal, and Lombard speech. Corresponding synthetic voices were created by training and adapting the GlottHMM system. Natural and synthetic speech along the continuum was assessed in listening tests that evaluate the intelligibility, quality, and suitability of speech in three different realistic multichannel noise conditions: silence, moderate street noise, and extreme street noise. The evaluation results were encouraging in showing that the synthesized voices with varying vocal effort were rated similarly to their natural counterparts both in terms of intelligibility and suitability (see Figures 3.0f and 3.0g).

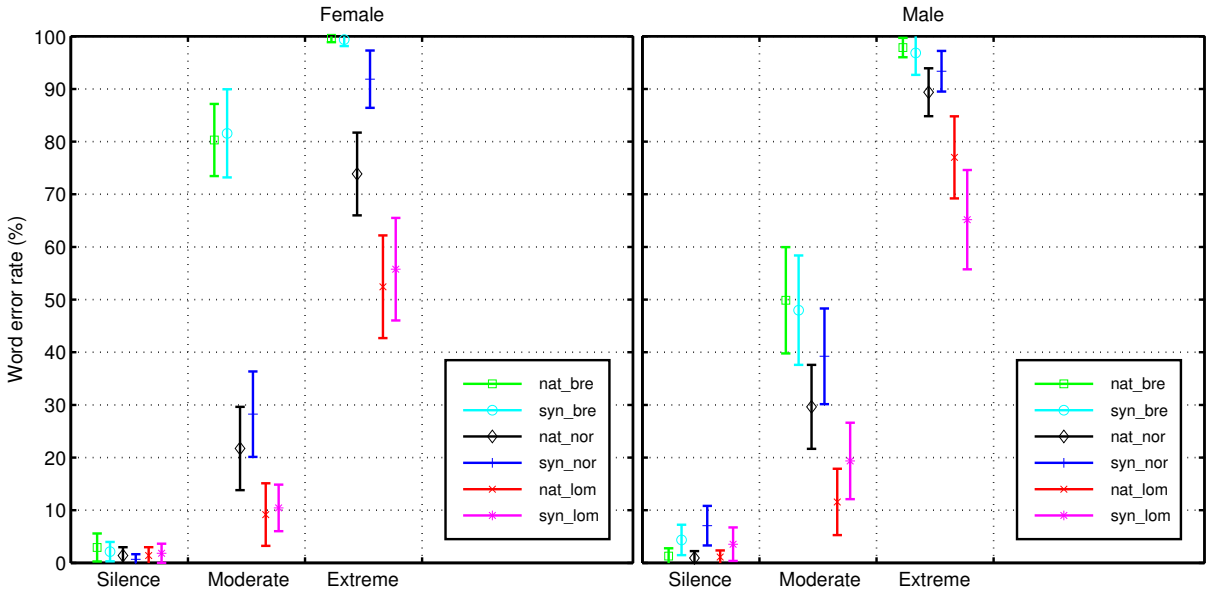

Figure 3.0f: Results of the intelligibility test for female (left) and male (right) voices in three noise conditions: silence, moderate street noise (63 dB, SNR $= -1$ dB), and extreme street noise (70 dB, SNR $= -8$ dB).

4 Conclusions

The GlottHMM vocoder is under constant development and evaluation in the project. New features of the vocoder have improved its capability of generating high quality and flexible synthetic speech and have enabled useful parameterization of speech in work aimed at the synthesis of expressive speech. The vocoder has been successfully used by other SIMPLE4ALL partners and intensive collaboration has proven to be especially useful in testing and evaluating the vocoder toolkit. The performance of the GlottHMM vocoder has been benchmarked in the Blizzard Challenge 2012 speech synthesis evaluation, showing that the GlottHMM system is among the top HMM-based speech synthesizers.

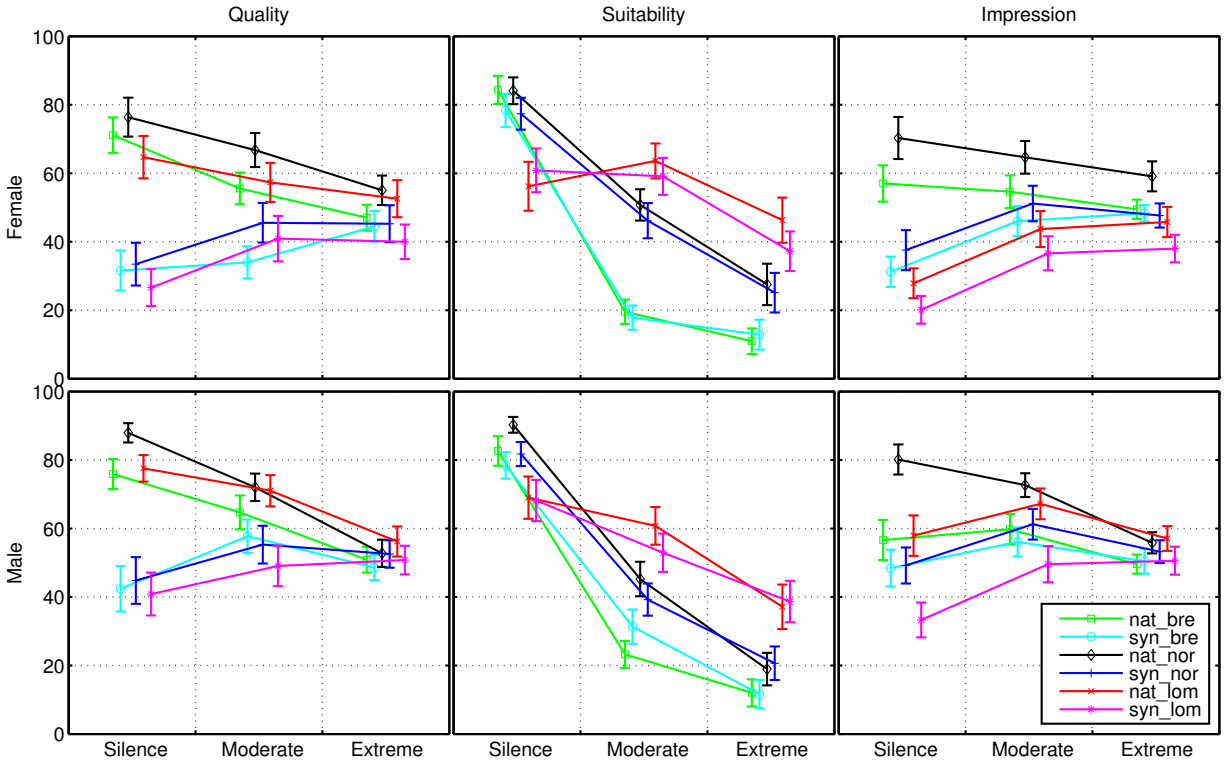

Figure 3.0g: Results of the subjective evaluation for female (upper) and male (lower) voices. The measured quantities are quality, suitability, and impression.

Publications

[1] Tuomo Raitio, Antti Suni, Martti Vainio, Paavo Alku: "HMM-Based Finnish Text-to-Speech System Utilizing Glottal Inverse Filtering". Proc. of Interspeech, Brisbane, Australia, Sept. 22–26, 2008.

[2] Tuomo Raitio, Marko Takanen, Olli Santala, Antti Suni, Martti Vainio, Paavo Alku: "On measuring the intelligibility of synthetic speech in noise – Do we need a realistic noise environment?". Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP'12), Kyoto, Japan, March 25–30, 2012.

[3] Tuomo Raitio, Antti Suni, Martti Vainio, Paavo Alku: "Wideband parametric speech synthesis using warped linear prediction". Proc. of Interspeech, Portland, Oregon, USA, Sept. 9–13, 2012.

[4] Antti Suni, Tuomo Raitio, Martti Vainio, Paavo Alku: "The GlottHMM Entry for Blizzard Challenge 2012: Hybrid Approach". Proc. of the Blizzard Challenge 2012 Workshop, Portland, Oregon, USA, Sept. 14, 2012.

[5] Jaime Lorenzo-Trueba, Roberto Barra-Chicote, Tuomo Raitio, Nicolas Obin, Paavo Alku, Junichi Yamagishi, Juan M. Montero: "Towards glottal source controllability in expressive speech synthesis". Proc. of Interspeech, Portland, Oregon, USA, Sept. 9–13, 2012.