Deliverable D5.1

Acoustic and prosodic analysis of genres and speaking styles

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	Acoustic and prosodic analysis of genres and speaking styles
Nature of the Deliverable	Report (R)
Dissemination Level	Public (PU)
This document is available from	`http://simple4all.org/publications/`
WP contributing to the deliverable	WP5 with contributions from WP2
WP / Task responsible	WP5 / T5.1
Editor	Martti Vainio (UH)
Editor address	martti.vainio@helsinki.fi
Author(s), in alphabetical order	Juan Manuel Montero Martínez, Roberto Barra–Chicote, Martti Vainio
EC Project Officer	Pierre Paul Sondag

Abstract

This report presents a collection of studies on acoustic analysis of speaking styles reflecting discourse genre, environmental factors, and emotion for the SIMPLE4ALL project. The purpose of the report is to identify the relevant acoustic and prosodic features for producing both neutral and expressive speech. Expressive speech includes both emotionally-coloured speech and speech that has been conditioned by environmental factors, such as noise, for example.

Contents

1 Intro

This report presents a collection of studies in which acoustic analysis is used to investigate the properties of various speaking styles reflecting spoken genres, environmental factors, and emotions for the SIMPLE4ALL . Since the project aims at producing synthetic voices in an unsupervised fashion, the acoustic features used for both training and controlling the synthesiser have to be accessible via automatic analyses. This naturally rules out not only hand labelling of data, but also probably labelling schemes that have been developed for use by humans (e.g., ToBI) rather than machines.

Different speaking styles relevant for speech synthesis – determined by text genre, speaking environment, and the speaker's emotional state – have complex consequences for the acoustic features of speech. The acoustics based on the vocal tract are, naturally, reflected in the segmental constitution of the messages in terms of vowels and consonants. The more relevant acoustic features related to speaking styles have to do with speech prosody and suprasegmental information.

There are two types of variability that affect the suprasegmental features of speech: intrinsic and extrinsic. Intrinsic variability has to do with utterance-internal prosody regarding chunking and prominence relations between words and is determined by the linguistic structure of the message, as well as variability determined by discourse functions; e.g., questions. Extrinsic variability, on the other hand, is determined by such factors as the environment (e.g., noise) and the emotional state of the speaker. Text genre-based variability is also extrinsic in nature as its effect on the speech is mostly global.

In this report we have divided the acoustic analysis into three separate parts: First, we present results for local, utterance internal prosody in terms of word prominence. Second, we present results from analysing speech in noise as opposed to speech in quiet. Preliminary analyses for near-whispered speech are included here. Third, we present acoustic analyses of different discourse genres as well as speech coloured by different emotions. The latter studies (Sections 4 and 5) were conducted by UPM and the first (Sections 2 and 3) by UH and AALTO .

2 Intrinsic prosody: Word prominence

Word prominence is a subjective feature that represents how listeners rate the relative salience of individual words in an utterance. It has been shown that people without expert knowledge of phonetics can still reliably determine the prominences in an utterance [1]. In this section we describe a study to determine the feasibility of prominence annotation by crowdsourcing.

The purpose of the experiment was to obtain information regarding the quality of our initial unsupervised annotation and the consistency between labellers. We were also interested in the number of distinct levels of prominence; i.e., can people can make four level distinctions consistently and, finally, how time consuming prominence annotation by naïve participants is.

The experiment utilised the unsupervised prominence tagging system described in Deliverable D2.1. The system employs an iterative scheme whereby a system is first trained with only limited prominence-related lexical information to serve as a sophisticated statistical reference. The word prominences are then estimated by aligning the training data with the output from the reference model and by calculating the differences between the two. Differences between all relevant prosodic parameters including the laryngeal ones are calculated. The prominences are then assigned on a four value scale from 0 (typically unaccented function words) to 3 (emphatic accent).

Altogether 900 utterances from a single female speaker were automatically tagged and then checked by ten participants who had no specific training in prominence tagging. Each utterance was checked by three people, resulting in four labels for each of around 7000 words in the data: the automatic label plus three human labels. The results were collated, along with information relevant for predicting the values.

With regard to the time required by the labellers, the results show that the participants on average labeled < 4 words/minute. However, these were phonetics students and may have taken the task too seriously! In any case, the system may be too time-consuming to be implemented as part of a new language development cycle. However,

crowd-sourcing can probably be used to obtain multi-lingual data for training a fully automatic labeler.

The participants agreed with the automatic labelling fairly well: 44.6% of the time all three labellers were in agreement, and 77.4% of the time two out of three agreed. Only 1.25% of the disagreements were bigger than one category and 26.4% of the disagreements were between adjacent categories. However, there was still over 20% disagreement between labellers and the algorithm. The discrepancy may be due to several factors that need further studying. For instance, it has been shown that abstract linguistic features (such as word order or word class) influence human prominence judgements [2, 1], whereas the automatic labelling depends solely on acoustic features and may thus be more accurate in phonetic terms.

As has been shown before, naïve listeners are able to make fairly reliable judgements of word prominences in multi-word utterances. This ability can be exploited in labelling prominences to construct prosodic training data. One of the main findings of the experiment is that labellers did not make any systematic labelling errors. The consistency of the obtained data should be helpful in further development of the automatic labelling scheme. Since this is ongoing work there are several open questions that need to studied in order to make the labelling scheme as efficient as possible. First, the experiment should be replicated using several languages in order to gain "universal" data. Second, a new, albeit more restricted study should be conducted with delexicalised speech since many linguistic variables are bound to affect prominence judgments outside the signal-based cues available to our automatic algorithm.

3 Speaking style: Context aware synthesis

With respect to different speaking styles, the dimension (or continuum) determined by the environment is the most important. Environmental noise has characteristics that can mask the information in the speech signals. Noise typically causes people to raise their voice in a manner that is amenable to modelling in speech synthesis – e.g., Lombard speech synthesis. On the other hand, a lack of noise and proximity of the speaker to the recipient causes changes in the speaking *style*. Here we summarise two studies that looked at the effect of different noise types and noise levels on speech prosody [3] and another one which was done when developing speech synthesis in a so called vocal effort continuum; that is, from quiet, near whispery, breathy speech to noise induced Lombard speech [4]

3.1 Effect of noise on laryngeal features

In Vainio et al [3] the emphasis was on the voice fundamental frequency (f_0) and the effect of different noise types as well as noise levels on the signaling of linguistic factors, namely prosodic focus. The main finding of the study was that f_0 does not simply increase as a function of louder speech, but that the main features signaling the stress and prominence relations in the utterances increase as a function of noise level and to a lesser degree vary with noise type. The three noise types used in the study were white noise, babble noise and low-pass filtered white noise with a frequency band of 0-1 kHz.

Twenty-one participants (11 female) were recorded in an anechoic chamber with loudness-calibrated noise played through headphones. The participants' own voice was also played back via the headphones. The participants read twelve different sentences with three different prosodic focus conditions.

The results of the study (with regard to f_0 are shown in Figures 3.1a and 3.1b. In summary, the results show that it is worthwhile modelling Lombard speech in synthesis. However, there are no prosodically-relevant reasons to model speech in different types of noise. Both the babble noise and white noise serve as efficient maskers and cause speakers to behave in a consistent manner with regard to Lombard speech.

We further studied ca. 700 vowel tokens for each gender in terms of laryngeal excitation. For this, a subset of the sentences containing only long [ɑ] vowels was used.

Glottal excitation was estimated from the speech pressure waveforms of the separated vowel segments by utilizing the Iterative Adaptive Inverse Filtering (IAIF) algorithm [5]. Analyses were conducted in the "TKK

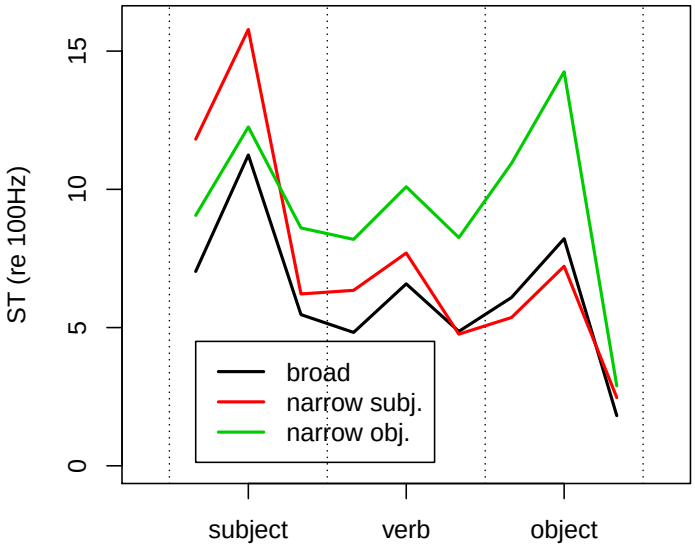

Figure 3.1a: Average f_0 contours calculated from three f_0 points per word for different focus conditions for all noise types and levels. Black = broad focus, red = narrow focus on subject, green = narrow focus on object.

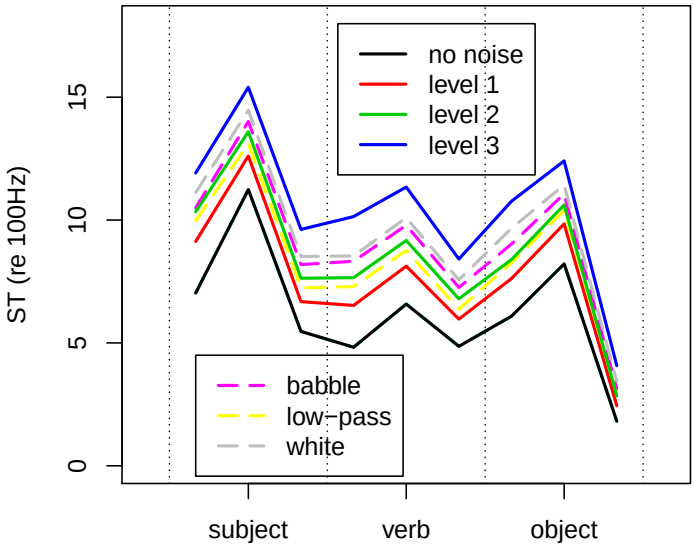

Figure 3.1b: Average f_0 contours for different noise levels and types. The dashed lines depict grand means in terms of different noise types; the solid lines stand for different noise levels.

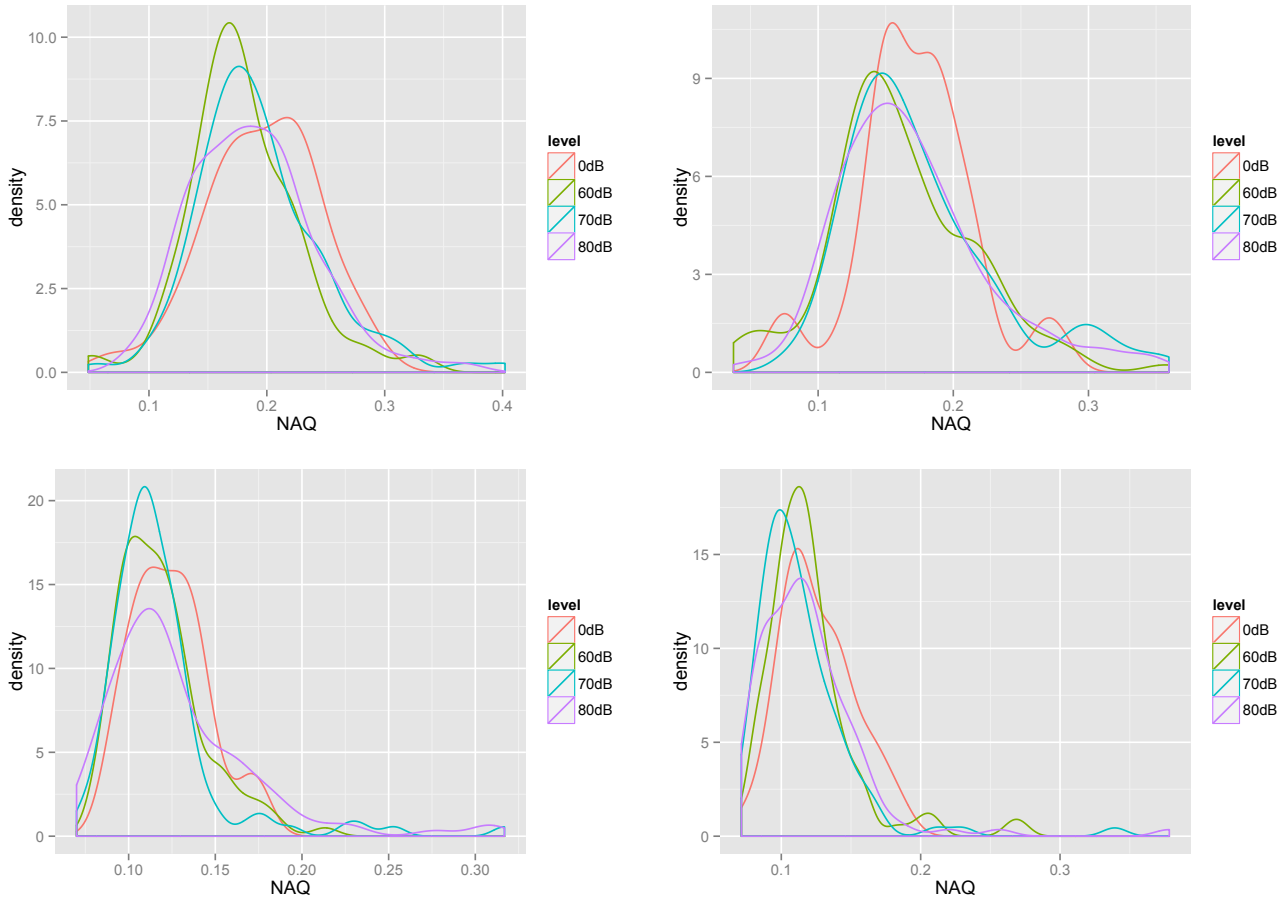

Figure 3.1c: Female (upper row) and male (lower row) NAQ distribution by noise level and the two word positions; subject (left column) and object (right column).

Aparat" environment, which enables the computation of the glottal source and the parameterization of the obtained waveform with a multitude of voice source parameters [6].

Because the estimation of the glottal source was to be conducted from challenging speech material represented by vowel segments cut from continuous speech, it was considered essential to select a parameterization method which is robust to distortions (such as formant ripple and aspiration noise). In addition, we expected the prosodic focus to be reflected by the behaviour of the glottal closing phase. Therefore, the voice source parameter to be selected was to focus on the parameterization of the time-domain features of the glottal source during its closing phase. Based on these rationales, we selected the normalized amplitude quotient (NAQ) [7, 8] as a parameterization method of the estimated glottal flow.

NAQ quantifies the time-domain characteristics of the glottal closing phase from two amplitude domain values: NAQ=A_{AC}/d_{min*T}, where A_{AC} is the maximum AC amplitude of the glottal flow, d_{min} is the minimum of the flow derivative, and T is the length of the fundamental period. It has been shown that NAQ correlates well with pressedness of voice [8]: a small NAQ value corresponds to a pressed phonation type where the relative duration of the glottal closing phase is short; while a large value of NAQ depicts a smooth glottal pulse typical in breathy phonation, where the relative duration of the glottal closing phase is long. In addition to NAQ, we also measured f_0 values from the estimated glottal pulse-forms. All values were calculated as means over the center 40 ms of the vowel span of the stressed vowel of the word.

Figures show the distributions of the NAQ parameter for each gender (females have a typically larger NAQ

values) in terms of noise level and focus type, as well as the position of the word in the utterance. As can be seen in the figure, both females and males have substantially lower NAQ values in the presence of noise (the red lines as opposed to other colours). Moreover, female speakers tend to regulate their voice quality more during the last word of the utterance. The smaller peaks in the distributions are due to different focus conditions.

We analysed the significance of the different factor using linear mixed-effects models (as implemented in the R "lme4" package). The NAQ values are strongly affected by the presence of noise ($t > 2.0$). However, the different noise levels did not differ from each other. The NAQ values for males are on average 0.05 lower than for females. There were also highly significant focus:word-position interactions ($t's > 4.0$) showing that the voice quality is controlled differently depending on the position of the word in an utterance.

3.2 Vocal effort continuum

The vocal effort continuum was directly studied in Raitio et al [4] using two speakers (one female, one male) by both acoustic analyses and synthesis. The effort continuum was modelled at two ends (breathy and Lombard speech) as well as in the middle (neutral). The data for the experiment was recorded in a sound-proof studio using headphones for both noise and speaker feedback. In the noise condition, ca. 80dB babble noise was used and the speaker feedback was kept constant. In the neutral case no headphones were used. In the breathy speech, headphones were used to increase the speakers' feedback and the participants were instructed to speak as quietly as possible but still using voiced speech – as opposed to whispering.

A prosodic analysis of the resulting corpora was carried out using 100 isolated sentences of each style. Both speakers' raw f_0 values were analyzed; the resulting distributions are shown in Figure 3.2a for the female and male speaker (left and right pane, respectively). The distributions are similar for both speakers with regard to the speaking style; the breathy style results in a relatively low average f_0 with a very narrow distribution, whereas the Lombard speech has a wider distribution with a relatively high mean. The normal speech is between the two extremes, but with a large overlap with the breathy style. The distributions of logarithmic segmental durations for both speakers were also studied. The female speaker has an overall slower speaking rate with an average duration of 80.4ms as opposed to the male with an average duration of 69.05ms. With respect to statistical significance, the male speakers distributions did not differ from each other between different speaking styles, whereas the female speakers' Lombard speech was significantly slower than the other styles (t-test, $p < 0.0001$) which in turn were not significantly different.

For all voice types, five sentences of the style-specific speech were used for extracting the pulse library. The number of pulses for calculating the mean was 1566, 1901, and 3562 for female, and 1279, 1488, and 1876 for male breathy, normal, and Lombard speech. The corresponding mean pulses for different vocal effort levels for the female and male speaker are shown in Figure 3.2b. The figure tentatively confirms the results from the NAQ analyses above, in that females tend to vary their laryngeal settings to a larger degree than males.

All in all the analyses using the glottal flow characteristics instead of measures than can be directly computed from the speech acoustics show that there is important prosodically- and environmentally-determined physiological changes that are important to model in speech synthesis.

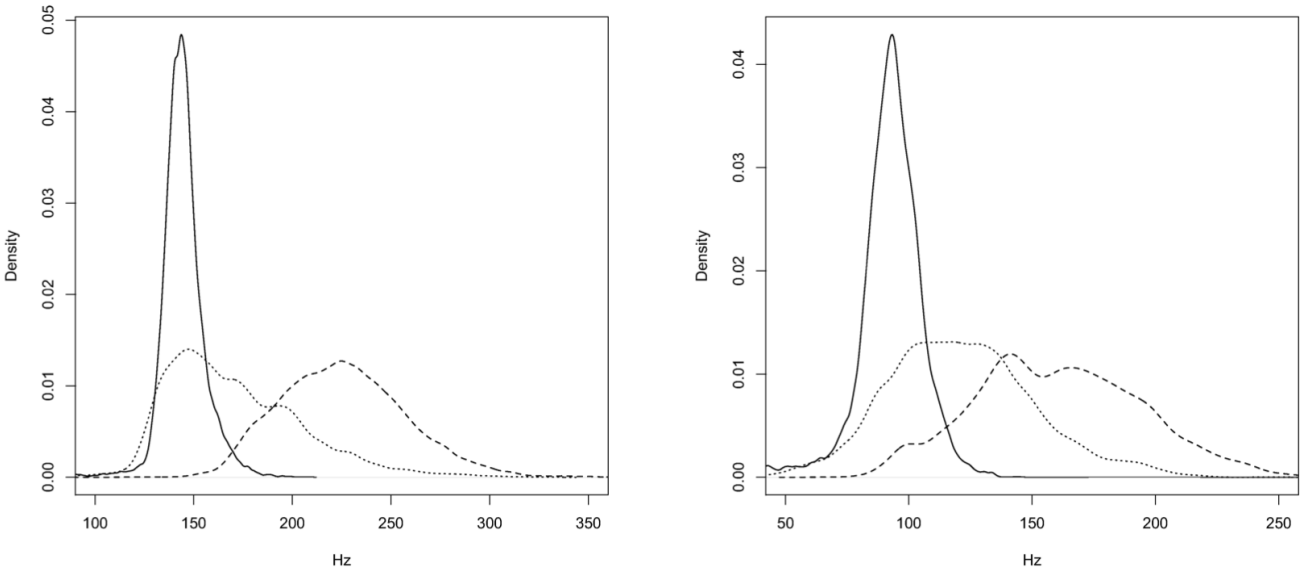

Figure 3.2a: Female (left) and male (right) f_0 distributions of the breathy (solid line), normal (dotted lind), and Lombard (dashed line) speech.

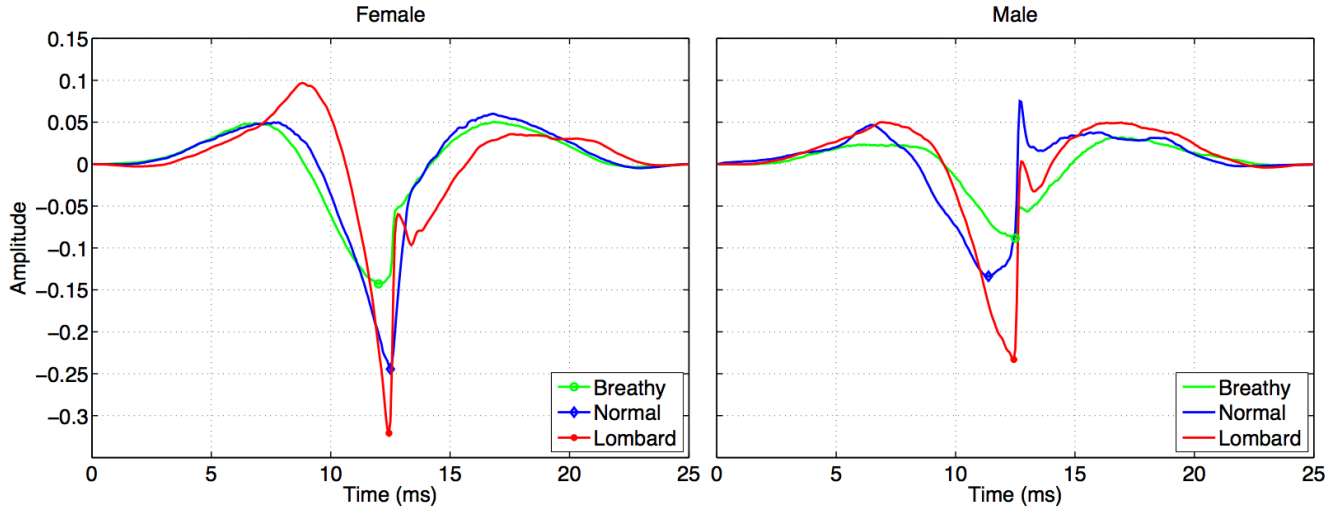

Figure 3.2b: Illustration of the mean of the windowed two-period glottal flow derivative waveforms (pulses) for different vocal effort levels for the female (left) and male (right) speakers.

4 Discourse genres and speaking styles

One of the deepest problems in current speech synthesis is the ubiquitous use of neutral speaking styles. These are 'designed' to be usable in any context, although in reality they are clearly inappropriate and unnatural in almost all real applications and listening situations. Conversational and expressive styles (as opposed to "read text") are therefore an important target in SIMPLE4ALL , as part of a suite of techniques and tools which can cost-effectively construct a speech synthesiser that fits a specific context of use.

Since one of the objectives in SIMPLE4ALL is the development of flexible models capable of producing a range of expressive or conversational speaking styles, this report will describe several speaking styles from prosodic and acoustic points of view, plus the results of our first experiments on speaking style identification.

Although one could argue that each speaker has a specific style, a kind of vocal signature which makes the speaker identifiable, this personal speaking style is modulated according to context in order to adapt it to that context. Different kind of texts (sometimes called discourse genres) have different lexical, syntactical and semantic features which make the genre identifiable, but they are also linked to a certain speaking style which best suits that genre [9, 10]. When reading a text or playing a role, speakers are able to identify the genre of the text, and they are also able to adapt their speaking style to the genre (most of the time at least) by adopting conventions generally associated to that particular genre.

The concept of genre has been part of Western literature and philosophy traditions at least since post-Socratic philosophers such as Aristotle in Poetics, and the concept of speaking style was already outlined by Aristotle in Rhetoric [11]. 20th century linguistics has dealt extensively with genre, discourse and style, mainly from qualitative point of view, relating genres and language functions [12].

In this section we present an analysis of several genres and their associated speaking styles as recorded in several speech corpora which will be briefly outlined. Then, some automatic speaking style identification experiments will be described, which used acoustic and prosodic features. Finally, we analyse some experiments on how the variability and diversity of speaking styles can affect other aspects of speech synthesis system creation, such as the diarization of training data.

4.1 Analysis of the IRCAM corpus

The IRCAM corpus [13] [14] is a multi-speaker transcribed, aligned and prosodically-annotated speech corpus especially designed for analysing and modelling speaking styles for French speech synthesis. The corpus contains four styles: Catholic mass preaching (subcorpus M), public political speeches (subcorpus P), journalistic broadcast news from the radio (subcorpus J), and live sports commentaries (subcorpus S). As the corpus has been the result of a careful selection, each subcorpus corresponds to a very specific genre and speaking style [15]. This one-to-one relationship between genre and speaking style makes the IRCAM corpus an exceptional resource for the analysis of speaking styles, despite being only monolingual. In addition to this intra-genre homogeneity, the corpus comprises male speakers only, in order to simplify the comparison between speakers.

The preaching and political sub-corpora are single-speaker monologues recorded in a natural environment. The first one was recorded from Christian sermons; the second one is based on a politician's New Year speech recorded specifically for broadcast, not in front of a live audience. Both sub-corpora contain no interaction [13]. The radio subcorpus contains press reviews and several types of chronicles: political, economical or technological. Most of the time, the subcorpus contains monologues, although there are a few interactions between the lead journalist and other participants. Finally, the live sports subcorpus involves two speakers engaged in quasi-monologues, commenting on a soccer match. The number of interactions is small, although there is some overlapped speech.

As the speech in the entire corpus was recorded in natural settings, audio quality is highly variable, with some background noise from the crowd or the audience in some cases, some recording noise, and room reverberation [13]. The corpus was especially selected to provide a relatively good balance of speaking-styles and speakers (between four and seven speakers per style), of duration per style (although sports commentaries are significantly

shorter: thirty-five minutes versus more than one hour for the other styles) and of duration per speaker (between nine minutes and fourteen minutes).

4.1.1 Prior work by Obin

The corpus was perceptually tested and assessed in [13] and the information in this part is a summary of his work. The aspects analysed first were the communication context and communication features. Following the conceptual scale proposed by Koch and Oesterreicher [13], three expert linguists described the situational context of the recordings using a three-degree scale. This scale contains ten aspects to be analysed: communication privacy, speakers' relationship, emotional strength, situational anchoring, referential anchoring in the situation, spatial and temporal distance, intensity of the communicative cooperation, dialogue vs. monologue, spontaneity and thematic freedom. The results of the analysis in [13] showed that broadcast news appeared to be the most formal style (the ten features were tagged as strictly formal: public communication, unknown speaker, weak emotions, situational detachment, referential detachment, great spatial and temporal distance, weak cooperation, monologue, prepared or scripted communication, thematically restricted) and sports commentaries seemed to be the most informal one (forty per cent of the features were tagged as informal: strongly emotional, action and situation anchoring, referential anchoring and spontaneity). However, all the recorded styles have some formal common features [13]: all of them were recorded from public media, thematically they are relatively restricted, the cooperation is weak, mostly based on relatively-independent monologues. Preaching and political speeches seem to be be very similar and mainly formal (more than sixty per cent features were tagged as purely formal). The main difference between these almost-formal styles was the distance to the audience (short for church preaching and long for the media-based political discourse).

4.1.2 Acoustic and prosodic analysis of the IRCAM corpus

Using the GlottHMM vocoder, we can analyse prosodic and acoustic properties of the IRCAM corpus, trying to determine the most important features which make each style identifiable.

Figure 4.1a shows the distribution of log f_0 and inverse of the speaking rate per style. In this carefully-designed multi-speaker corpus, the speaking styles are very pure and so exhibit significant differences between themselves. The highest f_0 in sports and church is due to the background noise in the stadium and the acoustic conditions of the church, which causes the speaker to speak louder, and consequently higher-pitched. As the sport event is a soccer match, the speaking rate in the sport style is very high, and this is followed by the broadcast news style (because the speakers try to maximise the amount of information conveyed per unit of time). Preaching and pre-recorded political speeches have lower speaking rates, because these styles do not try to transmit information as fast as the other styles.

Figure 4.1b summarizes these results in a prosodic map of the styles for the IRCAM corpus. In this map, the selected styles in the IRCAM corpus are clearly identifiable in spite of being a multi-speaker corpus, although some overlap exists. The green area in the lowest area is the area of the political speech, the area with lowest arousal (it is not a live event) and expressiveness (there is no live audience); the dark blue in the upper right corner is the area of preaching (pitch is high because of the acoustic environment and the speaking rate cannot be high because the discourse is at least partially improvised with some scripting); the light blue in the upper left corner is the sport area (conditioned by the event being described and by the adverse background noise to overcome); finally, the red area is the area of broadcast news (with a high speaking rate and some expressiveness, presumably to try to make the news more attractive).

In addition to these traditional prosodic features, one can analyse glottal features, such as those extracted by the GlottHMM vocoder. GlottHMM [16] is a vocoding technique developed for parametric speech synthesis in WP3 – see deliverable D3.1. It is based on decomposing speech into the glottal source and vocal tract through glottal inverse filtering. The vocal tract is parametrized as a Line Spectral Frequency (LSF) vector with 30 LSFs. The spectral tilt of the glottal source is also modelled using 10 LSFs. In addition, GlottHMM extracts the f_0 and

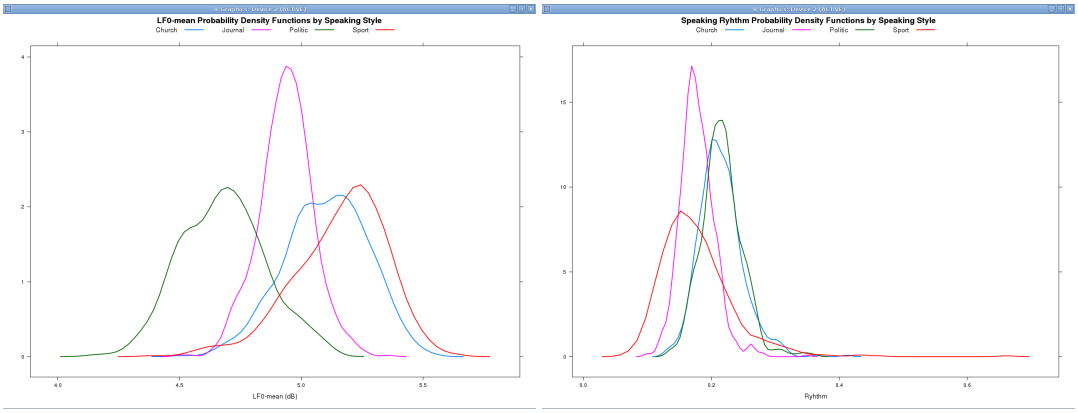

Figure 4.1a: *Log f_0 and the inverse of speaking rate, per style (IRCAM)*

harmonics to noise ratio (HNR) of the glottal source. The information of f_0 is used to separate voiced and unvoiced frames. The HNR measures the strength of the cepstral peaks averaged throughout their respective frequency bands, amounting to five in the present version. Finally, in addition to the standard features utilized in GlottHMM, we added two further features: Normalized Amplitude Quotient (NAQ) [8] and the magnitude differences between the first ten harmonics of the voice source.

Figure 4.1c, figure 4.1c and figure 4.1d show the distribution of average HNR3 and HNR5 and the variance of HNR4 and HNR5 per style, respectively. Although recording conditions may be different for the different styles in this corpus, news and political speeches were recorded in similar recording conditions, and nevertheless there are significant differences (in both average values and variance) between these two styles, suggesting these glottal features are style-dependent.

Figure 4.1e shows the distribution of NAQ per style. NAQ seems not to be conditioned by recording conditions, as styles are grouped not by recording conditions but in a different way: preaching and news on one side, sports and politics on the other.

4.2 Analysis of the C-ORAL-ROM corpus

The C-ORAL-ROM corpus is a multilingual multi-genre multi-speaker multi-style corpus of spontaneous natural speech for the main Romance languages. The transcribed and aligned corpus was recorded from the media (public formal styles) or in informal private contexts. Some additional details about this corpus are described in [17] and in deliverable D1.1.

C-ORAL-ROM is an exceptional corpus for the purpose of our work on analysis of speaking styles, as it provides resources in several languages being used in the project, it comprises a whole range of styles representing the main possible ones and it contains many speakers which means that any conclusions are more likely to generalise to new data. Another relevant feature of the corpus is naturalness: the corpus has been recorded in natural contexts (not only for the public resources recorded from the media, but also for the private recordings). Finally, the last relevant feature is spontaneity: the scenarios or the texts have not been especially designed for recording the corpus, although some of the recordings are based on professional scripts created for the media, because this is the standard procedure in the media for these kind of situations (for example: broadcast news or weather reports).

C-ORAL-ROM was not designed for speech synthesis tasks: it is a broad collection of recorded genres and styles, and the relationship between genre and style is rather complex – not as simple as in the IRCAM corpus. Some items of the same genre or theme (for example, sports), exhibit very different styles for different languages (sports news versus sports event commentaries in Portuguese, for example) or even for the same language (soccer commentaries versus cycling commentaries in the Italian subcorpus). The structure of C-ORAL-ROM is mostly thematic, rather than grouping by speaking style, therefore some style selection must first be carried out in order to

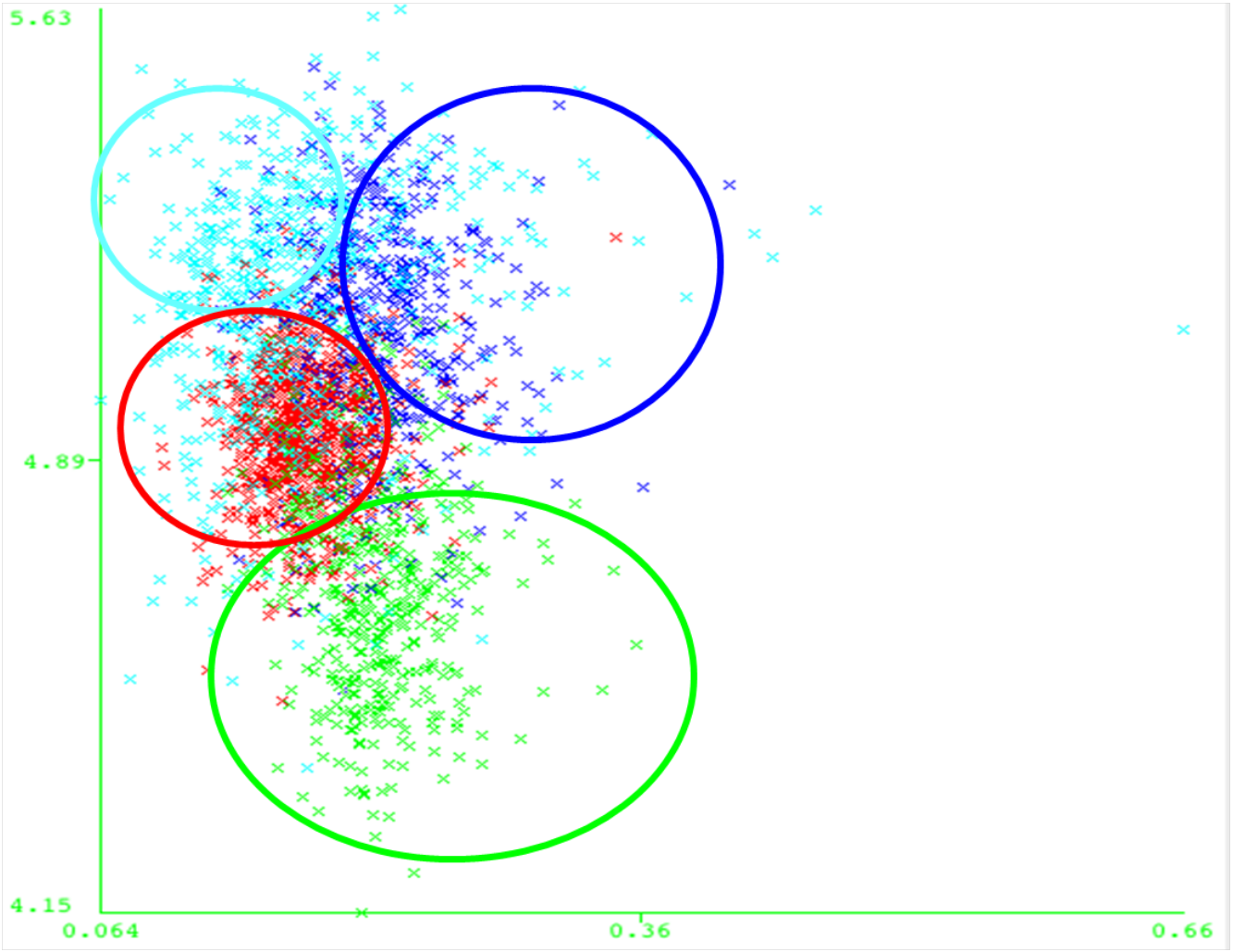

Figure 4.1b: *Prosodic map of the speaking styles in the IRCAM corpus. The horizontal axis shows the range of the log of the inverse speaking rate in syllables per second, and the vertical axis shows the log f_0 range.*

obtain homogeneous data sets for analysis.

4.2.1 Communication analysis of formal styles in C-ORAL-ROM

Although the developers of the corpus divide the subcorpora into formal and informal ones, a careful analysis reveals more details. The subcorpora tagged as formal (interviews, weather reports, broadcast news, reportage, scientific press, sport news or events, talk shows, political speeches or debate, preaching...), are fortunately mainly formal. They are characterised by being public, most of the times the listeners or the speakers are unknown (media monologues), emotional strength is relatively weak (except for live sport events, some talk shows and some political debates), there is situational and referential anchoring, they are thematically restricted, with some spatial and temporal distance to the audience (except for interviews, debates, talk show and preaching), communicative cooperation is rather weak (except for talks shows, debates, and, partially, in interviews), mostly based on monologues (even in interviews) and mostly with some preparation (when not fully scripted, except for sports and, partially, in interviews).

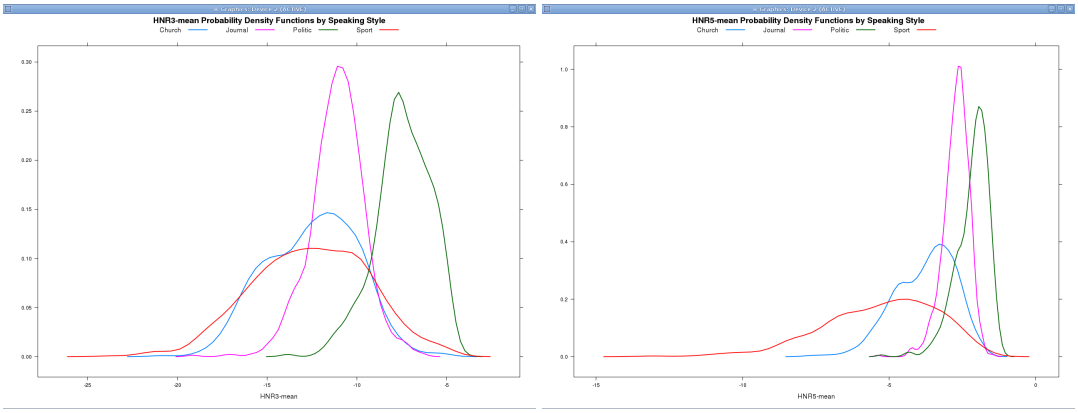

Figure 4.1c: *Distribution of average HNR3 and HNR5 per style in the IRCAM corpus*

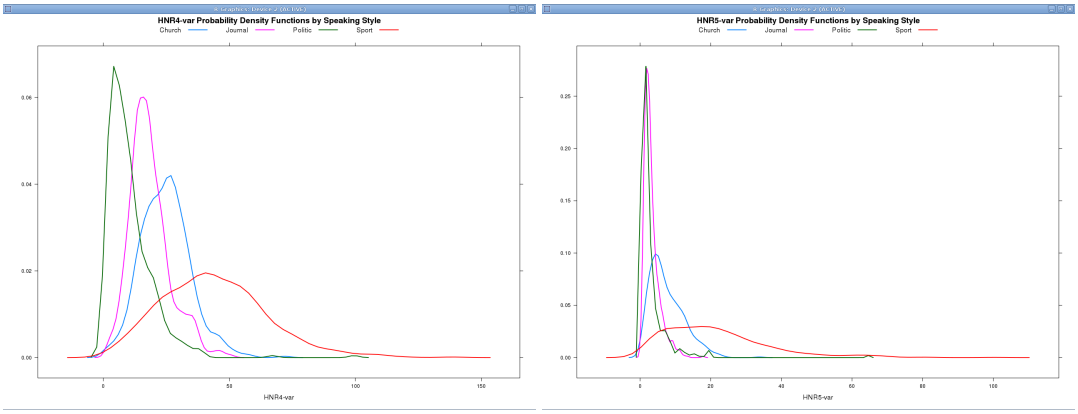

Figure 4.1d: *distribution of HNR4 and HNR5 variance per style in the IRCAM corpus*

4.2.2 Prosodic and glottal analysis of formal styles in C-ORAL-ROM

Figure 4.2a shows an f_0 boxplot per style, language and sex. Looking at the male data in the lower part of the figure, the speaking style of the interviews, news and talk shows is quite similar across Spanish, Italian and Portuguese, suggesting some possible language independence and therefore the possibility of having a cross-language model for these styles in the media.

Sports recordings show a great cross-language variability and the speaking styles associated with sports seem to be very different (in spite of the thematic similarities), because the genres recorded in different languages appear to be very different, ranging from live soccer broadcasting in Portuguese, to sport talk shows in Spanish (with some soccer broadcasting) or cycling broadcasting in Italian (live sport broadcasting but with a lower number of events to report per unit of time). The natural speed and spontaneity and the background noise of the events are completely different, causing the speaking styles to be significantly different.

The Spanish scientific recordings contain interviews in a very quiet situation, conditioned by the theme, but mainly by the interviewers. Finally, Italian reportage is really a scripted quiet documentary style, quite different from the non-scripted and more spontaneous speaking style in the Spanish and Portuguese recordings, explaining the different prosody of reportage in the Italian subcorpus.

Figure 4.2b shows the f_0 distribution per style and sex in the C-ORAL-ROM Spanish subcorpus. Focusing just on male data, sports is the most spontaneous style with greater arousal and emotional intensity (higher f_0), even in studio recording conditions. Interviews and scientific interviews are very similar and the genres with lower associated arousal. Finally, talk shows, reportage and broadcast news, seem to be mid-expressive. When compared

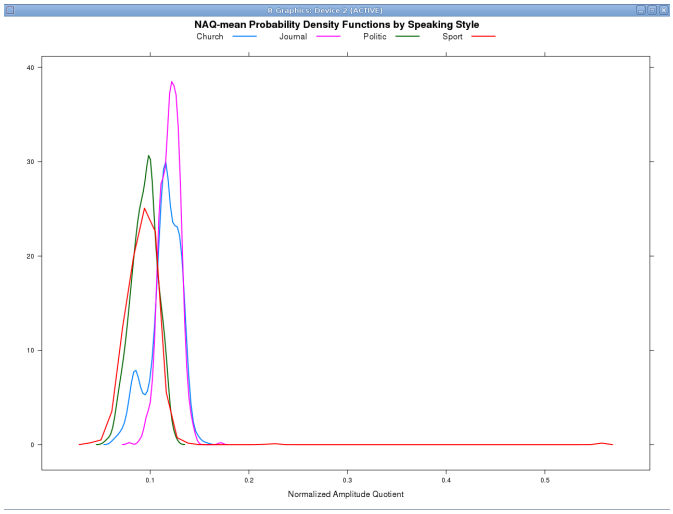

Figure 4.1e: *NAQ distribution per style in the IRCAM corpus*

to the IRCAM corpus and other languages in the C-ORAL-ROM corpus (such as Portuguese in Figure 4.2c), one important speaking style is missing: live sport broadcasting, with its high arousal and extreme prosody (opposite to the rather relaxed interviews).

Figure 4.2d shows the f_0 distribution per role and sex in Spanish broadcast news. The roles of professional male speakers such as conductor o interviewer exhibit narrower distributions than the other roles, showing the homogeneity of the style of this kind of speaker.

Figure 4.2e shows the f_0 distribution per role and sex in Spanish talk shows. Telephone callers exhibit a significantly different prosody, because of the acoustic conditions and the arousal of the caller. Other non-professional participants have also a certain tendency to a higher pitch.

Figure 4.2e shows the f_0 distribution per role and sex in the C-ORAL-ROM Spanish talk shows subcorpus. Calls over the telephone show the broadest distribution because this is the style dealing with more adverse acoustic conditions (less feedback when talking).

Figure 4.2f shows a boxplot of f_0 distributions per role and style in the C-ORAL-ROM Spanish subcorpus. Professional speakers have a lower average F0. In news, interviews and scientific press, professional speaker roles such as interviewer or conductor) consistently have lower average fundamental frequency when compared to other roles not played by professional media speakers (such participant or interviewed). However, in sports, the arousal associated to this topic can even reverse this relationship.

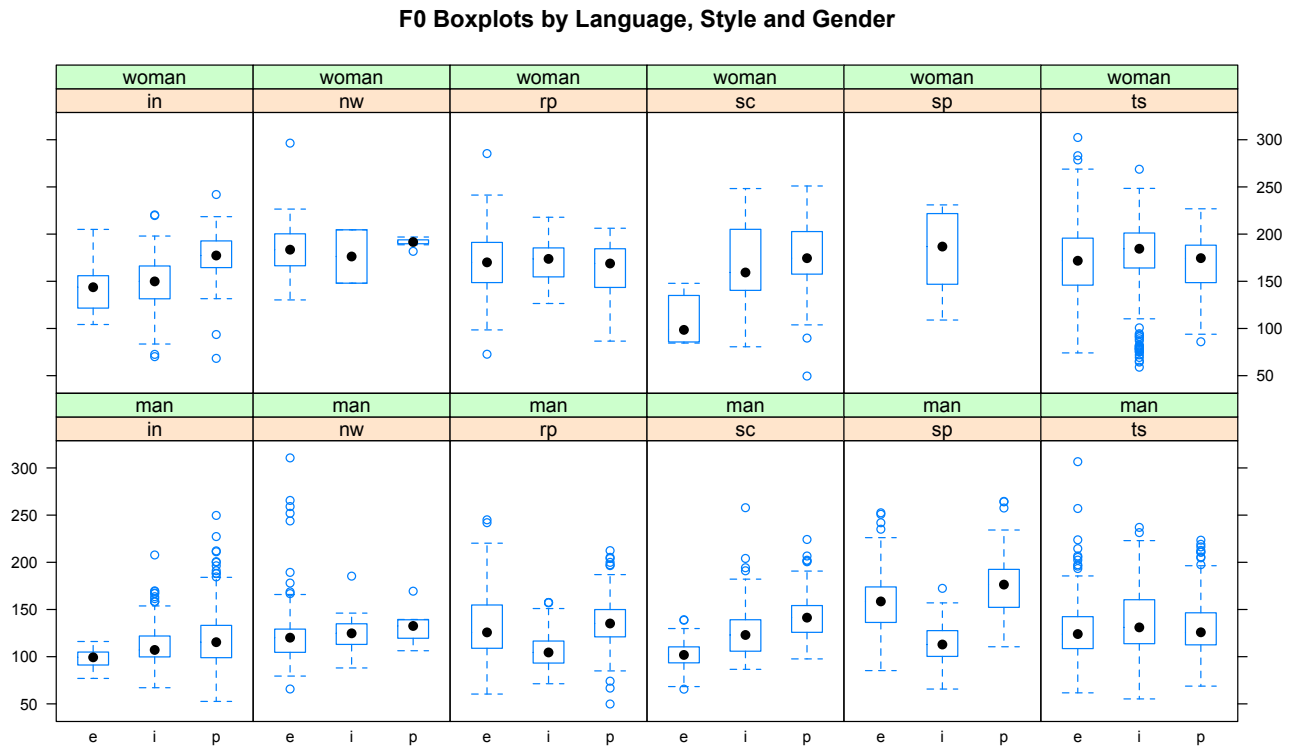

Figure 4.2a: f_0 *boxplot per style, language and sex in the C-ORAL-ROM corpus*

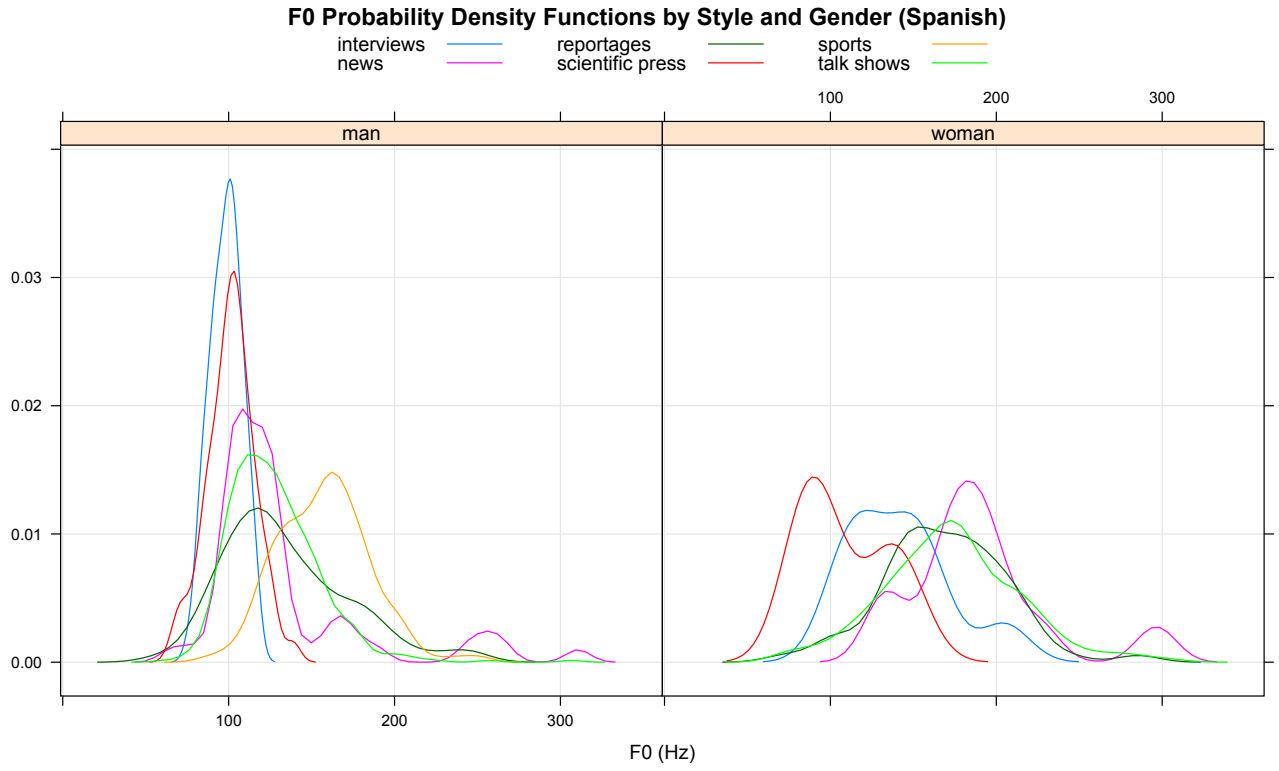

Figure 4.2b: f_0 *distribution per style and sex in the C-ORAL-ROM Spanish subcorpus*

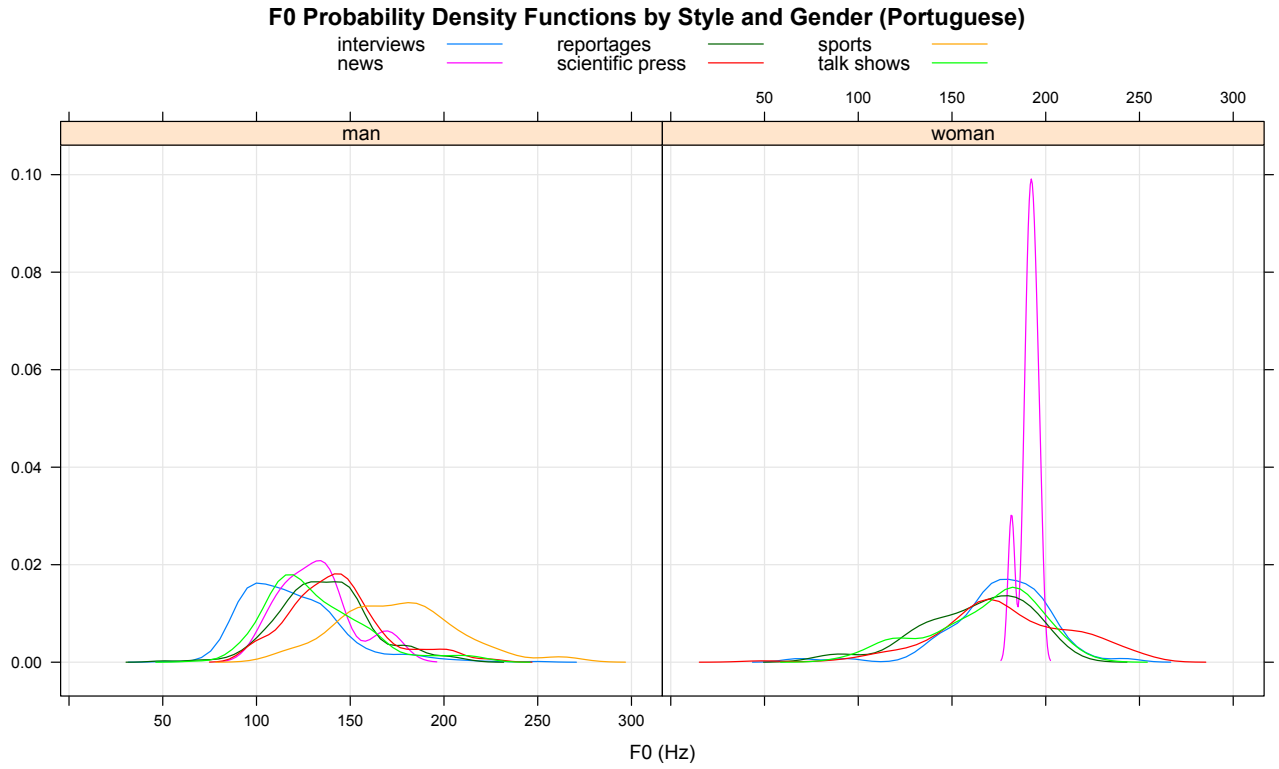

Figure 4.2c: f_0 *distribution per style and sex in the C-ORAL-ROM Portuguese subcorpus*

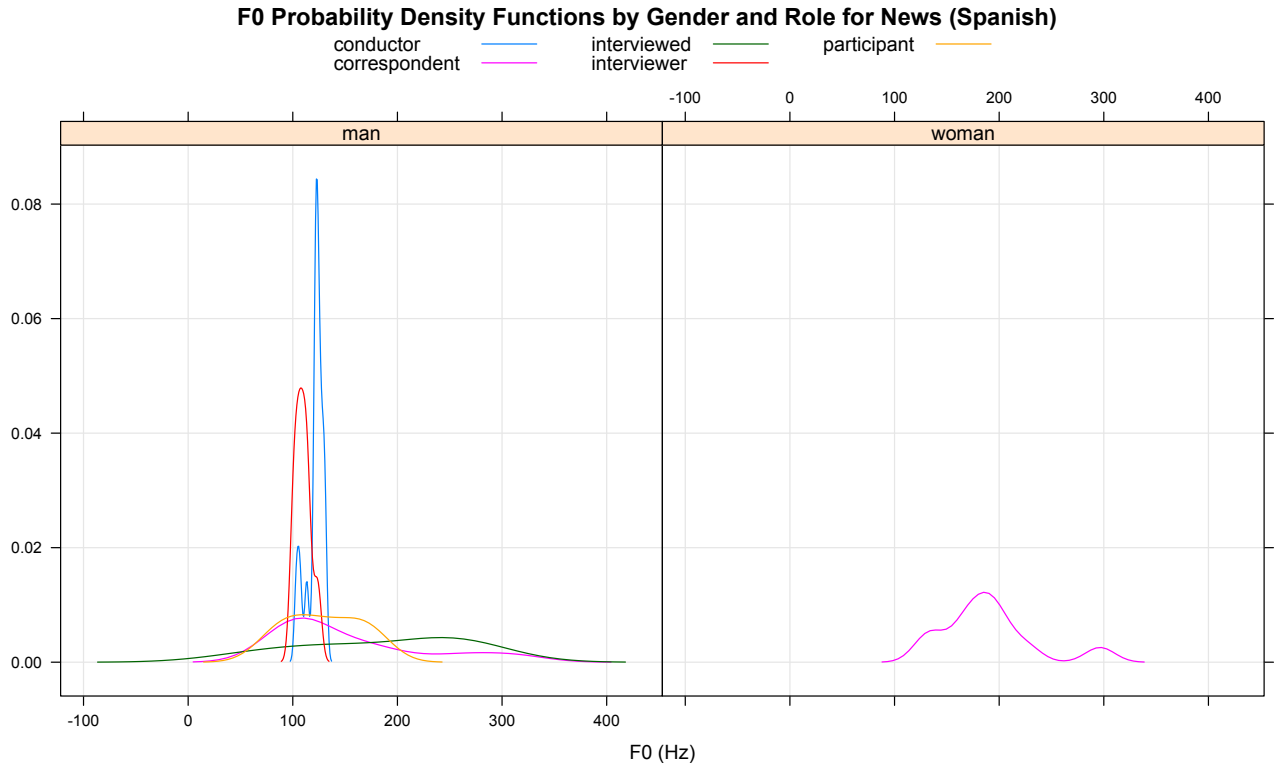

Figure 4.2d: f_0 *distribution per style and sex in the C-ORAL-ROM Spanish news subcorpus*

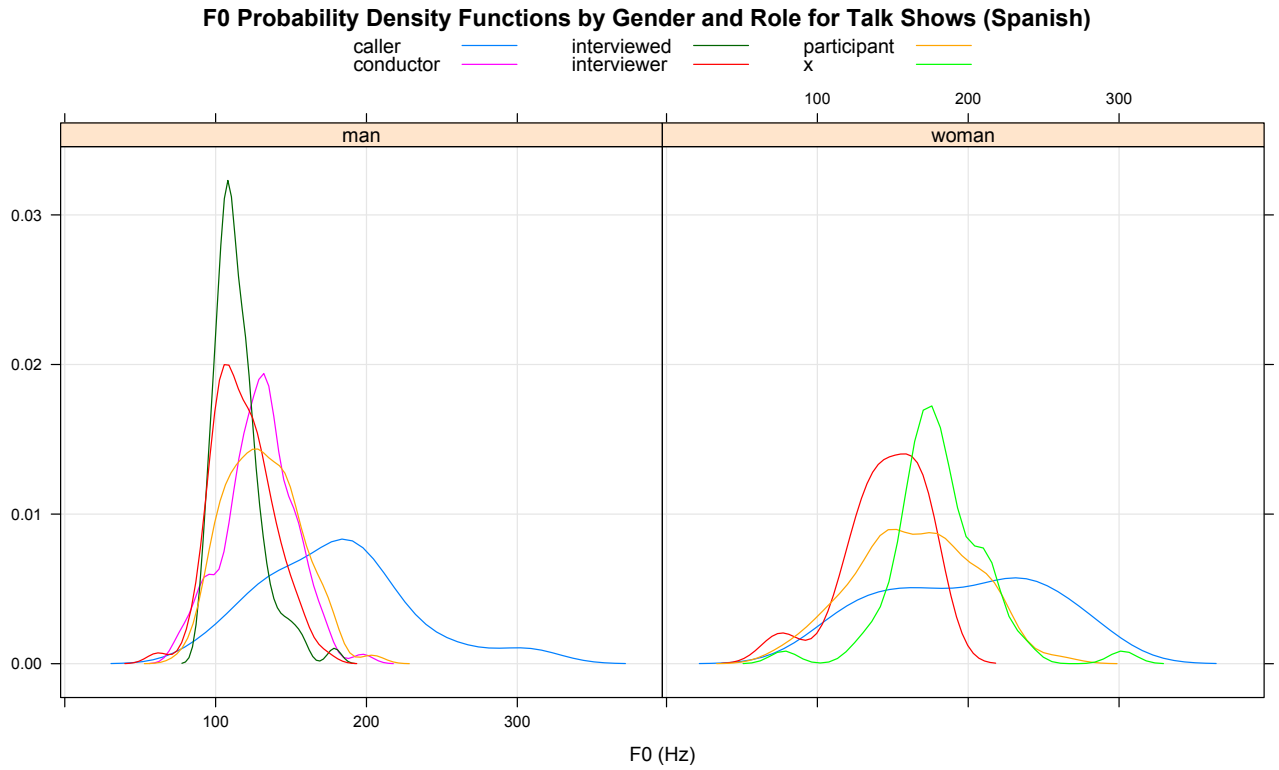

Figure 4.2e: *f_0 distribution per style and sex in the C-ORAL-ROM Spanish talk shows subcorpus*

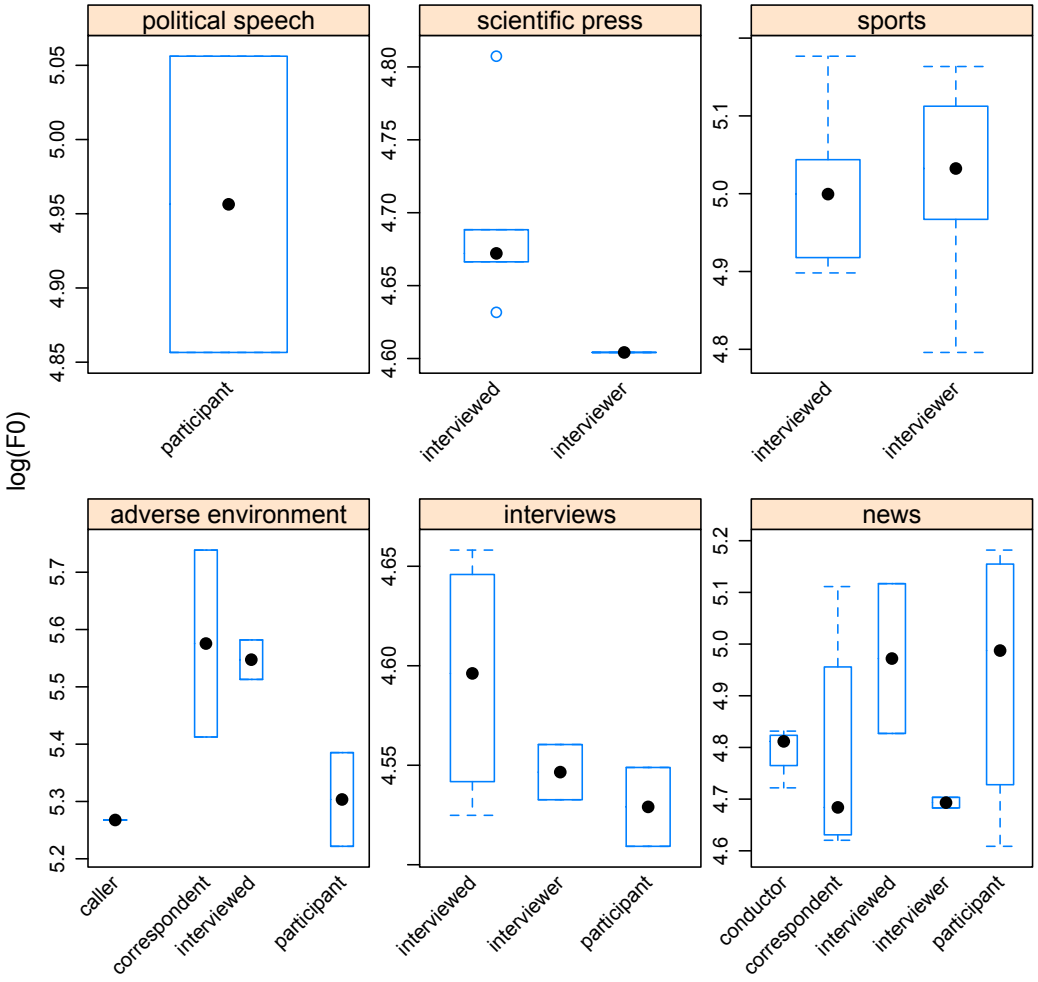

Figure 4.2f: f_0 *boxplot per role and style in the C-ORAL-ROM Spanish subcorpus*

4.3 Identification of speaking styles

4.3.1 Perceptual experiments (prior work of Obin)

A perceptual experiment was carried out on the IRCAM corpus in order to assess the subjective identifiability of the styles [13]. The recordings were delexicalised (some words are clearly associated to one specific style and should be removed from this multiple-choice test) by low-pass filtering, in order to avoid thematic cues. Short and long sentences were selected because the test should be balanced. Speech recording quality and volume were normalised to avoid any bias.

Speaking style identification rate was remarkable. The overall Kappa coefficient was significantly higher than random selection not only for the native speakers (0.58) but even for non-French-speaking listeners (0.26) [13]. Of course, some styles are more easily identifiable than others. For example, sports commentaries are clearly identified with great consensus (0.78). Radio broadcast news are mostly identified (0.63), while the more difficult styles are preaching and political speeches (0.48 and 0.45, respectively)[13]. It is important to notice that the test was in a sentence-by-sentence basis, not based on paragraphs.

The confusion matrix in [13] shows a certain degree of confusion between preaching and political speeches, matching the results of the communication analysis on formal versus informal styles. Maybe this finding is suggesting that these two styles are just two sub-styles of a more general style: public speech to an audience (either on a live event or in previously-recorded one).

This subjective experiment by Obin demonstrates that we will be able to select a certain set of identifiable styles for future modelling and synthesis purposes. In addition to this, he has shown that we could take selected recordings from the media, which is consistent with objectives of SIMPLE4ALL : to use publicly-available material (through the media or on the Internet), instead of using special-purpose studio recordings for creating new voices or speaking style models.

4.3.2 Automatic identification experiments

We have seen that people can identify some general styles even in a language-independent fashion through non-lexical supra-segmental cues. However, we have not yet proved whether a machine-learning procedure is able to identify those styles or similar ones in a fully automatic way.

In order to obtain more human-sounding speech synthesis we need expressive capabilities in the way of emotional and stylistic features so as to closely adapt them to the intended task. If we want to replicate those features, it is not enough to merely replicate the prosodic information of fundamental frequency and speaking rhythm. Given the tools developed in SIMPLE4ALL project, our proposal is to base these expressive capabilities on the modification of GlottHMM glottal parameters.

One must analyse the viability of such an approach by verifying that the expressive nuances are captured by the aforementioned features (as the perceptual subjective experiments suggest); we would wish to see high recognition rates (from automatic classifiers) on multi-style recordings and on emotional speech. Then we should evaluate the effect of speaker bias and recording environment on the source modelling in order to quantify possible problems when analysing multi-speaker corpora.

We will carry out this verification not only using prosodic features, but also glottal source parameters, known to be able to better capture the nuances of speaking styles and emotions [18]. This section analyses the viability and shortcomings of glottal source modelling for the identification of expressive speech, on the basis that high accuracy in such automatic recognition experiments is a necessary pre-condition before proceeding to use these parameters in synthesis. First, we provide recognition rates on the IRCAM corpus [13]. Finally, we suggest a style separation for the Spanish language based on the data from the C-ORAL-ROM corpus [17].

4.3.3 Prosodic versus glottal modelling

This modelling experiment aims to analyse the usefulness of glottal features for the identification and characterization of expressive speech. It is known [19] that speech production features extracted from the source carry relevant information for the expressiveness of speech that is not fully present in the prosodic features. If simply rely on traditional prosodic information such as pitch and rhythm, the identification rate is just about 74%. On the contrary, by using the GlottHMM features, we get an accuracy for the same style classification problem of 95.4% (Table 4.3a). These experiments were carried out using the SVM–SVO classifier in Weka, in a 10–fold cross–validation way.

Precision	Recall	F-Measure	Class
93.4	90.1	91.8	PREACHING
96.1	94.9	95.5	NEWS
95.6	98.7	97.1	POLITICS
96.3	98.9	97.6	SPORTS
95.4	95.5	95.4	Average

Table 4.3a: *Style recognition results for the IRCAM corpus, using glottal features.*

To confirm the hypothesis that expressiveness information is present in the glottal features, we applied an Information Gain analysis to all the features that showed which of the features are more individually relevant for the detection process. The results that can be seen in Table 4.3b show that there are a set of parameters even more informative than f_0, with speaking rhythm placed at a comparatively low position.

Table 4.3b: *Information gain of the best glottal features compared to prosodic features for the IRCAM corpus.*

Ranked	Feature	Ranked	Feature
0.8865	LSF2-mean	0.5962	HNR5-var
0.8097	LSF3-mean	0.5628	HNR4-var
0.7545	LSF1-mean	0.5239	LSF10-mean
0.6922	LSF4-var	0.5119	NAQ-mean
0.6892	HNR5-mean	0.5093	HNR3-mean
0.6031	LF0-mean	0.3194	Rhythm

These results do not imply that the pair of f_0 and rhythm will not produce a reasonable classifier, but they do show that considering glottal parameters for modelling expressiveness is a good idea. In fact, applying a greedy stepwise procedure to the features showed that, in addition to the six top features shown in Table 4.3b, an additional number of LSFs together with rhythm and NAQ lead to the optimal recognition rate.

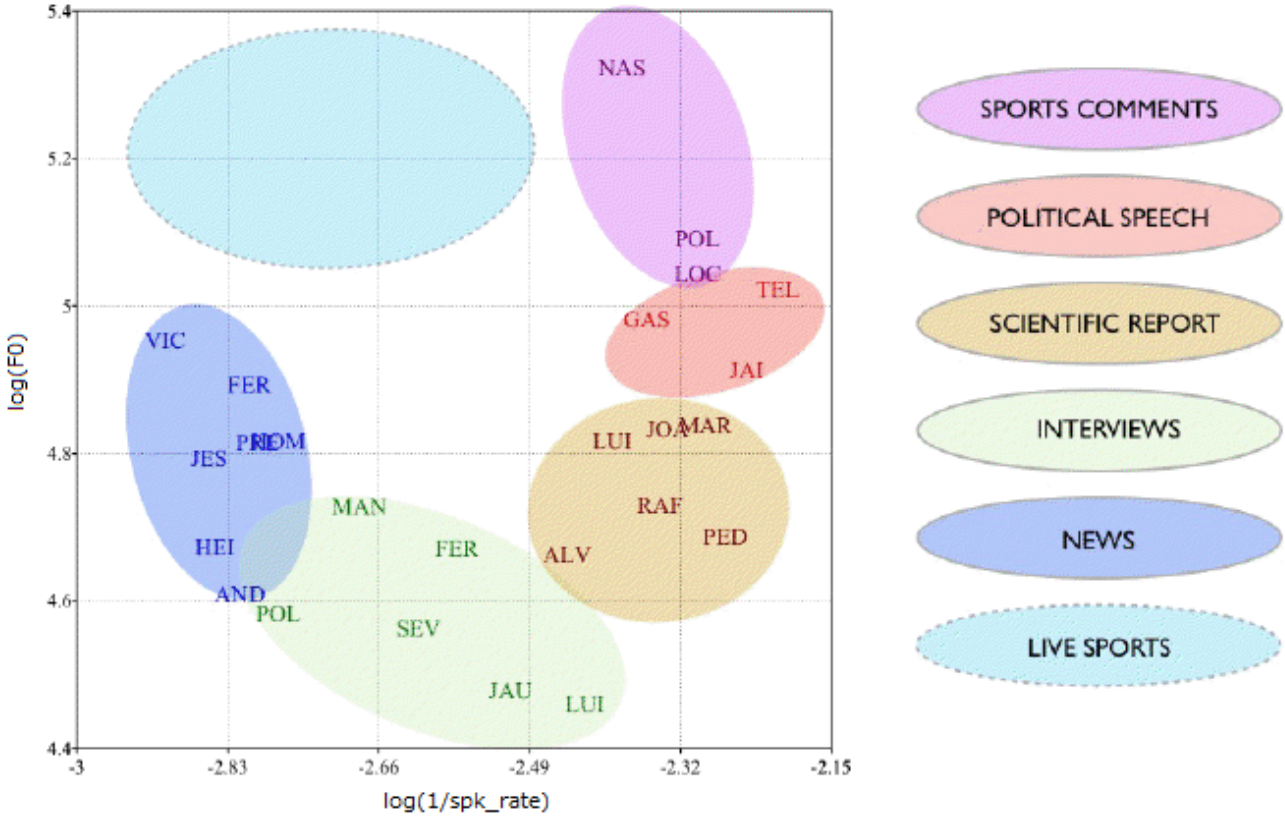

Figure 4.3a: *Proposed speaking styles distribution*

4.3.4 Spanish speaking styles space

The next experiment performed was to discover the definition of a set of separable styles in Spanish by analysing the C-ORAL-ROM speech. Because of the nature of the corpus, with many different speakers mostly uttering only a handful of sentences each, combined with highly variable recording environments, we decided to focus first on analysing two proven robust features: f_0 and speaking rate (there will be a future analysis of the new glottal features).

The first task was to remove from the corpus any non-separable styles and choose representatives for the separable ones. The chosen styles and their distribution can be seen in Figure 4.3a. This figure deserves some discussion on the meaning of the axes from a perceptual point of view: The f_0 axis correlates with the cleanliness of the speaking environment: the noisier the environment, the more the speaker will have to strain his or her voice and increase the pitch. An example of this would be live sports commentaries where the broadcaster will have to speak over the noise of the crowd (as live sports are hardly present in the Spanish C–ORAL–ROM, the drawn area is based on a selection of a few consistent examples of truly live sports commentaries, not summaries of sport matches). The opposing situation can be seen in news broadcasts, where the newscaster speaks from a studio in which the recording environment is perfectly controlled and quiet.

The speaking rate axis reflects the spontaneity of the speech: the more the speaker has prepared the speech, the faster he or she will be able to talk, as there is no need to pause and think the following utterance. The defining example is the news realm, where newscasters typically read the news in a style which allows them to fit as much content as possible in the allotted time. On the contrary, political speech shows much more improvisation and emotional load, introducing pauses, greatly reducing the effective talking speed. Finally, sports commentaries are not fully spontaneous, as the discourse is guided by the events of the sports to be commented, and those events

occur at high speed in many sports such as football.

4.3.5 Conclusions regarding style identification

In this section we have shown how the use of glottal model features greatly increase recognition rates of styled speech when compared to purely prosodic analysis, obtaining identification rates of 95% for styled speech. Finally we proposed a style separation for Spanish formal speaking styles that is based on considerations of spontaneity and environmental circumstances correlating with prosodic features (F0 and speaking rate).

4.4 Speaking styles and speaker diarization

4.4.1 Why diarization is required

One of the goals of SIMPLE4ALL is the automatic generation of appropriate speaking styles, and the approach we are taking to this is the modification of a neutral or expressive voice without needing to record a new speaker under each target expressivity condition, thus maintaining the quality of the original models but achieving a high style (or emotion) identification rate by listeners, and being able to control the intensity of the expressivity in a continuous fashion. A first step to being able to successfully control the speaking style of synthetic speech using this method, is to obtain enough data from speakers in different speaking styles, with which we could build speaking style-dependent average voice models.

Another goal of SIMPLE4ALL is to create the most portable speech synthesis system possible: one that could automatically (or with minimal supervision) be applied to many domains and tasks, which implies dealing with a wide variety of expressive situations and domains. In order to use speech collected from the media, or other 'found' data, speech synthesis systems must be robust to variation in the acoustic and environmental conditions. The system mush be able to robustly cope with noisy corpora and with challenging data such as interviews, debates, home recordings, political speeches, etc. The use of diarization techniques for speaker-turn segmentation and clustering is a useful processing step, because it allows the identification of homogeneous subsets of speech from heterogeneous recordings. Speaker diarization techniques are inherently unsupervised and language-independent; they can automatically label recordings with 'who spoke when' [20, 21].

The data processing pipeline proposed in D1.4 includes speaker diarization, where appropriate. However, there is (to the best of our knowledge) no existing analysis of how speaking style affects the performance of speaker diarization (i.e., a combined segmentation and clustering).

4.4.2 Speaker diarization vs. style diarization

It is important to note here that we are still addressing *speaker* diarization: the splitting of longer recordings containing multiple speakers into approximately single-speaker clusters. It would also be interesting to perform diarization to discover *style* clusters from multi-style recordings, but we have not tried this yet. A potential strategy for addressing this issue could be based on a multi-pass diarization process, where speaker diarization is carried out in the first pass, the resulting speaker models are used to normalize spectral and prosodic information (speaker normalization) in order to carry out a second pass where the speaking style diarization is addressed. In this second pass, the diarization process would be probably biased towards the prosodic information rather than the spectral information, which would be more relevant in the first pass. The experiments reported next concern speaker diarization. The goal is to examine how robust a standard state-of-the-art diarization system is to different speaking styles, and to consider what performance metrics are most appropriate for measuring this.

4.4.3 Data used in speaker diarization experiments

The data were taken from C-ORAL-ROM (media broadcasts of different stations), and they present a great deal of variability in the recording environments and a high number of speakers (up to 124). This results in some

Table 4.4a: *Features of the speaking style sessions in the C-ORAL-ROM corpus (ses. stands for session).*

Style	# ses.	SNR	#spk/ses.	time/ses.
interviews	5	25	4	8 min
meteorology	3	26	1	3 min
news	9	29	6	5 min
reportage	15	29	7	5 min
scientific press	4	27	5	9 min
sports	5	33	4	11 min
talk shows	12	29	5	8 min

speakers uttering only a few short sentences, making them fairly irrelevant from a statistical point of view. The main Spanish formal media styles will be analysed: *news broadcasts, sports, meteorological reports, reportage, talk-shows, scientific press and interviews*.

Long recordings in C-ORAL-ROM corpus were split into medium-length sessions. The number of speakers in each session is variable (between 1 and 9 speakers). The maximum length for a specific speaker in one session is 5 minutes. Table 4.4a summarises the average characteristics of the processed sessions for each speaking style.

4.4.4 Speaker diarization system

Clusters found in an unsupervised way (i.e., "pseudo-speakers") were generated using the UPM speaker diarization system described in [21]. The system has been adapted to this task in order to use only the feature streams modelling 19 Mel Frequency Cepstral Coefficients (MFCC) and the F0, but not the inter-channel delay feature vectors that are used when microphone array recordings are available. We are assuming that only one channel is available in the data to be used in SIMPLE4ALL . It carries out speaker segmentation and agglomerative clustering of segments, on speech that has already been filtered through a Voice Activity Detector.

4.4.5 Speaker Diarization Results

First, we evaluated the performance by measuring the Diarization Error Rate (DER). Table 4.4b shows that the system achieves a very low DER (lower than 5%) for the styles with speakers that have prepared their discourse, or have some kind of prompt (such as in *interviews*, *meteorology* and *news*). More spontaneous speaking styles (such as *reportage*, *sports*, *scientific press* and *talk shows*) obtain a moderate DER (between 10% and 25%).

However, the purpose of speaker diarization in SIMPLE4ALL is to identify clusters of speech data with high precision (or speaker purity) which contain enough speech to train a speaking-style average voice or to be used as a specific target voice. In this task, miss errors (MISS) are not as problematic as in other tasks (e.g., close captioning or meeting diarization) because these errors do not directly degrade the models – they will simply reduce the amount of data used (and probably by a relatively small amount); if, for very challenging data, a high MISS rate is the only way to achieve sufficient cluster purity for our purposes, this can be counteracted by collecting more data. Similarly, False Alarms (FA - where non-speech is labelled as speech) are probably also not too problematic, since we anticipate that the speech-text time alignment module will recover from these errors. Of the standard performance measures commonly used in speaker diarization, Speaker Error Rate (SER) is the most relevant to us.

4.4.6 Unsupervised Pseudo-Speakers for Speaking Style Average Voices

For our purposes, the larger and the purer (i.e., number of actual speakers present is as close to one as possible) a cluster is, the better. This is why the Speaker Error Rate (SER) commonly used in speaker diarization tasks [20] will in fact not be a good enough measure to evaluate performance, and why other metrics such as recall, precision

Style	MISS	FA	SER	DER
interviews	0.00	0.30	6.60	6.93
meteorology	0.00	0.40	0.70	1.14
news	0.00	0.30	4.40	4.66
reportage	0.00	1.40	22.40	23.78
scientific press	0.00	0.30	15.70	16.01
sports	1.30	0.20	11.80	13.27
talk shows	2.30	0.30	13.20	15.78

Table 4.4b: *Speaker diarization results (%) for each speaking style.*

and F-score (for each cluster) will be required to evaluate the available data and purity of each found cluster (i.e, pseudo-speaker).

Figure 4.4a shows the recall as a function of the size (i.e., duration of speech) of the agglomerated clusters. The recall of a cluster has been estimated as the ratio between the duration of the speech associated to the speaker that contributed the most speech frames to that cluster to the duration of all the speech in that session for that speaker. We obtain a recall in excess of 70%, which is high considering that this is an unsupervised task over realistic data recorded from the media, not using any phonetic transcription. The recall sharply decreases for clusters smaller than thirty seconds ($t <= 30$), indicating that these small clusters cannot contain all the speech for a single speaker.

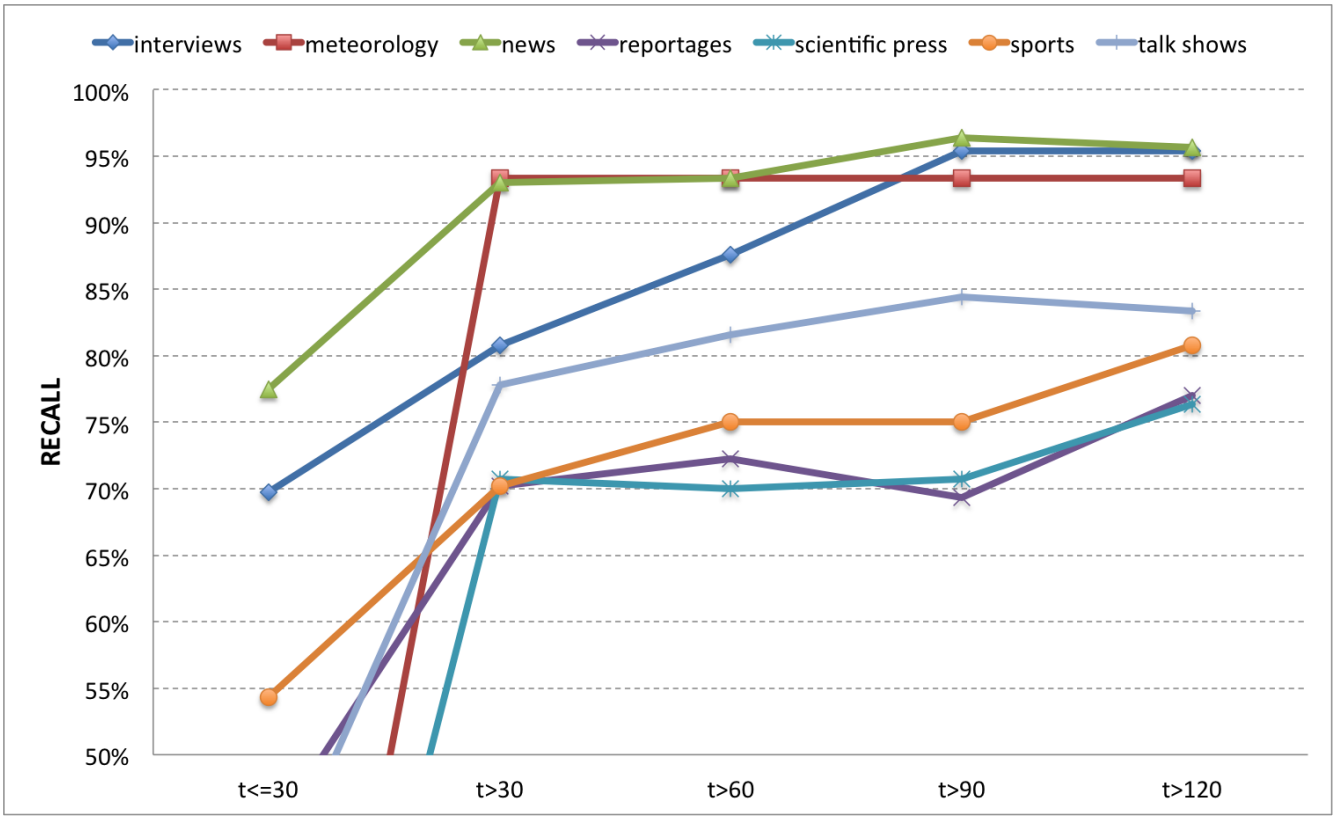

Figure 4.4a: *Recall (%) as a function of the size of the generated clusters (duration, in seconds) for each speaking style*

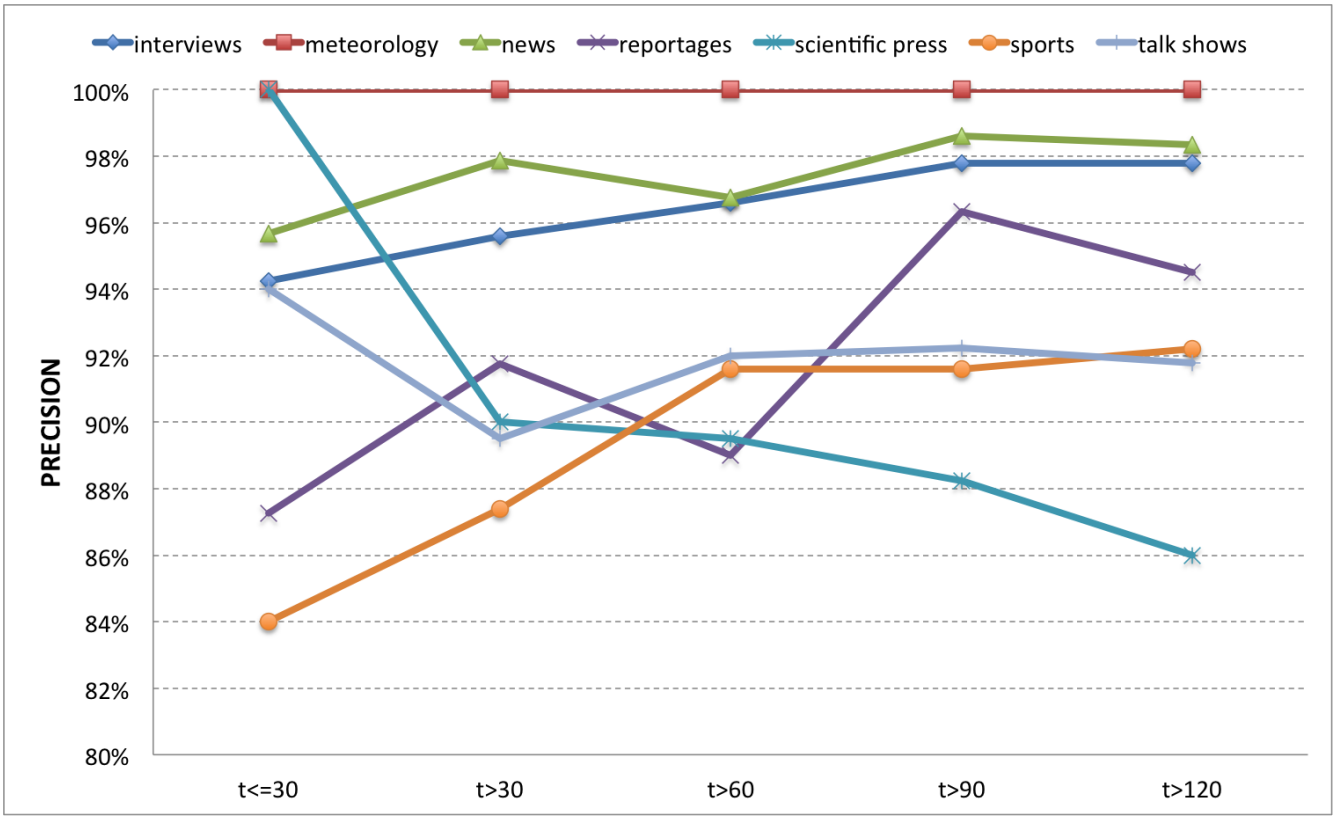

Figure 4.4b: *Precision (in %) as a function of the size of the clusters (in seconds) for each speaking style*

Figure 4.4b shows the precision as a function of the size of the clusters. High precision scores are obtained for all the speaking styles (greater than 85% for most lengths). *Meteorology* precision may appear trivial since there is only one speaker in each session; however, the diarization system has no prior information about the real number of speakers in each session and nevertheless it has been able to guess that there is only one speaker using purely unsupervised techniques. *Interviews* and *news* obtain better precision scores than the other speaking styles. Background music when certain speakers are talking in *reportage* and *scientific press* sessions may explain the lower precision for the longest clusters (especially in the case of *scientific press*. This suggests the need for a speech/music segmentation module [22] (as proposed in the SIMPLE4ALL framework) which, combined with the Voice Activity Detection module (VAD), should filter out music-corrupted speech data before diarization.

F-measure results shown in Figure 4.4c combine both recall and precision in one performance metric. When a cluster is longer than thirty seconds, recall is higher than 75% for every speaking style. This result suggest that a 30sec threshold should be applied before selecting clusters for use in training acoustic models. In addition to this, it has been verified that those clusters with high scores correspond to professional speakers (interviewers, journalists, etc.) or speakers that are used to speaking in the media or in public (such as politicians). On the contrary, *sports* is the most spontaneous style and most of the speakers (excluding the leading journalist) are not used to talking in the media, making them more difficult to diarize.

From these results, we see that the system performance is markedly higher when the speakers have prepared their discourses than when the spontaneity is higher.

We carried out an experiment to measure the correlation between the performance of the system and the Signal to Noise Ratio (SNR) or the number of speakers in each session: the performance is only weakly affected by these two features (Pearson correlation coefficients lower than 0.1).

It is expected that the precision (or speaker purity) of the pseudo-speaker clusters found here will be good

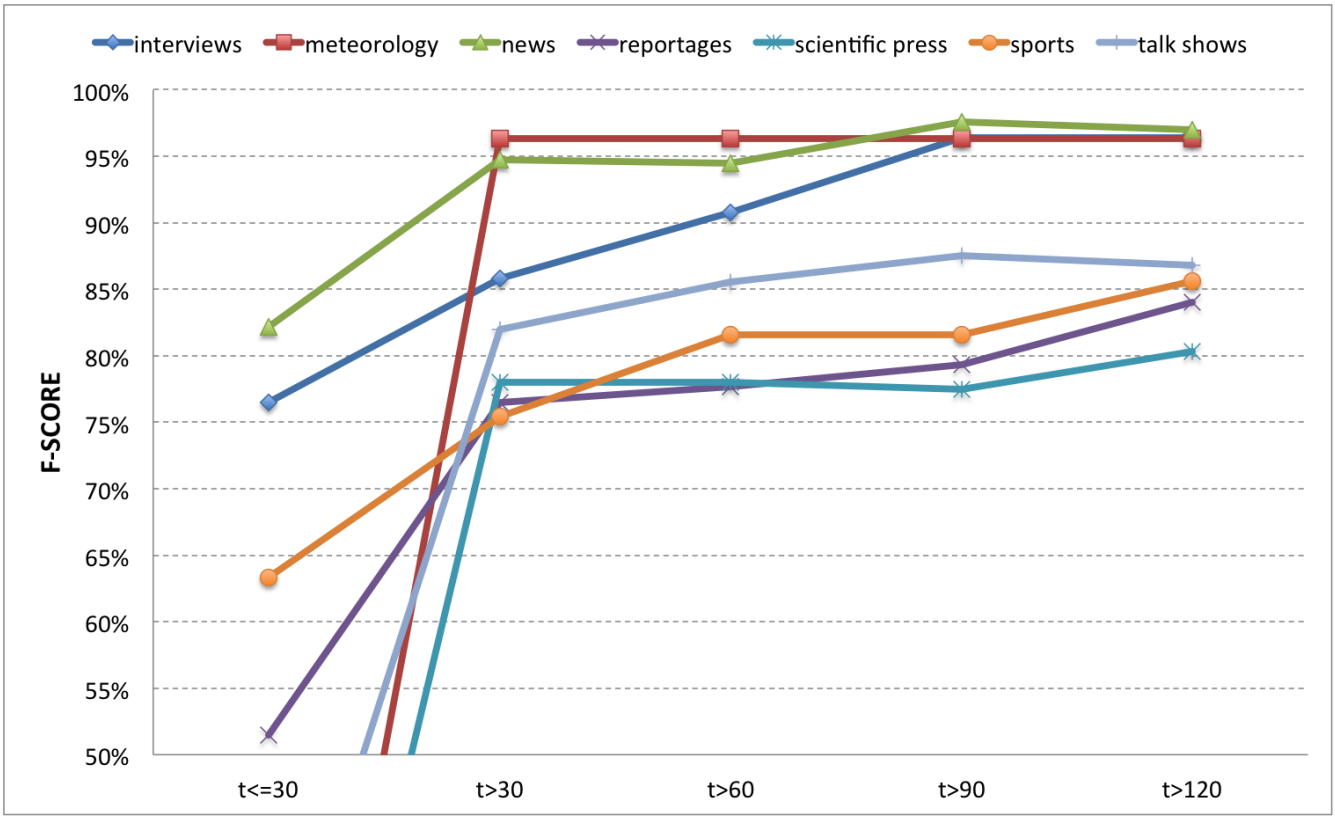

Figure 4.4c: *F-SCORE (in %) as a function of the size of the generated clusters (in seconds) for each speaking style*

enough to build accurate styled average voices with a high similarity with the original speaker and style. In further research it will be necessary to perceptually evaluate the similarity of the synthetic voices with the original speakers, and examine this as a function of the precision scores of the clusters used in the building process of every speaking style average voice.

4.4.7 Conclusions for speaking style and speaker diarization

We have found that a high performance is achieved by our speaker diarization system in terms of DER (average of 12%). However, DER is unlikely to be the best way to evaluate the quality of the pseudo-speakers found by the system, so we analysed performance in terms of recall, precision and F-score. We found that speaking styles with speakers used to talking in public (interviewers, politicians, etc.) tend to lead to pure pseudo-speaker clusters with high precision and recall (higher than 90%). Similarly, speaking styles in which the speakers read from text prompts or otherwise prepare their discourse, obtain higher F-scores than for more spontaneous speaking styles (sports, reportage and talk shows).

Eventually, it will be necessary to carry out a perceptual evaluation of the accuracy of the speaking style average voices and the similarity of generated expressive synthetic voices of the target speakers.

RANKED	FEATURE
0.6712	LF0-mean
0.50493	LSF3-mean
0.45145	LSFSOURCE10-var
0.43291	GAIN-mean
0.42299	HNR3-var
0.40886	HNR4-mean
0.4065	HNR1-mean
0.40348	LSF24-var
0.39706	HNR4-var
0.39682	HNR3-mean
0.39359	HNR2-mean
0.38533	LSFSOURCE1-mean
0.37928	LSF2-mean
0.37206	LSF23-var
0.36877	HNR5-mean
0.35747	LSF1-var
0.34863	LSF1-mean
0.34751	HNR5-var
0.34586	HNR2-var
0.33746	LSF22-var
0.33618	LSF25-var
0.32097	LSF14-var
0.31175	LSF24-mean
0.30936	LSF18-var
0.30461	LSF17-var
0.30411	LSF16-var

Figure 5.0d: *Feature selection for emotion classification*

5 Expressive speech: Spanish emotional voices

To complement the analysis of the power of glottal features to discriminate speaking styles from one another, we now analyse a specific type of expressive speech: emotional speech, and determine if the expressivity can be modelled in a speaker-independent way.

Emotions can be considered an extreme case of styled speech. SES and SEV, two Spanish corpora of emotional speech meet the requirement of being parallel multi-style and multi-speaker databases, recorded in a controlled environment, therefore can be used for an analysis of the power of the glottal features to discriminate emotional speaking styles. Both emotions and styles are aspects of expressive speech.

In addition to the earlier analysis on media-based genres, one can apply the GlottHMM feature analysis system to SES and SEV, two emotional speech corpora described in D1.1. Three speakers and seven emotional states were used: happiness (A), cold anger (E), hot anger (K), surprise (S), sadness (T), fear (M), disgust (C) and neutral (N).

Preliminary 10–fold cross–validation emotion recognition tests using J48 decision tress (as implemented in the Weka framework) showed a 81.2% recognition rate (see Table 5.0c) with the rank seen in Table 5.0d and Figure 5.0d. The problem then was to check whether the parameters and their distributions are grouped similarly for the different speakers. With that objective in mind we normalized every speaker's emotion with a z-score normalization algorithm respective to each of their speaker's neutral voice.

After normalization, we computed the Kullback-Leibler distance between all the multidimensional distributions and applied a MDS algorithm in order to project the distances into a two dimensional plane. The results of the scaling can be seen in Figure 5.0f; MDS of the un-normalized distributions is also provided in Figure 5.0e.

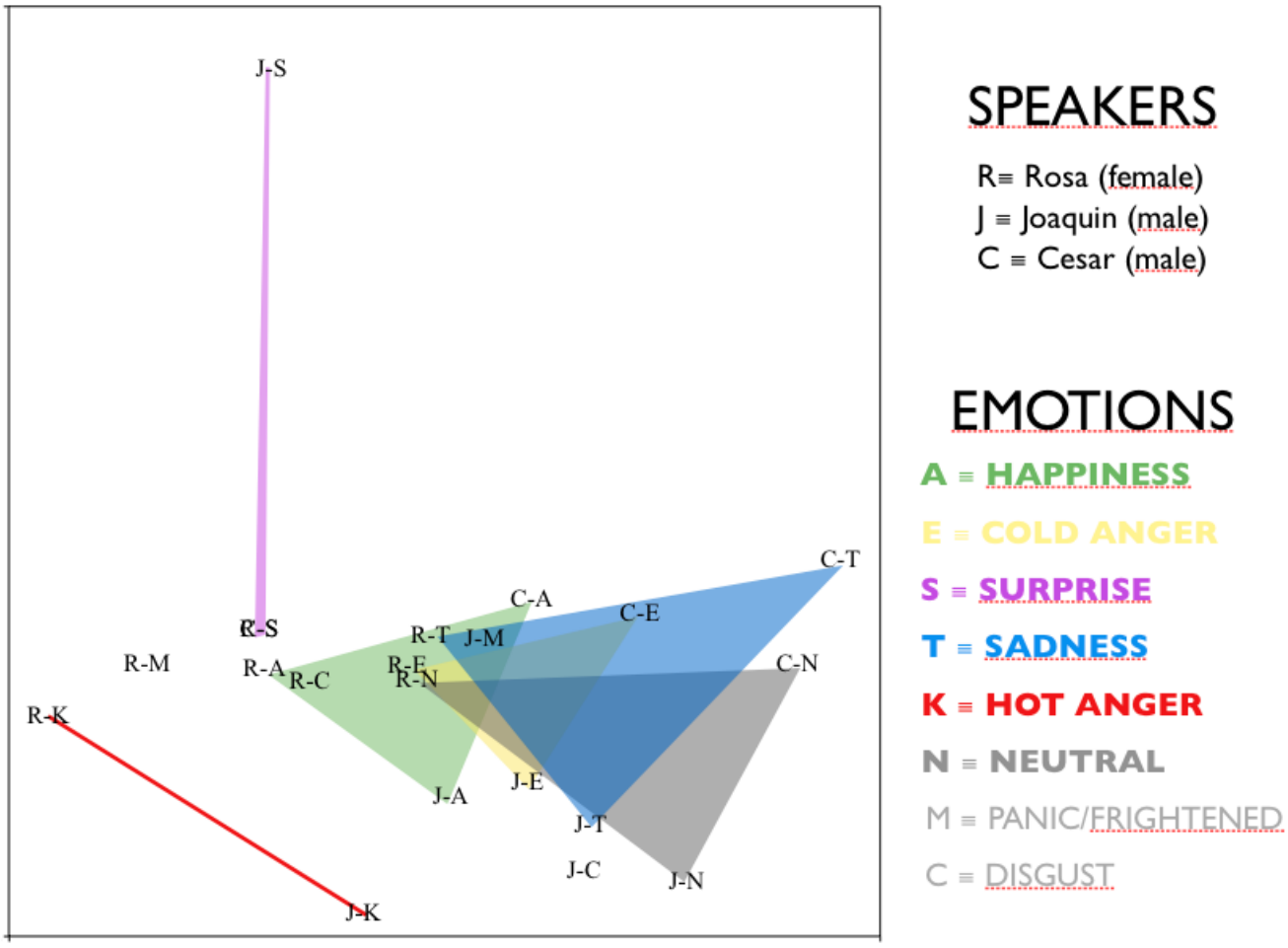

Figure 5.0e: *MDS of the non-normalized emotions*

Despite the moderate overlap between the neutral and sadness subspaces, the emotions present a clearly separable emotion space in which distances to the respective speaker's neutral voice is consistent across emotions. This results clearly support the theory that glottal features not only capture expressiveness information reliably but also that they are consistent between speakers, removing suspicions of bias. Additionally it is expected that normalization with an average neutral voice would help in the recognition process.

5.1 Conclusions on emotion identification

In this section we have shown how the use of glottal model features greatly increases recognition rates of expressive speech when compared to a purely prosodic analysis, obtaining rates of 82% for emotional speech. The usefulness of this approach was backed up by further analysis that showed that these glottal features do not suffer from speaker bias: our multi-speaker analysis showed clear distinctions between emotions, independent of speaker, when applying a MDS analysis. Although the results cannot be fully extrapolated to other speaking styles until we obtain and analyze a parallel corpus of several speaking styles from the same speaker or speakers, the approach is clearly promising.

Table 5.0c: *Recognition results of the emotional corpora using glottal features.*

Precision	Recall	F-Measure	Class
78.6	81.0	79.8	Neutral
86.6	86.4	86.5	Fear
80.9	80.9	80.9	Happiness
76.9	77.1	77.0	Disgust
89.0	88.4	88.7	Sadness
85.5	85.5	85.5	Surprise
72.1	70.1	71.1	Cold Anger
67.2	69.6	68.4	Hot Anger
81.2	81.2	81.2	Average

Table 5.0d: *Information gain of the best glottal features compared to prosodic features for SES and SEV corpora.*

Ranked	Feature	Ranked	Feature
0.8865	LSF2-mean	0.5962	HNR5-var
0.8097	LSF3-mean	0.5628	HNR4-var
0.7545	LSF1-mean	0.5239	LSF10-mean
0.6922	LSF4-var	0.5119	NAQ-mean
0.6892	HNR5-mean	0.5093	HNR3-mean
0.6031	LF0-mean	0.3194	Rythm

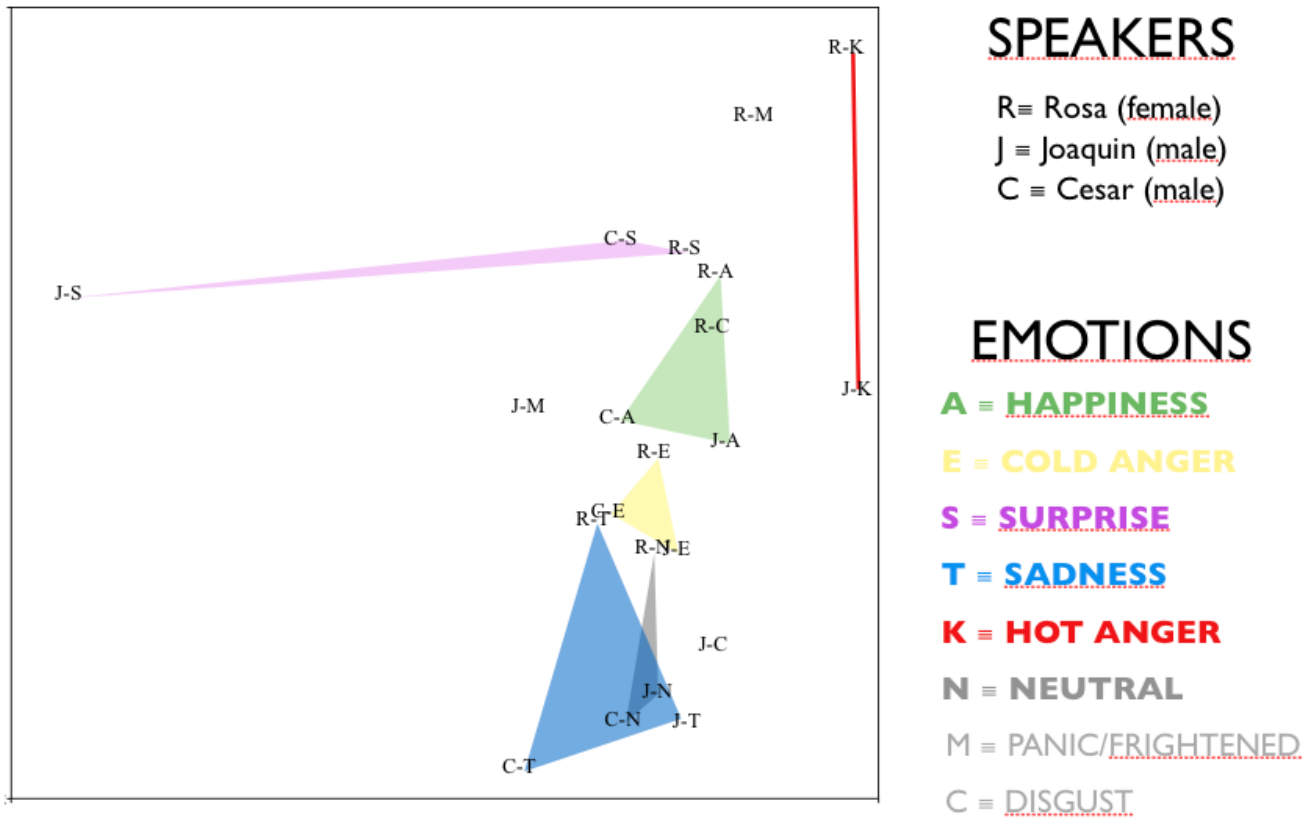

Figure 5.0f: *MDS of the emotions normalized by the neutral voice of each speaker*

References

[1] M. Vainio, A.S. Suni, T. Raitio, J. Nurminen, J. Järvikivi, P. Alku, et al. New method for delexicalization and its application to prosodic tagging for text-to-speech synthesis. In *Interspeech 2009 Proceedings of the 10th Annual Conference of the International Speech Communication Association, Brighton, UK, 6-10 Sept 2009*, 2009.

[2] Martti Vainio and Juhani Jarvikivi. Tonal features, intensity, and word order in the perception of prominence. *Journal of Phonetics*, 34:319 – 342, 2006.

[3] M. Vainio, D. Aalto, A. Suni, A. Arnhold, T. Raitio, H. Seijo, J. Järvikivi, and P. Alku. Effect of noise type and level on focus related fundamental frequency changes. In *Interspeech, Portland, Oregon*, 2012.

[4] T. Raitio, A. Suni, M. Vainio, and P. Alku. Synthesis and perception of breathy, normal, and lombard speech in the presence of noise. under review.

[5] P. Alku. Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering. *Speech Communication*, 11(2-3):109–118, 1992.

[6] M. Airas. TKK Aparat: An environment for voice inverse filtering and parameterization. *Logopedics Phoniatrics Vocology*, 33(1):1–16, 2008.

[7] P. Alku and E. Vilkman. Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering. *Speech communication*, 18(2):131–138, 1996.

[8] T. Beckstrom P. Alku and E. Vilkman. Normalized amplitude quotient for parameterization of the glottal flow. *Journal of Acoustic Society of America*, 112(2):701 – 710, 2002.

[9] Obin N. Lacheret, A. and M. Avanzi. Design and evaluation of shared prosodic annotation for spontaneous french speech: From expert knowledge to non-expert annotation. In *Proceedings of the Linguistic Annotation Workshop*, 2010.

[10] L. Degand and A.C. Simon. *Mapping prosody and syntax as discourse strategies: how basic discourse units vary across genres*, pages 81 – 107. 2009.

[11] Grard Genette. 1979.

[12] Roman Jakobson. Linguistics and poetics. 1960.

[13] N. Obin. 2011. PhD Thesis. Universite Paris VI - Pierre et Marie Curie.

[14] N. Obin, A. Lacheret, C. Veaux, X. Rodet, and A.C. Simon. A method for automatic and dynamic estimation of discourse genre typology with prosodic features. In *Proceedings of Interspeech 2008*, pages 1204 –1207, 2008.

[15] N. Obin, A. Lacheret, and X. Rodet. Expectations for speaking style identication: a prosodic study. In *Proceedings of Interspeech 2010*, pages 3070 – 3073, 2010.

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

[17] E. Cresti, F. do Nascimento, A. Sandoval, J. Veronis, P. Martin, and K. Choukri. The c-oral-rom corpus a multilingual resource of spontaneous speech for romance languages. In *Proceedings of 6th international conference on Language Resources and Evaluation Corpora*, 2004.

[18] D.G. Childers and CK. Lee. Vocal quality factors: Analysis, synthesis, and perception. *Journal of Acoustic Society of America*, 90(2):2394 – 2410, 1991.

[19] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis. In *Procedings of Eurospeech 1999*, pages 2374 – 2350, 1999.

[20] X. Anguera, Simon Bozonnet, Nicholas W.D. Evans, C. Fredouille, O. Friedland, and O Vinyals. Speaker diarization : A review of recent research. *IEEE Transactions On Audio, Speech, and Language Processing, February 2012, Volume 20, NÂ2, ISSN: 1558-7916*, 05 2011.

[21] J. M. Pardo, R. Barra-Chicote, R. San-Segundo, R. de Cordoba, and B. Martinez-Gonzalez. Speaker diarization features: The upm contribution to the rt09 evaluation. *IEEE Trransactions on Audio, Speech and Language Processing*, 20(2):426–435, 2012.

[22] A. Gallardo and R. San-Segundo. Upm-uc3m system for music and speech segmentation. In *Jornadas de Tecnologa del Habla FALA 2010*, November 2010.