Towards Glottal Source Controllability in Expressive Speech Synthesis

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

Jaime Lorenzo-Trueba[1], *Roberto Barra-Chicote*[1], *Tuomo Raitio*[2], *Nicolas Obin*[3], *Paavo Alku*[2], *Junichi Yamagishi*[4], *Juan M Montero*[1]

[1]Speech Technology Group, ETSI Telecomunicacion, Universidad Politecnica de Madrid, Spain
[2]Department of Signal Processing and Acoustics, Aalto University, Finland
[3]Sound Analysis and Synthesis, IRCAM, Paris, France
[4]CSTR, University of Edinburgh, United Kingdom

jaime.lorenzo@die.upm.es, barra@die.upm.es

Abstract

In order to obtain more human like sounding human-machine interfaces we must first be able to give them expressive capabilities in the way of emotional and stylistic features so as to closely adequate 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. The proposed additional layer is the modification of the glottal model, for which we make use of the GlottHMM parameters. This paper analyzes the viability of such an approach by verifying that the expressive nuances are captured by the aforementioned features, obtaining 95% recognition rates on styled speaking and 82% on emotional speech. Then we evaluate the effect of speaker bias and recording environment on the source modeling in order to quantify possible problems when analyzing multi-speaker databases. Finally we propose a speaking styles separation for Spanish based on prosodic features and check its perceptual significance.

Index Terms: expressive speech synthesis, speaking style, glottal source modeling.

1. Introduction

The task at hand is enclosed in the Simple4All project [1], which aims to improve synthetic speech with expressive cues to adapt the produced voice to the desired task. This is done by using not only prosodic features, such as fundamental frequency (F0) or rhythm, but also glottal source parameters, known to be able to better capture the nuances of speaking styles and emotions [6]. When considering expressive speech synthesis one should not forget that it requires a certain degree of adaptability automatically discarding unit selection based technologies in favor of HMM-based ones, as they are too rigid to be of use in this situation [5].

The present paper analyzes the viability and shortcomings of implementing glottal source modeling for the definition and identification of expressive speech. First we compare recognition rates with Nicolas Obin's style recognition results in Ircam database [10]. Secondly, to quantify the effects of speaker induced bias, we apply speaker normalization to the SES and SEV databases and apply a multi-dimensional scaling (MDS) algorithm [7] to the results in order to detect possible inconsistencies in the parametrization. Finally we suggest a style separation for the Spanish language based in C-oral-ROM database [2].

2. Corpora

For the present study we utilized 4 databases, 2 of them emotional speech databases (SES and SEV) with the other 2 being speaking style databases (Ircam and C-oral-ROM).

2.1. Emotional speech databases

Spanish Emotional Speech [8] and Spanish Expressive Voices [9] (SES and SEV) are two acted emotional speech databases featuring a total of 2 male speakers and 1 female speaker. In SEV, which consists of the female speaker and one of the males, the 6 basic emotions are represented: Happiness, sadness, cold anger, fear, disgust, surprise and hot anger, with the addition of neutral voices too. SES only considers a smaller subset of them, namely happiness, anger, sadness, surprise and the neutral voice.

2.2. Styled speech databases

C-oral-ROM [2] is a multi-language and multi-style database covering a wide spectrum of formal and informal speaking styles. The languages included are French, Italian, Portuguese and Spanish, with styles ranging from formal to informal, extracted either from the media, telephonic conversations or natural speaking. For the task considered in this paper we only analyzed the Spanish formal media styles: news broadcasts, sports, meteorological reports, reportages, talk shows, scientific press

and interviews. This data, having been extracted from media broadcasts of different stations presents a substantial variability in the recording environments and a high number of speakers (124). This implies that some of the speakers utter only a few short sentences, making them irrelevant in parametrical analysis.

Ircam [3] is a French multi-speaker multi-style database featuring 4 formal styles: news, live sports, political speeches and mass sermons. For each style it is guaranteed that multiple speakers and multiple environments are recorded, adding completeness to the database.

3. GlottHMM

GlottHMM [4] is a vocoding technique that was recently developed for parametric speech synthesis. It is based on decomposing speech into the glottal source and vocal tract through glottal inverse filtering. The vocal tract is parameterized by using a Line Spectral Frequency (LSF) vector (with 30 LSFs). The spectral tilt of the glottal source is also modeled using LSFs (with 10 LSFs). In addition, GlottHMM also extracts the F0 and harmonics to noise ratio (HNR) of the glottal source. The information of the F0 is used to separate between voiced and unvoiced frames. HNR is evaluated based on the ratio between the upper and lower smoothed spectral envelopes (defined by the harmonic peaks and interharmonic valleys) and averaged across five frequency bands according to the equivalent rectangular bandwidth (ERB) scale. Finally, in addition to the standard features utilized in GlottHMM, we added two additional features: Normalized Amplitude Quotient (NAQ) [11] and the magnitude differences between the 10 first harmonics of the voice source.

4. Experiments

4.1. Prosodical versus glottal modeling

This experiment aims to analyze the usefulness of glottal feature extracted from the source to the identification and characterization of expressive speech. It is known [5] that speech production features extracted from the source carry relevant information toward the expressivity of speech that is not fully present in the prosodic features. For this analysis we parted from a previous study carried on by Nicolas Obin on speaking style identification [10] in the Ircam database. This previous study relied on purely prosodic information: pitch and rhythm, obtaining identification rates of about 74%. On the contrary, the attempted study consisted on obtaining the GlottHMM features and attempting to solve the same classification problem that Obin attempted. The results showed a 95.4% recognition rate (Table 1).

To confirm the hypothesis that expressivity information is present in the glottal features we applied an InfoGain analysis on 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 2 show that there are a set of parameters even more informative than the F0, with speaking rhythm placing at a comparatively low position.

Table 1: *Recognition results of Ircam database using glottal features.*

Precision	Recall	F-Measure	Class
93.4	90.1	91.8	CHURCH
96.1	94.9	95.5	JOURNAL
95.6	98.7	97.1	POLITIC
96.3	98.9	97.6	SPORT
95.4	95.5	95.4	AVERAGE

Table 2: *Information gain of the best glottal features compared to prosodic features for Ircam database.*

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 f0 and rhythm is not a good classifier, but they show that considering glottal parameters for modeling expressivity is a pursuable idea. In fact, applying a greedy stepwise search of the features showed that in addition to the 6 top features shown in Table 2, an additional number of LSF together with rhythm and NAQ conformed the optimal recognition. An additional consideration that arose from this analysis is whether this set of features is too sensitive to noise or speaker bias to be useful in non-prepared databases or recording environments. To clarify this situation we realized the second set of experiments.

4.2. Speaker bias quantification

To quantify the effects of speaker bias we applied the GlottHMM feature analysis system to the pair of emotional speech databases SES and SEV. The change from a style-oriented database to an emotional speech analysis is mainly due to the requirements of a multi-speaker, controlled recording environment database for the considered analysis. Even so, as both emotion and style are aspects of expressive speech there is not a lack of generalization.

Preliminary emotion recognition tests showed a 81.2% recognition rate (see Table 3) with the rank seen in Table 4. 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

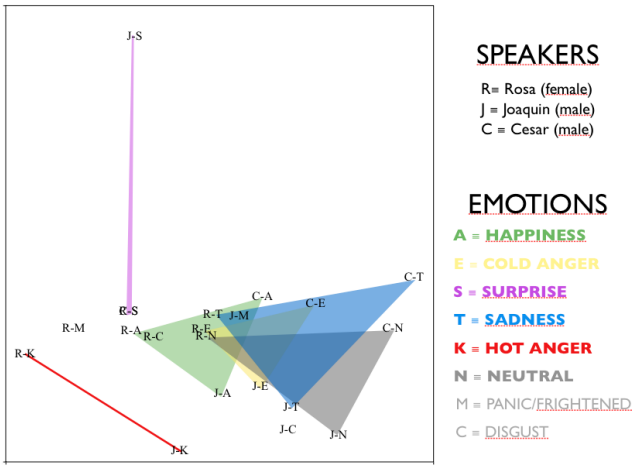

Figure 1: *MDS of the non-normalized emotions*

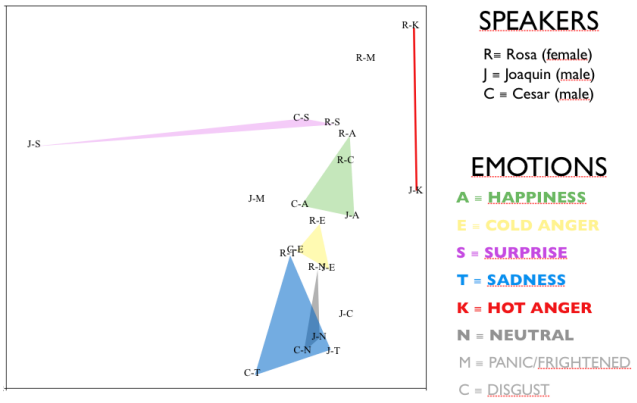

Figure 2: *MDS of the emotions normalized by the neutral voice of each speaker*

normalized every speaker's emotion with a z-score normalization algorithm respective to each of their speaker's neutral voice.

Following the normalization we obtained 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 2, MDS of the non-normalized distributions is also included in Figure 1.

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

Table 3: *Recognition results of the emotional databases 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

4.3. Spanish speaking styles space

The third experiment considered was the definition of a set of separable styles in Spanish media by analyzing C-oral-ROM. Because of the nature of the database, with many different speakers and most of them uttering only a handful of sentences on top of the greatly variable recording environments, we decided to only make use of proven robust features: F0 and speaking rate.

The first task was to focus on the distinctive examples of the main styles and choose representatives. The chosen styles and their distribution can be seen in figure 3. This figure allows a profound discussion on the meaning of the axes from a perceptual point of view: The F0 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 caster will have to speak over the noise of the crowd. The opposing situation can be seen in news broadcasts, where the caster speaks from a studio in which the recording environment is perfectly controlled.

The speaking rate axis reflects the spontaneity of the speech: the more the speaker has prepared the speech the faster he will be able to talk, as there is no need to pause and think the following phrase. The defining example

Table 4: *Information gain of the best glottal features compared to prosodic features for SES and SEV database.*

Ranked	Feature	Ranked	Feature
0.6712	LF0-mean	0.4065	HNR1-mean
0.5049	LSF3-mean	0.4035	LSF24-var
0.4515	LSFSOURCE10-var	0.3971	HNR4-var
0.4329	NAQ-mean	0.3968	HNR3-mean
0.4223	HNR3-var	0.3936	HNR2-mean
0.4088	HNR4-mean	0.3853	LSFSOURCE1-mean

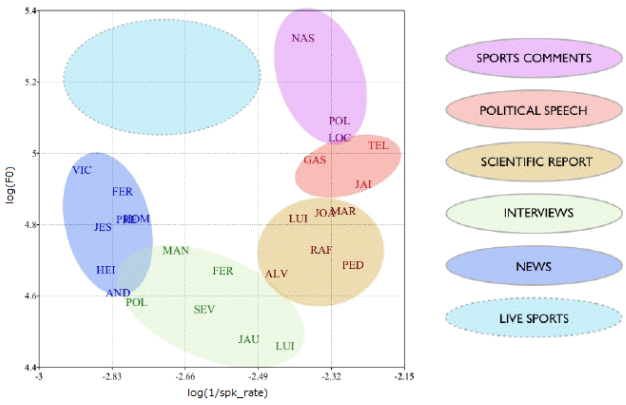

Figure 3: *Proposed speaking styles distribution*

is the news realm, where newscasters typically read the news instead of thinking on the phrasing so as to be able 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.

5. Discussion

The results of the first two experiments strongly suggest that glottal models can be successfully applied to expressive speech characterization. It was also seen that under controlled recording environments the parameters do not suffer from speaker bias. The problem arises when trying to analyze style databases, as current databases are not prepared for this task and tend not to show consistent style subsets or stable recording conditions.

A second problem is that even if different sources are using the same speaking style the nuances of the style are not necessarily consistent between them, introducing variability in the data. Ultimately this implies that establishing clear representatives of the different separable styles becomes a difficult problem.

Possible solutions to these problems could be either obtaining enough consistent data so that the effects of speaker variability become irrelevant or trying to record a database in which we control the environment and the speakers. The second solution would imply the use of acted data in a field in which there is a lot of raw data. The first solution, despite the vast quantity of raw data present, would lack adequate labeling.

6. Conclusions

In this paper we have shown how the use of glottal model features greatly increase recognition rates of expressive speech when comparing with purely prosodic analysis, obtaining rates of 95% for styled speech and 82% for emotional speech. Their usefulness was backed up by further analysis that showed that they do not suffer from speaker induced bias, as our multi-speaker analysis showed clear distinctions when applying a MDS analysis. Finally we proposed a style separation for Spanish formal speaking styles that is backed up on considerations of spontaienty and environmental circumstances correlating with prosodical features (F0 and speaking rate).

7. 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.

8. References

[1] Rob Clark and Simon King, "Simple4All", 2011; http://simple4all.org

[2] Emanuela Cresti, Fernanda Bacelar do Nascimento, Antonio Moreno Sandoval, Jean Veronis, Philippe Martin, Kalid Choukri, "The C-ORAL-ROM CORPUS A Multilingual Resource of Spontaneous Speech for Romance Languages", In Proc. of LREC, 2004.

[3] Nicolas Obin, Pierre Lanchantin, Anne Lacheret, Xavier Rodet,"Discrete/Continuous Modelling of Speaking Style in HMM-based Speech Synthesis: Design and Evaluation", in Proc. Interspeech 2011, pp. 2785-2788.

[4] Raitio, T. and Suni, A. and Yamagishi, J. and Pulakka, H. and Nurminen, J. and Vainio, M. and Alku, P., "HMM-based speech synthesis utilizing glottal inverse filtering", Audio, Speech, and Language Processing, IEEE Transactions on, 9:153-165, IEEE, 2011.

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

[6] Childers, D.G. and Lee, CK, "Vocal quality factors: Analysis, synthesis, and perception", J. Acoust. Soc. Am 90, pp. 2394-2410, 1991.

[7] Schiffman, S.S. and Reynolds, M.L. and Young, F.W. and Carroll, J.D., "Introduction to multidimensional scaling: Theory, methods, and applications", by Academic Press New York, 1981.

[8] J.M. Montero and J. Gutierrez-Arriola and S. Palazuelos and E. Enriquez and J.M. Pardo, "Spanish emotional speech from database to TTS", in Proc. of ICSLP 1998, pp. 923-925.

[9] R. Barra-Chicote and J. M. Montero and J. Macias-Guarasa and S. Lufti and J. M. Lucas and F. Fernandez and L. F. D'haro and R. San-Segundo and J. Ferreiros and R. Cordoba and J. M. Pardo, "Spanish Expressive Voices: Corpus for Emotion Research in Spanish", in Proc. of LREC 2008.

[10] Nicolas Obin, "MeLos: Analysis and Modelling of Speech Prosody and Speaking Style", PhD, Ircam-UPMC, Paris, 2011.

[11] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parameterization of the glottal flow", J. Acoust. Soc. Amer., vol. 112, no. 2, pp. 701-710, 2002.