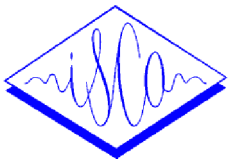

Speech Prosody 2010
Chicago, IL, USA
May 10-14, 2010

Expressive Speech Style Transformation: Voice Quality and Prosody Modification Using a Harmonic plus Noise Model

Carlos Monzo, Àngel Calzada, Ignasi Iriondo, Joan Claudi Socoró

GTM – Grup de Recerca en Tecnologies Mèdia
LA SALLE – UNIVERSITAT RAMON LLULL,
Quatre Camins 2, 08022 Barcelona (Spain)
{cmonzo, acalzada, iriondo, jclaudi}@salle.url.edu

Abstract

This paper proposes an approach to transform speech from a neutral style into other expressive styles using both prosody and voice quality (VoQ). The main aim is to validate the usefulness of VoQ in the enhancement of expressive synthetic speech. A Harmonic plus Noise Model (HNM) is used to modify speech following a set of rules extracted from an expressive speech corpus with five categories (neutral, happy, sensual, aggressive and sad). Finally, modified speech utterances were used to perform a perceptual test. These results indicate that listeners prefer prosody together with VoQ transformation instead of only prosody modification.

Index Terms: *Expressive speech transformation, voice quality, prosody, Harmonic plus Noise Model*

1. Introduction

The research fields of automatic speech recognition and text-to-speech (TTS) synthesis use expressive speech to make human-machine interaction more natural, e.g., in terms of emotion recognition [1] and voice transformation [2][3]. Voice quality (henceforth VoQ) and prosody parameters are used to represent the emotional content of speech [4]. In spite of the fact that VoQ has been less explored than prosody, recent works propose using both types of data to improve the acoustic modeling of expressive speech [4][5]. Other studies relate perceived speech features to VoQ parameters [6], dealing with the association of phonation type (e.g., whispery voice) and affective speaking [7][8].

In recent years, interest has increased in using the Harmonic plus Noise Model (HNM) for speech transformation [9], since a high quality and versatility is achieved. The parameterization of speech in both harmonic and stochastic components allows for a flexible manipulation of VoQ and the time and pitch-scale, maintaining a natural speech quality.

This work has two main aims: i) to show that when transformation from neutral to an expressive speech style (happy, sensual, aggressive and sad) is required, VoQ plus prosody modification improves the results over those obtained using only prosody modification; ii) to propose a methodology for measurement and modification of VoQ parameters using HNM parameterization of speech. A large expressive styles corpus is used to extract the relationships among different expressive styles from prosody and VoQ parameters designed especially for the analysis and transformation of expressive speech styles.

The paper is organized as follows. Section 2 discusses the previous work performed in this area, including the construction of the expressive speech corpus, VoQ and prosody parameterization, and HNM. Section 3 presents the methodology proposed for transformation of expressive speech style. Section 4 reports and discusses the conducted experiment, and finally, Section 5 concludes the paper.

This work has been developed under SALERO project (IST FP6-2004-027122) of the European Community.

2. Previous work

2.1. Expressive speech corpus

An expressive oral corpus for Spanish speech synthesis was developed in the previous works of our research group with two main purposes: first, to be used in the acoustic modeling (prosody and VoQ) of emotional speech, and second, as a speech unit database for a TTS synthesizer. For the recording, texts semantically related to different expressive styles were selected from a textual database of advertisements and read by a female professional speaker. The voices in audio-visual advertisements were studied for five categories of goods and the most suitable expressive style was assigned to each one: new technologies (neutral-mature, NEU), education (happy-elation, HAP), cosmetics (sensual-sweet, SEN), automobiles (aggressive-hard, AGG), and trips (sad-melancholic, SAD). The recorded corpus has 4638 phrases and is 5 h 12 min long. The expressive styles in this corpus were evaluated using a subjective test (see confusion matrix in Table 1) where the possible answers were the five styles of the corpus plus the additional option of Do not know/Another (Dk/A) to avoid introducing a bias to the results from confusion or doubt between two or more options. Detailed information about this corpus can be found in [5].

Table 1. *Confusion matrix for the subjective test* [5].

%	NEU	HAP	SEN	AGG	SAD	Dk/A
NEU	**86.4**	1.3	3.6	5.3	0.7	2.7
HAP	1.9	**81.0**	0.2	15.6	0.1	1.2
SEN	4.7	0.1	**86.8**	0.0	5.7	2.6
AGG	1.8	14.2	0.1	**82.7**	0.1	1.1
SAD	0.5	0.0	0.6	0.0	**98.8**	0.1

2.2. Voice quality parameters

VoQ parameters involved in this work, some previously proposed by other authors [2], were analyzed and used in previous experiments and their ability to discriminate expressive speech styles was demonstrated [5][8]. In addition, to fit the behavior of these parameters to the application of transforming expressive speech styles, some of the parameters were redesigned and then their usefulness was demonstrated [10]. Thus, the following subset of possible parameters was considered in this expressive speech styles transformation experiment:

- *Jitter and Shimmer*, these parameters describe the cycle-to-cycle variations of the fundamental period and waveform amplitude, respectively, describing frequency and amplitude modulation noise. These parameters were developed taking into account that they would be used in expressive speech applications [10].
- *Harmonic-to-Noise Ratio* (*HNR*) describes the ratio between harmonic and stochastic components.
- *Hammarberg Index* (*hammi*), defined as the difference between the maximum energy in the 0-2000 Hz and 2000-5000 Hz frequency bands.
- *Relative amount of energy* in the high (above 1000 Hz) versus the low frequency range of the voice spectrum (*pe1000*).
- *Drop-off of spectral energy* above 1000 Hz (*do1000*), a linear approximation of spectral tilt above 1000 Hz.

2.3. Prosody parameters

Prosody parameters are extracted from the expressive speech corpus and modeled using Case Based Reasoning (CBR), a data mining technique with utility for expressive speech synthesis applications [11]. When speech transformation is conducted, the CBR returns the best case that fits with the target. Given the input text, the predicted parameters are: *contour of fundamental frequency* (F_0), *contour of energy*, and *segmental duration*.

2.4. Harmonic plus Noise Model

HNM divides the speech signal into the harmonic and stochastic components [12]. The lower spectrum band is mainly modeled by a sum of harmonically related sinusoids (harmonic component) characterizing the voiced part of speech. This is completely characterized by the time-variant *amplitudes*, *frequencies*, and *phases* of these sinusoids.

Unvoiced sounds and all non-periodic events in speech are modeled by the stochastic component, a time-variant autoregressive (AR) process where both spectral and temporal fluctuations are represented by *Linear Predictive Coding (LPC) coefficients* and *energy*.

The implementation of our HNM analysis is based on the Depalle's frequency domain algorithm [13]. Although HNM implies that the lower band of the signal spectrum is harmonic, Depalle's algorithm does not guarantee the harmonicity of the estimated frequencies. To ensure harmonicity, we modified the frequency estimation algorithm using the Lagrange multiplier optimization procedure [14].

3. Proposed method

This section proposes a method for expressive speech transformation. The benefit of adding VoQ to the transformation process to improve the perception of the transformed speech was analyzed and compared with prosody-only transformation. Previous works have already demonstrated the usefulness of modeling prosody in transformation of expressive speech styles, identifying various problems during the modeling of those expressive styles [11]. In addition, VoQ and prosody have been proposed to be used together to allow them to complement each other [5][8]. HNM-based synthesis was selected for carrying out all transformations due to its flexibility. In previous experiments conducted using Pitch-Synchronous Overlap and Add (PSOLA) based TTS [10], not all VoQ parameters could be modified because the stochastic component was unknown. Moreover, transformation of spectral bands using PSOLA is difficult without introducing distortion in the resulting speech, since this spectrum information is not available during the speech creation.

As shown in Figure 1, our system is divided into three main parts. First, HNM analysis and resynthesis blocks extract the original speech information and regenerate it when the transformation is conducted. Second, prosody and VoQ modeling is performed. Prosody is predicted by means of CBR, obtaining the target information for each phoneme: *contour of* F_0, *contour of energy*, and *segmental duration*. For VoQ modeling, VoQ parameters involved in the transformations and the related target values are selected. Finally, the speech transformation is carried out based on the results of the HNM analysis and the prosody and VoQ modeling.

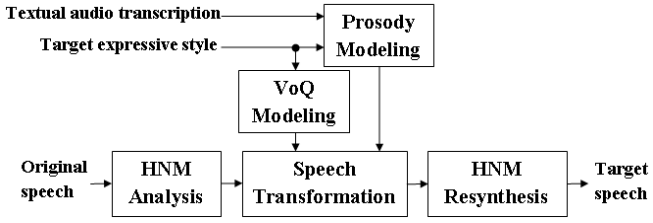

Figure 1: *Block diagram of the transformation process.*

Various considerations must be made to determine which parameters and values should be involved in the transformations. For prosody modifications, all parameters were involved in all transformations. The number and selection of VoQ parameters to be used during the transformation were chosen as follows: i) from a previous study about the use of VoQ parameters in the discrimination of expressive speech styles [8]; ii) as a continuation of this work, an exhaustive classification was conducted to obtain different configurations for all parameters and expressivities; iii) from descriptive statistics calculated for all expressive styles of the corpus and all involved VoQ parameters, comparing the results for neutral expressive style with the obtained for the rest of them; and iv) from heuristics tests performed during the design process to adjust the final parameters involved in the different transformations.

Regarding the target values for prosody and VoQ, the target values for prosody were obtained using a new contour of F_0, an *energy contour*, and *segmental durations* predicted from CBR. For VoQ modification, a factor per parameter was predicted from the extracted corpus values for the different expressive styles. Based

Table 2. *Set of selected VoQ parameters for the neutral-to-target expressive speech style transformation and the % factor to be applied for every involved parameter ("*": factor has been fined tuned; "--": parameters not involved in the transformation).*

% factor	jitter	shimmer	HNR	hammi	pe1000	do1000
Happy	--	--	--	-60	110*	--
Sensual	--	85	-50*	155	-50*	--
Aggressive	-20*	-45	--	-70*	220*	--
Sad	-45*	90	--	655	-75	--

on the heuristics tests carried out during the design process for VoQ, fine tuning was carried out to improve the expressiveness perception while minimizing the distortion due to these modifications. Table 2 presents the VoQ parameters used for eachtransformation and the percent variation of the original factor from neutral, calculated from the mean values of the corpus. Note that the "*" symbol indicates that this factor was fined-tuned to maximize perception and to avoid decreasing the final speech quality, whereas the symbol "--" shows that those parameters are not involved in the transformation.These VoQ factors were applied directly as a linear transformation of the HNM parameters. During the transformation, each factor was considered as a value that multiplies the original parameter.

This speech transformation methodology is applied to HNM parameters, i.e., *frequencies, amplitudes,* and *phases* for the harmonic component; and *energies* for the stochastic component. Two kinds of modifications are carried out. The first transformation is related to the prosody, which modifies all vectors from the CBR prediction. All HNM parameters are involved in this transformation due to *frequencies*, *amplitudes*, and *phases* are modified to produce the required *contour of F_0, contour of energy*, and *segmental duration*. Also, the VoQ parameters have different degrees of necessity. *The jitter* parameter is needed to modify the *F_0 contour*; in this way, the vector *frequencies* and *amplitudes* and *phases* are also affected, since *amplitudes* and *phases* are highly related to the *frequency*. In the *shimmer* case, all HNM parameters are involved because modifications are carried out directly on the time domain, thus HNM parameters have to be reestimated. The *hammi* and *pe1000* parameters are related to the energy in different frequency bands of the harmonic component, therefore the *amplitudes* vector must be modified and the *phases* must be reestimated. To vary the *HNR* parameter, both harmonic and stochastic components are modified by modifying the vector *amplitudes* (so that *phases* are also reestimated) and *energy*, respectively. For *hammi*, *pe1000*, and *HNR*, the variation factor is distributed between the spectral bands, ensuring that the final speech energy is the same as the original when the transformation is performed. Finally, *do1000* modifies the *amplitudes* vector (so that *phases* are also reestimated), although no different spectral bands are involved.

Notice that *do1000* does not appear in the selected parameters. This is due to the fact that modifications to this parameter did not yield expected results, with artifacts appearing in the final speech producing an unpleasant and unnatural voice.

4. Experiment

To evaluate the effect of introducing VoQ into the prosody-based expressive speech style transformation, a perceptual test was carried out using a web platform designed for this type of experiment [15].

In this evaluation, 8 neutral utterances were collected from the corpus. For each one, HNM analysis was conducted on the audio, and both prosody and VoQ were modeled. For all utterances, VoQ parameters and prosody were modified. The transformation for each target expressive speech style resulted in 32 new utterances with prosody modification and 32 new utterances with both prosody and VoQ modification.

For each pair of new utterances, the question, "*From which utterance is the target emotion best perceived*?" was asked, and the listeners answered according to their preferences. There were seven possible answers based on the Comparative Mean Opinion Scale (CMOS) test: "*much more,*" "*more,*" or "*slightly more*" than the other one, or "*the same,*" with scores of 3, 2, 1, 0, -1, -2 and -3. Positive values were assigned to cases where VoQ increased the ability to perceive the expressive style. A total of 15 volunteer listeners performed the test (5 females and 10 males). Of these volunteers, 40% were experts in speech technologies, whereas 60% were non-experts. Comments could be provided when the test was completed, allowing for the collection of information about each listener's criteria and opinions about the difficulty and quality of the utterances, which will be useful for future work.

The obtained results presented in the boxplots in Figure 2 demonstrate that adding VoQ to prosody during the modification is preferred. While the addition of VoQ improves the perception of all expressive speech styles, difficulties exist in some cases. The main reasons for perceptual difficulties are: i) each listener has a different criterion for deciding on the intensity of their perception; ii) the semantic content of the utterance unconsciously affects its perception; and iii) parameter modifications can produce voice degradation, while listeners tend to prefer a more natural voice.
Moreover, analysis of the median and confidence intervals was performed using the Wilcoxon test (see Table 3) [16]. These results indicate that using VoQ along with prosody is preferred over using prosody alone. This is interesting because this finding holds true both in cases where prosody typically obtained good results (sad) and in those where prosody alone resulted in problems in identifying the expressive style (aggressive). In addition, for the sensual case, the best results were obtained with a median equal to 2, corresponding to an intermediate preference ("*more*") for the addition of VoQ, whereas the remaining expressive styles have a median value between 0.5 (aggressive) and 1 (happy and sad), indicating a "*slight*" preference for the VoQ transformation. Finally, for all expressive styles, the confidence interval shows a maximum dispersion with a confidence level of 95%, which indicates that the best results are obtained using VoQ, and shows the preference for using prosody together with VoQ.

To address the voice effects of VoQ use, the sensual style is improved by using a whispering voice effect. For happy and aggressive transformations, the best results are obtained when the voice tension effect is present. Finally, for the sad modification case, the best results are obtained with trembling voices.

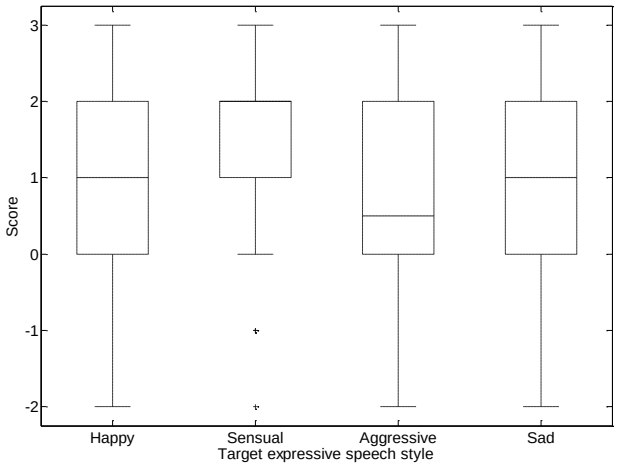

Figure 2: *CMOS evaluation.*

Table 3. *Median and confidence interval with a confidence level of 95%.*

		HAP	SEN	AGG	SAD
Median		1	2	0.5	1
Confidence interval	Min	1	1.5	0.5	0.5
	Max	1.5	2	1	1.5

5. Conclusions

This work has presented and analyzed a method for expressive speech style transformation using prosody and voice quality (VoQ) modification based on HNM. First, a flexible HNM parameterization is used to extract the fundamental speech parameters that must be used during the performed prosody and voice quality (VoQ) modifications. Second, prosody is predicted by means of CBR and modified using the HNM parameters, representing a first attempt at the expressive speech style transformation. Finally, VoQ is taken into account, and once prosody modifications are applied, the VoQ parameters are transformed by varying the HNM parameters. To select which VoQ parameters should be modified, an analysis of the best configurations was performed.

All parameters involved in this method have been presented, explained, and discussed, and the modified methodology using HNM parameterization has been shown. Whereas prosody modeling and modification is common for all expressive styles examined here, VoQ depends on the target expressive speech style.

The use of VoQ improves perception of expressive speech styles. A perceptual test demonstrated that all expressivities were perceived better when VoQ parameters were combined with prosody. Despite these positive results, more work remains, for example, the analysis of the best place to apply the VoQ modification. The use of CBR could be an interesting approach to value prediction. Moreover, more work can be done in parameterization to improve the characterization of each expressive speech style and the quality of the transformation.

Finally, this work shows the usefulness of using prosody together with VoQ in expressive speech style transformations. In addition, this work proposes a method to be used in speech transformation, which allows for easy modification of all the parameters.

7. References

[1] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. and Taylor, J. “Emotion recognition In human-computer interaction”, IEEE Signal Processing Magazine, 18(1):32–80, 2001.

[2] Drioli, C., Tisato, G., Cosi, P. and Tesser, F. “Emotions and voice quality: experiments with sinusoidal modeling”, In Proceedings of Voqual 2003, Voice Quality: Functions, Analysis and Synthesis, ISCA Workshop, Geneva, 127–132, 2003.

[3] Turk, O., Schröder, M., Bozkurt, B. and Arslan, L.M. “Voice quality interpolation for emotional text-to-speech synthesis”, In Interspeech, Lisbon, 797–800, 2005.

[4] Cabral, J. and Oliveira, L. “Pitch-synchronous time-scaling for high-frequency excitation regeneration”, In Interspeech, Lisbon, 1513–1516, 2005.

[5] Iriondo, I., Planet, S., Socoró, J.C., Martínez, E., Alías, F. and Monzo, C. “Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification”, Speech Communication, 51:744–758, 2009.

[6] Bänziger, T. and Scherer, K. “A study of perceived vocal features in emotional speech”, In Proceedings of Voqual 2003, Voice Quality: Functions, Analysis and Synthesis, ISCA Workshop, Geneva, 169–17, 2003.

[7] Gobl, C. and Ní Chasaide, A. “The role of voice quality in communicating emotion, mood and attitude”, Speech Communication, 40:189–212, 2003.

[8] Monzo, C., Alías, F., Iriondo, I., Gonzalvo, X. and Planet, S. “Discriminating Expressive Speech Styles by Voice Quality Parameterization”, ICPhS, Saarbrücken, 2081–2084, 2007.

[9] Laroche, J., Stylianou, Y. and Moulines, E. “HNS: Speech modification based on a harmonic+noise model,” Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on, 2(27-30):550–553, 1993.

[10] Monzo, C., Iriondo, I. and Martínez, E. “Jitter and Shimmer Measurement and Modification Procedure Applied to Expressive Speech Synthesis” V Jornadas en Tecnología del Habla (JTH2008), Bilbao, 58–61, 2008. (*In Spanish*).

[11] Iriondo, I., Socoró, J.C., Alías, F. “Prosody modelling of Spanish for expressive speech synthesis”, In Proceedings of 32nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Honolulu, 4:821–824, 2007.

[12] Laroche, J., Stylianou, Y. and Moulines, E. “Hnm: a simple, efficient harmonic+noise model for speech”, Applications of Signal Processing to Audio and Acoustics, 1993, Final Program and Paper Summaries, 1993. IEEE Workshop on, New Paltz, 169–172, 1993.

[13] Depalle, P. and Helie, T. “Extraction of spectral peak parameters using a short-time fourier transform modeling and no sidelobe windows”, Applications of Signal Processing to Audio and Acoustics, 1997, IEEE ASSP Workshop on, 4, 1997.

[14] Moon, T.K. and Stirling, W.C. “Mathematical methods and algorithms for Signal Processing”, Pap cdr Prentice Hall, 1999.

[15] Planet, S., Iriondo, I., Martínez, E. and Montero, J.A. “TRUE: an online testing platform for multimedia evaluation” In Proc. 2nd International Workshop on EMOTION: Corpora for Research on Emotion and Affect at LREC’08, Marrakech, 2008.

[16] Hollander, M. and Wolfe, D.A. “Nonparametric Statistical Methods” New York: John Wiley & Sons, 1999.