On the limitations of voice conversion techniques in emotion identification tasks

R. Barra, J.M. Montero, J. Macias-Guarasa, J. Gutiérrez-Arriola, J. Ferreiros, J.M. Pardo

Speech Technology Group
Universidad Politécnica de Madrid, Spain
{barra,juancho,macias}@die.upm.es, jmga@ics.upm.es, {jfl,pardo}@die.upm.es

Abstract

The growing interest in emotional speech synthesis urges effective emotion conversion techniques to be explored. This paper estimates the relevance of three speech components (spectral envelope, residual excitation and prosody) for synthesizing identifiable emotional speech, in order to be able to customize voice conversion techniques to the specific characteristics of each emotion. The analysis has been based on a listening test with a set of synthetic mixed-emotion utterances that draw their speech components from emotional and neutral recordings. Results prove the importance of transforming residual excitation for the identification of emotions that are not fully conveyed through prosodic means (such as cold anger or sadness in our Spanish corpus).

Index Terms: speech synthesis, voice conversion, emotional speech, perceptual test

1. Introduction

One of the current research lines on Speech Technologies is expressive synthesis. As in spoken dialog systems request for friendlier human-machine interfaces, the output of Text To Speech systems need to be richer and richer. One of the most requested improvements is to include emotional speech capabilities.

One strategy for the generation of emotional speech is based on unit selection techniques, which can provide high quality synthetic speech. Nevertheless, these techniques can only copy emotional speech from a corpus, without any generalization on how to synthesize each emotion.

An alternative technique is to build a model of the emotions to be synthesized and implement spectral voice conversion algorithms [1][2] to carry out emotion conversion [3][4]. This technique mainly tries to transform the segmental information, modeling the spectral envelope by LPC [5], LSF [2] or MFCC [6]. Recently, residual excitation conversion has been studied [7][8]. The role of prosody in emotional speech has also been analyzed [9] and used in expressive synthesis [10][11].

In this work we analyze the relevance of each speech component (spectral envelope, residual excitation and prosody) for each specific emotion in a Spanish emotional speech corpus. The analysis will be based on the perceptual evaluation of a set of synthetic mixed-emotional utterances that draw their speech components from emotional and neutral recordings. Results provide clues to adapt the algorithms of the emotional conversion synthesis module to the needs of each emotion.

2. Spanish Emotional Speech Corpus

In this work, the Spanish Emotional Speech corpus (SES) has been used. It contains two emotional speech recording sessions played by a professional male actor in an acoustically-treated studio. Each recorded session includes thirty words, fifteen short sentences and three paragraphs, simulating three basic or primary emotions (*sadness*, *happiness* and *cold anger*), one secondary emotion (*surprise*) and a *neutral* speaking style. The text uttered by the actor did not convey any explicit emotional content.

This parallel corpus was phonetically labeled in a semiautomatic way. An automatic pitch epoch extraction software was used, but the outcome was manually revised using a graphical audio-editor program, which was also used for phoneme location and labeling.

The assessment of the emotional voice was aimed at evaluating the speech corpus as a model for recognizable emotional speech [12]. Perceptual copy-synthesis tests (mixing emotional phoneme durations and linearized F0 contours with neutral diphones or vice versa), showed the segmental or non segmental nature of each emotion [10].

Emotional patterns were also evaluated by means of automatic identification experiments [13]. Emotional information was analyzed using segmental (MFCC) and prosodic information (F0-related statistics). When both sources of information were combined, better classification rates were achieved, even for prosodic emotions.

3. Speech synthesis model

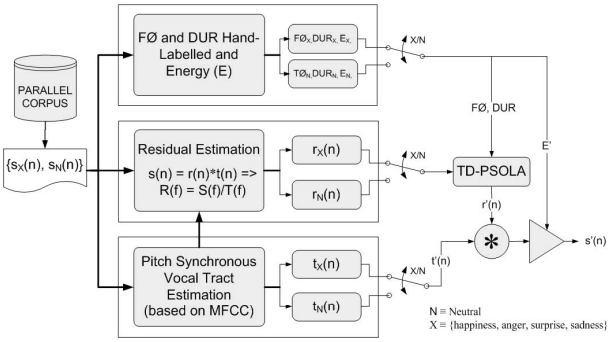
The pitch-synchronous residuum-filter synthesis model used in this work is based on the classical source-filter model [14]. The filter models the spectral envelope of the vocal tract in a two pitch-period window. The source is the excitation signal that can regenerate the speech by means of overlap-add techniques. This residual excitation models not only the signal as generated by the vocal cords, but also the behavior of the vocal tract that has not been modeled by the filter.

3.1. Vocal tract estimation

The MFCC-based spectral conversion technique is one of the most usual strategies in Voice Conversion. Genrally speaking, first MFCC coefficients model the slow variation of the spectral envelope, which represents an estimation of vocal track spectral information. Once we have computed the target MFCC, we could reconstruct the target spectral envelope.

We calculated 21 MFCC, that could potentially be transformed by standard voice conversion techniques [1]. A high-

Figure 1: *Synthesis Process*



resolution filtered spectrum is calculated by an over-sampled DCT over the zero-padded MFCC vector [14]. Then, the logarithm operation in MFCC extraction is inverted. Finally, a N-points FFT of the spectral envelop is obtained by means of interpolation.

3.2. Residuum estimation

Once the vocal tract spectrum is estimated, we calculate the residual spectrum as:

$$\begin{aligned} s(n) = t(n) * r(n) \Rightarrow S(f) = T(f) \cdot R(f) \\ \Rightarrow R(f) = S(f)/T(f) \end{aligned} \quad (1)$$

where $s(n) =$ Hanning-windowed signal frame
$t(n) =$ vocal tract, $r(n) =$ residual excitation

3.3. Prosody Modeling

Phoneme durations, F0 and energy model emotional prosody. Durations and F0 are modified by applying Residual-Domain PSOLA to $r(n)$; when a pitch reduction or increment is needed, the algorithm adds zeros or removes side samples in each pitch-synchronous frame. Source energy value is computed from the whole frame and it is modified once the target frame has been generated.

Figure 1 shows the synthesis scheme adopted in this work. Every pair of utterances ($\{S_N, S_X\}$ comprises one neutral and one emotional version of the same specific sentence). They are splitted into the components of the speech synthesis model, in order to select what emotional information (named as X in the figure) is combined with neutral components to create the synthesized utterance.

4. Perceptual Experiments

In previous experiments on the same SES database [3], the relevance of stylized F0 and tempo was established, but other segmental and prosodic features (such as spectral envelope, energy or F0 micro-prosody) were not accounted for. The mixed-emotion synthesized utterances of this test allow analyzing the contribution of three components (vocal tract envelope, residual excitation and prosody), to the identification of the intended emotion by human listeners.

4.1. Description

After selecting three short sentences from the SES corpus, the emotional and neutral recordings of these sentences were analyzed to extract the speech components of each utterance. For generating each mixed-emotion utterance in the test, we have combined some components from one emotional SES recording and other components from a phonetically-aligned neutral utterance. Four possible combination schemes are available as shown in Figure 1 (X represents {*happiness (H)*, *anger (A)*, *surprise (Su)*, *sadness (Sa)*} and N represents *Neutral*):

- $T_X - R_X - P_X$**:** the three components (vocal **T**ract envelope, **R**esidual excitation and **P**rosody) were taken from an emotional utterance to be re-synthesized. This minimum-distortion scheme provides a high-performance reference for each emotion.
- $T_N - R_N - P_X$**:** The prosody of an emotional utterance is applied to neutral segmental components (both vocal tract envelope and residual excitation). This schemes intends evaluating the relevance of each emotional prosodic pattern.
- $T_X - R_X - P_N$**:** the emotional segmental components were combined with neutral prosody to estimate the contribution of segmental information to emotion identification rates.
- $T_X - R_N - P_N$**:** Only vocal tract envelope is taken from emotional recordings and the remaining components (residual excitation and prosody) were extracted from a neutral utterance. As MFCC transformation is a standard voice conversion technique, we are estimating whether MFCC conversion can convey most segmental information which characterizes each emotion, especially on those emotions with poorly-recognized prosodic patterns

Finally, thirty-four listeners have listened to the 48 utterances in the test set (3 sentences · 4 emotions · 4 schemes):

- Select the emotion intended on each utterance from a limited set: {*H, A, Su, Sa, N, other*}.
- Evaluate voice quality in a MOS scale (from poor quality -1- to high quality -5-).

Before making a decision, a utterance could be played as many times as listeners needed, but they could never re-play previous utterances. Every listener used headphones for the test and they never heard any SES emotional recording before this test.

4.2. Re-synthesis Evaluation

Table 1 shows the confusion matrix for these re-synthesis utterances. High-quality re-synthesized speech (MOS from 4.42 to 4.57) was so similar to natural recordings that listeners were able to identify every emotion (average identification rate is 82.6% and average precision is even higher: 89.5%) in spite of not being familiar to the database . The high identification rates obtained validate the emotional patterns used by the actor to simulate each emotion.

There is a clear distinction between the two positive emotions. The highest confusion rate was caused by identifying *happiness* as *surprise* (10.8%) or vice versa (20.6%), because both high-pitched positive emotions are the closest ones in our emotional space. The negative emotions (*anger* and *sadness*), achieved the highest scores both in identification rate (87.5%) and precision (97%): their patterns were easy to recognize and exhibited no confusion with other emotional patterns (indeed, most of their confusion rate was due to the *other* option, 8.8%, not to confusion with the positive emotions).

Table 1: *Identification rates for re-synthesis utterances.*

SYNTHESIS SCHEME	INDENTIFIED EMOTION					
T_X-R_X-P_X	Happ.	Anger	Surprise	Sadness	Neutral	Other
T_H-R_H-P_H	81.4%	1.0%	10.8%	2.0%	2.9%	2.0%
T_A-R_A-P_A	1.0%	87.3%	2.9%			8.8%
T_{Su}-R_{Su}-P_{Su}	20.6%	2.9%	74.5%			2.0%
T_{Sa}-R_{Sa}-P_{Sa}				87.3%	3.9%	8.8%
PRECISION	79%	97%	84%	98%		

Table 2: *Identification rates for emotional prosody utterances.*

SYNTHESIS SCHEME	INDENTIFIED EMOTION					
T_X-R_X-P_N	Happ.	Anger	Surprise	Sadness	Neutral	Other
T_N-R_N-P_H	18.6%		31.4%	24.5%	13.7%	11.8%
T_N-R_N-P_A		25.5%	4.9%	36.3%	22.5%	10.8%
T_N-R_N-P_{Su}	15.7%	4.9%	73.5%	1.0%		4.9%
T_N-R_N-P_{Sa}	2.0%			89.2%	5.9%	2.9%
PRECISION	51%	84%	67%	59%		

4.3. Emotional prosody

The results for those utterances with emotional prosody applied to neutral segments (Table 2) show the clear prosodic identifiability of *sadness* (89.2%) and *surprise* (73.5%). This is confirmed by the Pearson correlation coefficient between these two rows and the equivalent rows in the re-synthesis experiment, which is higher than 0.9950. As these emotions are strongly prosodic, they are robustly identified even when segmental synthesis distortion noise is high and the MOS quality is as low as 1.99 (for *surprise*). This shows that a clear prosodic pattern is hardly affected by segmental noise.

Nevertheless, *sadness* achieved relatively-low precision in this prosodic part of the test (59%), because the prosodic modification of the emotions that lack an easy-to-identify prosodic pattern (*happiness* and *anger*), was linked by listeners have to *sadness*. This confusion was not exhibited when testing the re-synthesized samples. This is especially paradoxical for *happiness* (a 24.5% confusion with *sadness*). However, as it will be shown, *happiness* is the emotion most sensitive to distortion and noise.

In the experiments of previous papers [12], *cold anger* was almost never identified by means of stylized pitch and tempo. Emotional micro-prosody and energy contours have now increased the identification rate of *cold anger* prosody, although confusion rates with *neutral* and *sadness* are still very high (22.5% and 36.3%). However, high identification precision for *cold anger* confirms that some of its prosodic patterns have been identified by listeners.

When compared to the re-synthesis table, the confusion between positive emotions (*happiness* and *surprise*) has now been reversed: happy prosody is identified as surprised (31.4%) more times than the opposite way, because prosodic patterns of *happiness* are not as distinctive as *surprise* patterns. Although precision for these positive emotions is lower, when *happiness* and *surprise* are grouped, their global precision is almost the same as in re-synthesis.

Generally speaking, mismatches between residual excitation and prosody in these utterances result in a poor overall quality (2.4 in a MOS scale). One could think that the higher prosodic modification, the higher degradation (for *surprise* MOS decreases from 4.57 to 1.99, and for *happiness* from 4.54 to 2.72). Nevertheless, the highest degradation is associated to *cold anger* prosody (from 4.42 to 1.6), because of the application of the micro-prosodic jitter of *angry* utterances, without the complementary *angry* segmental component. Although jitter does not characterize our *sad* recordings, it can be interpret as a representative feature of highly negative emotions (such as *fear* or extreme *sadness*), which lead to the observed confusion.

4.4. Segmental relevance

Table 3 shows the results for the utterances with a combination of neutral prosody and emotional vocal tract and residual excitation. The clear segmental nature of *cold anger* [10] is easily identified by listeners: the score (99.0%) is higher than the corresponding re-synthesis score (87.3%), and quality (3.3) is higher than the average quality (2.8). This significant recognition improvement was caused by the prosody-excitation mismatches which favor the confusion with a negative emotion (*anger*).

Although overall quality (2.8) is higher than in the preceding emotional-prosody section, degradation seriously affects *happiness* identification (10.8%), leading listeners to identify it as *neutral* (25.5%), or even as *anger* (27.5%). This confirms the sensitivity of happiness to noise and distortion.

The segmental component of *surprise* was mostly identified as *neutral* and its precision is lower than 30%, reassuring the relevance of their prosodic components. Only a marginal segmental similarity between *surprise* and *happiness* was observed (12.7%), because most of their similarity is prosodic, not segmental.

Although *sadness* was highly confused with *neutral* (46.1%), the precision is rather high (65%), suggesting that the segmental component of *sadness* is not irrelevant and can complement the prosodic patterns.

Table 3: *Identification rates when emotional segmental information is used.*

SYNTHESIS SCHEME	INDENTIFIED EMOTION					
T_X-R_X-P_N	Happ.	Anger	Surprise	Sadness	Neutral	Other
T_H-R_H-P_N	10.8%	27.5%	7.8%	12.7%	26.5%	15.7%
T_A-R_A-P_N	1.0%	99.0%				1.0%
T_{Su}-R_{Su}-P_N	12.7%	12.7%	5.9%	2.9%	43.1%	22.5%
T_{Sa}-R_{Sa}-P_N	1.0%	5.9%	6.9%	29.4%	46.1%	10.8%
PRECISION	44%	84%	29%	65%		

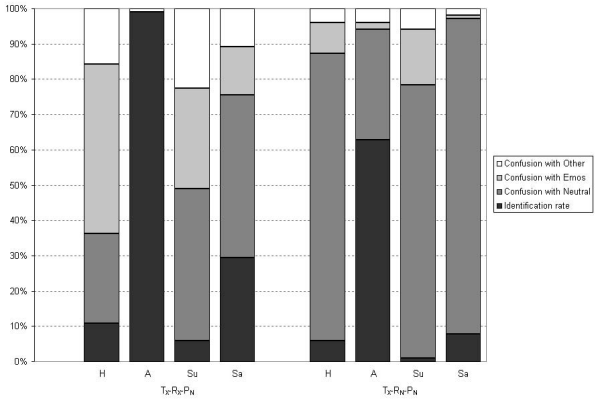
4.4.1. MFCC Relevance for Emotion Conversion

In order to analyze the relevance of modeling vocal tract envelope in an emotion conversion system, we included some utterances in the test with an emotional MFCC-estimated envelope, copying the prosody and residual information from a neutral recording.

Figure 2 plots emotion identification rates and the confusion with *neutral*, with *other* and with all the other emotions,

when using all the segmental emotional components ($T_X - R_X$) and when using only the emotional residual ($T_X - R_N - P_N$).

Figure 2: *Identification and confusion rates when using emotional segments or emotional spectrum envelope.*



For $T_X - R_X - P_N$, confusion with *neutral* is 25.5%, 43.1%, 46.1% for *happiness*, *surprise*, and *sadness* respectively. *Anger*, the best identified, is never confused with neutral. Quality is less than 3 (2.72) in *happiness* and induces listeners to identify it as any other emotion (48.0%), as results obtained when applying emotional prosody to neutral segments.

For $T_X - R_X - P_N$, identification rates and confusion with *other* and remaining emotions decrease in favor of the confusion with *neutral* (81.4% for *happiness*, 31.4% for *anger*, 77.5% for *surprise* and 89.2% for *sadness*).

5. Conclusions

We have analyzed emotional and neutral speech, to estimate the relevance of three components (vocal tract, residual excitation and prosody) to identify four emotions (*happiness, anger, surprise and sadness*). Mixed-emotion utterances in the evaluation were synthesized by combining speech components extracted from emotional and neutral recordings in our Spanish corpus.

We can classify our corpus emotions as: non-segmental (*surprise*), mainly prosodic (*sadness*), mainly segmental (*cold anger*) or complex (*happiness*). In contrast to previous studies, no purely-segmental emotion has been identified (*cold anger* energy curve and micro-prosody proved to be precisely identifiable by listeners).

We have verified that standard MFCC-based vocal tract modeling does not fully convey emotion-identification information. Therefore, emotion conversion systems should also transform residual excitation as it increases significantly emotion identification rates on segmental and mixed-nature emotions such as *anger* and *sadness* (up to 40%), while decreasing confusion to *neutral* equivalently.

Prosodic patterns of mainly prosodic emotions (such as *surprise* or *sadness*) are clearly identifiable and quite robust to synthesis artifacts. However, segmental information should be consistently included to keep quality high.

On the contrary, mismatches between the prosodic and segmental components severely affect identification of segmental emotions. Minimal synthesis distortion favors the identification of negative emotions such as *sadness* and, especially, *cold anger* (better identified than in re-synthesis utterances).

As *happiness* is complex and non-linear, all emotional components must be included to clearly transmit this emotion.

6. Acknowledgement

This work has been partially funded by the Spanish Ministry of Education and Science under contracts DPI2004-07908-C02-02 (ROBINT) and TIN2005-08660- C04-04 (EDECAN-UPM) and by UPM-CAM under contract CCG06-UPM/CAM-516 (ATINA). Special thanks to the students of the *Laboratorio de Sistemas Electrónicos Digitales*, who participated in the tests.

7. References

[1] Stylianou, Y., Cappe, O. and Moulines, E., "Continuous probabilistic transform for voice conversion", IEEE Tr. on Speech and Audio Process., Vol. 6(2), pp. 131-42, 1998.

[2] Kain, A. and Macon, M.W., "Spectral voice conversion for text-to-speech synthesis", Proc. of ICASSP, vol. 1, pp. 285-288, 1998.

[3] Kawanami,H., Iwani, Y. ,Toda, T., Saruwatari, H. and Shikano, K., "GMM-based Voice Conversion Applied to Emotional Speech Synthesis", Proc. of Eurospeech, pp. 2401-2404, 2003.

[4] Chung-Hsien, W., Chi-Chun, H., Te-Hsien, L. and Jhing-Fa, W., "Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis", IEEE Tr. on Audio, Sp. & Lang. Proc., vol. 14(4), pp. 1109-1116, 2006.

[5] Wouter,J. and Macon, W., "Control of Spectral Dynamics in Concatenative Speech Synthesis", IEEE Trans. on Speech and Audio Proc., vol. 9(1), pp. 30-38, 2001.

[6] Toda, T., Lu, J., Saruwatari, H. and Shikano, K. "Straight-Based Voice Conversion Algorithm Based On Gaussian Mixture Model", in ICSLP, vol. 3, pp. 279-282, 2000.

[7] Arslan, L.M. , "Speaker Transformation Algorithm using Segmental Codebooks (STASC)", Speech Communication, vol. 28, pp. 211-226, 1999.

[8] Duxans, H., "Voice Conversion applied to Text-to-Speech systems", Ph.D. Thesis. Universitat Politécnica de Catalunya, May 2006.

[9] Bänziger, T. and Scherer, K., "The role of intonation in emotional expressions", Speech Communication: 46, pp. 252-267, 2005.

[10] Montero, J.M., Gutiérrez-Arriola, J., Córdoba, R., Enríquez, E. and Pardo, J.M. "The role of pitch and tempo in emotional speech" in Improvements in speech synthesis, pp. 246-251. Ed. Wiley & Sons, 2002.

[11] Jianhua, T., Yongguo, K. and Aijun, L., "Prosody conversion from neutral speech to emotional speech", IEEE Tr. on Audio, Speech and Language Processing, Vol. 14(4), pp. 1145-1154, 2006.

[12] Montero, J.M., Gutiérrez-Arriola, J., Palazuelos, S., Enríquez, E. and Pardo, J.M., "Spanish emotional speech: from database to TTS" Proc. of ICSLP, pp. 923-925, 1998.

[13] Barra, R., Montero, J.M.,Macías, J., D'Haro, L.F., San Segundo, R. and de Córdoba, R. "Prosodic and segmental rubrics in emotion identification", Proc. of ICASSP, pp. 1085-1088, 2006.

[14] Milner, B. and Shao, X., "Speech Reconstruction from Mel-Frequency Cepstral coefficients using a source-filter model" in Proc. ICSLP, pp. 2421-2424, 2001.