EmoVoice: a System to Generate Emotions in Speech

João P. Cabral and Luís C. Oliveira

L^2F Spoken Language Systems Lab.
INESC-ID/IST,
Rua Alves Redol 9, 1000-029 Lisbon, Portugal
{jpcabral,lco}@l2f.inesc-id.pt
http://www.l2f.inesc-id.pt

Abstract

Generating emotions in speech is currently a hot topic of research given the requirement of modern human-machine interaction systems to produce expressive speech.

We present the EmoVoice system, which implements acoustic rules to simulate seven basic emotions in neutral speech. It uses the pitch-synchronous time-scaling (PSTS) of the excitation signal to change the prosody and the most relevant glottal source parameters related to voice quality. The system also transforms other parameters of the vocal source signal to produce the irregular voicing quality. The correlation of the speech parameters with the basic emotions was derived from measurements of the glottal parameters and from results reported by other authors. The evaluation of the system showed that it can generate recognizable emotions but improvements are still necessary to discriminate some pairs of emotions.

Index Terms: speech synthesis, emotions, evaluation.

1. Introduction

Concatenation-based speech synthesizers provide more natural sounding when compared with rule-based methods. However, concatenation techniques are characterized by low flexibility to modify the speech signal which is a drawback when it is intended to generate expressive speech. The most successful systems in generating emotions in speech are unit-selection speech synthesizers including speech recordings with acted emotions. Since the complexity of the system increases with the size of the corpus, the variability of the produced speech is limited. A more flexible approach is to derive acoustic rules to impose emotional content on the speech signal using signal processing techniques. This method has been widely experimented using the concatenation of diphone units and simple methods to transform prosodic parameters, which introduce acceptable degradation in speech quality, as in [1]. However, in such systems it is usually impossible to control voice quality which is also an important aspect to produce emotions in speech.

Recently, it has been shown that the glottal parameters related to voice quality are relevant to express emotions [2]. In this work we developed a system named EmoVoice, which performs high-quality prosodic transformations and changes the main parameters of the glottal source, conveying emotions in speech. It can be used to produce emotions in natural speech or synthetic speech generated by any system.

2. Acoustic rules of emotions

Typically, to derive acoustic rules of emotions, a corpus of emotional speech is recorded using professional speakers or actors. Then, the acoustic features are measured and statistically analyzed.

In this work, the acoustic correlates of emotions were derived from measurements of the glottal flow parameters of voice quality and from results published by other researchers.

2.1. Prosodic parameters

Generally, the process of measuring and examining the speech features takes a significant time and requires the appropriate software tools. In this work, the pitch, duration and energy rules were obtained based on acoustic profiles suggested in literature [3, 4, 5, 6], since there were several studies that correlated the prosodic parameters with emotions. We found some difficulties because the studies were developed in different conditions. For example, the size of the speech corpus, the speakers, the language, the type of emotions and the type of feature analysis (e.g. at the syllable or utterance level) varies. These factors were considered in the investigation and the values of the transformations were also adjusted by listening to the modified speech. The results of this work are summarized in Table 1.

Emotions	Mean F0	F0 range	Duration	Intensity
Anger	15%	30%	-16%	70%
Happiness	18%	30%	-19%	30%
Sadness	-16%	-38%	16.5%	-50%
Fear	25%	30%	-10%	0%
Surprise	15%	40%	-5%	2%
Boredom	-16%	0%	10%	-20%
Disgust	-20%	20%	-10%	-20%

Table 1: Percent variations of the prosodic parameters obtained from the literature and adjusted with our perceptual evaluations.

2.2. Voice quality parameters

Voice quality aspects are usually described in terms of qualitative descriptions or spectral patterns [6]. For example, anger and happiness is usually described as tense, blaring and breathy. Sadness is typically lax, and fear is characterized by tense and irregular voicing. In the spectral domain, happiness and sadness present increased energy in the high-frequency band while fear shows a decrease of the energy in high frequencies.

In this work, the continuous variables used to control the voice

quality were the aspiration noise, jitter, shimmer and the glottal source parameters: open quotient (OQ), speed quotient (SQ) and return quotient(RQ) [7]. The aspiration noise creates the breathy effect while the jitter and shimmer give the sensation of harsh voice. These parameters have been used to generate speech with emotions by [5] and [3]. In these studies the acoustic measures were computed from recorded emotional speech. We used these results and our perceptual evaluation of the transformed speech to derive the variation values of the jitter, shimmer and aspiration noise, presented in Table 2.

Emotions	Jitter (%)	Shimmer (%)	Aspiration noise (dB)	OQ	SQ	RQ
Anger	6	40	0	0.9	0.9	1.5
Happy	3	0	9	0.8	1	0.7
Sad	0	0	0	0.8	1.2	0.9
Fear	0	0	0	0.5	1	1
Surprise	7	0	7	1	1	1
Boredom	0	0	0	1	1	1
Disgust	0	0	0	1	1	1

Table 2: Transformation factors of the speech parameters related to voice quality.

In case of the glottal parameters, there were studies that correlated these parameters for different types of voice quality, e.g. [8, 9], but we have not found any reported measurements of these features for emotional speech.

In order to derive the acoustic variations of the OQ, SQ and RQ, these parameters were measured from an emotional speech database [10]. This way, the hard and time consuming task of building a speech corpus with emotions was avoided. However, this approach might be considered in future investigations. The speech corpus was in German and contained seven emotional states: neutral, anger, happy, sad, fear, boredom and disgust. We used the Aparat software [11] to perform the acoustic measurements. The results were obtained from 27 speech files, spoken by a male speaker, with one of the seven emotions each, and are presented in Table 3.

Emotions	**OQ**		**SQ**		**RQ**	
	Mean	Error	Mean	Error	Mean	Error
Neutral	0,891	0,013	2,288	0,260	0,076	0,024
Anger	0,885	0,022	1,921	0,360	0,097	0,045
Happy	**0,859**	0,017	**1,639**	0,310	0,097	0,038
Sad	0,869	0,012	2,023	0,214	0,070	0,026
Fear	**0,848**	0,023	**1,695**	0,410	0,096	0,042
Boredom	0,895	0,010	2,540	0,295	**0,034**	0,022
Disgust	0,881	0,015	2,168	0,610	0,068	0,042

Table 3: Mean values of the glottal flow parameters calculated for each emotion, from the speech files.

The results in Table 3 show that the OQ decreases specially for happiness and fear, which is consistent with the spectral effect of increasing energy at higher frequencies. The variation of the spectral energy with OQ is explained in [12]. Sadness also presents a decrease in the OQ which was not expected because it is normally characterized by an increase in energy of the higher order harmonics.

The decrease of the SQ for anger and happiness can be explained by the breathy quality associated with these emotions [8]. In the case of sadness the decrease of the SQ produces the lax voice. However, the emotion fear which is associated with a tense voice presents an unpredictable decrease of this parameter [9].

Finally, the return quotient is higher for angry which supports the breathy quality of this emotion [8].

The mean errors of the measures are relatively high so it is not possible to draw conclusions for all cases. Despite this, we obtained approximate values for the relevant variations of the glottal flow parameters with emotions, which are presented in Table 2.

3. System description

A schematic diagram of the EmoVoice system to transform speech emotions is shown in Figure 1.

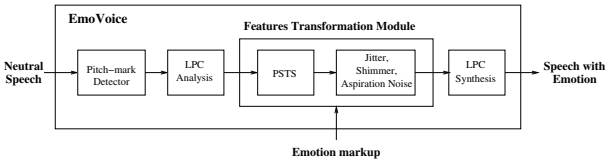

Figure 1: Diagram of the EmoVoice system.

First, the glottal epochs are computed using a robust pitch marking algorithm from the ESPS (Entropic Signal Processing System) tools. The linear prediction coefficients (LPCs) are calculated using the autocorrelation method for the Hanning-windowed segments, with 20 ms length and centered in the estimated pitch marks. The residual signal is obtained by filtering the speech signal by the time-varying error prediction filter $A(z)$.

All of the modifications are performed on the residual signal according to the transformation factors of the acoustic features that are mapped to the desired emotion. The emotions are indicated to the system by markup labels on the speech signal.

The Pitch-Synchronous Time-Scaling (PSTS) [7] method transforms the energy, duration, pitch and the glottal source parameters (OQ, SQ and RQ). This technique allows large time-scale transformations and the modification of different aspects of voice quality, preserving the speech naturalness.

Jitter is produced by adding a random component to the pitch period of the short-time signal. Shimmer is introduced by multiplying the energy envelope of the short-time signal by a random factor. To generate the aspiration noise, white gaussian noise, filtered in the perceptually significative band of 2-4 kHz, is amplitude modulated by an Hanning window with duration of the pitch period and added pitch-synchronous to the residual signal.

Finally, the transformed residual signal is passed through the time-varying all-pole filter $A(z)^{-1}$ using the LPC coefficients associated with the synthesis pitch marks to obtain the speech waveform with emotion.

A web interface was developed for demonstration of the EmoVoice system. It is accessible at `http://www.l2f.inesc-id.pt/~jpcabral/EmoVoice/`.

4. Perceptual test

To evaluate the seven basic emotions (anger, happy, sad, fear, surprise, boredom and disgust) produced by the EmoVoice system a double A-B forced choice test was conducted. Subjects were presented with an emotion name and two different stimuli (A and B), where A and B were random choices of two speech signals with emotions generated by the system. The subjects had to indicate

which of the two stimuli better matched the pretended emotion. The dual choice test was preferred over a 8-way choice in order to simplify the subjects task: we could perform more tests keeping the subjects focused.

4.1. Stimuli

Two different utterances in European Portuguese, with approximately 2 seconds each, were chosen from a speech corpus spoken by a male speaker. The corpus was recorded in a studio environment. From the two neutral signals, seven speech file were generated with induced emotions by using the EmoVoice system.

4.2. Subjects

Thirty listeners participated in the test. There were 22 males and 8 females. They consisted mainly of students and professors working at the $L2F$ Lab and other departments of INESC-ID. They were all native speakers of European Portuguese.

4.3. Experiment

This test involved eight different emotions, including the neutral emotion. Subjects were presented with 56 pairs of stimuli, corresponding to all possible combinations of one utterance with two randomly chosen emotions. They could play the signals any time they wanted but they were forced to select the stimulus that better matched the emotion name indicated in the interface panel. The same pair of stimuli was presented twice (in random order) to test the two emotions of the pair. An easy-to-use computer interface was developed for the test.

5. Results

Figure 2 show the recognition rates and the 95% confidence intervals, obtained for each emotion. For example, Figure **??** shows the rate of listeners responses that matched the intended anger emotion when compared with the remaining emotions.

In general, the simulations of the anger, sad and fear emotions were well recognized by the listeners. For these emotions, the lowest rates were obtained when anger was compared with fear (80% for anger and 75% for fear) and when the happy emotion was compared with anger and surprise (around 60%).

The results for surprise, boredom and disgust emotions were not as good, given that the lowest accuracy rates obtained were under 50%. Subjects distinguished very well the boredom emotion from all emotions but sad. Surprise was confused with the emotions with high activation level, achieving only 30% and 35% when compared with anger and happy. In general, subjects did not perceive the disgust emotion, as shown by its recognitions rates (between 40% and 60%).

6. Discussion

The anger, happiness, surprise and fear are characterized by similar prosodic variations (increased pitch, wider pitch range, slower speech rate and higher energy) which can explain the lower recognition rates obtained for some pairs of these emotions such as fear-anger, happy-anger and surprise-happiness. Also, there are emotions typically more difficult to perceive, such as happiness and disgust. It is very likely that other aspects of human expression, such as facial cues [13], are determinant to distinguish these emotions.

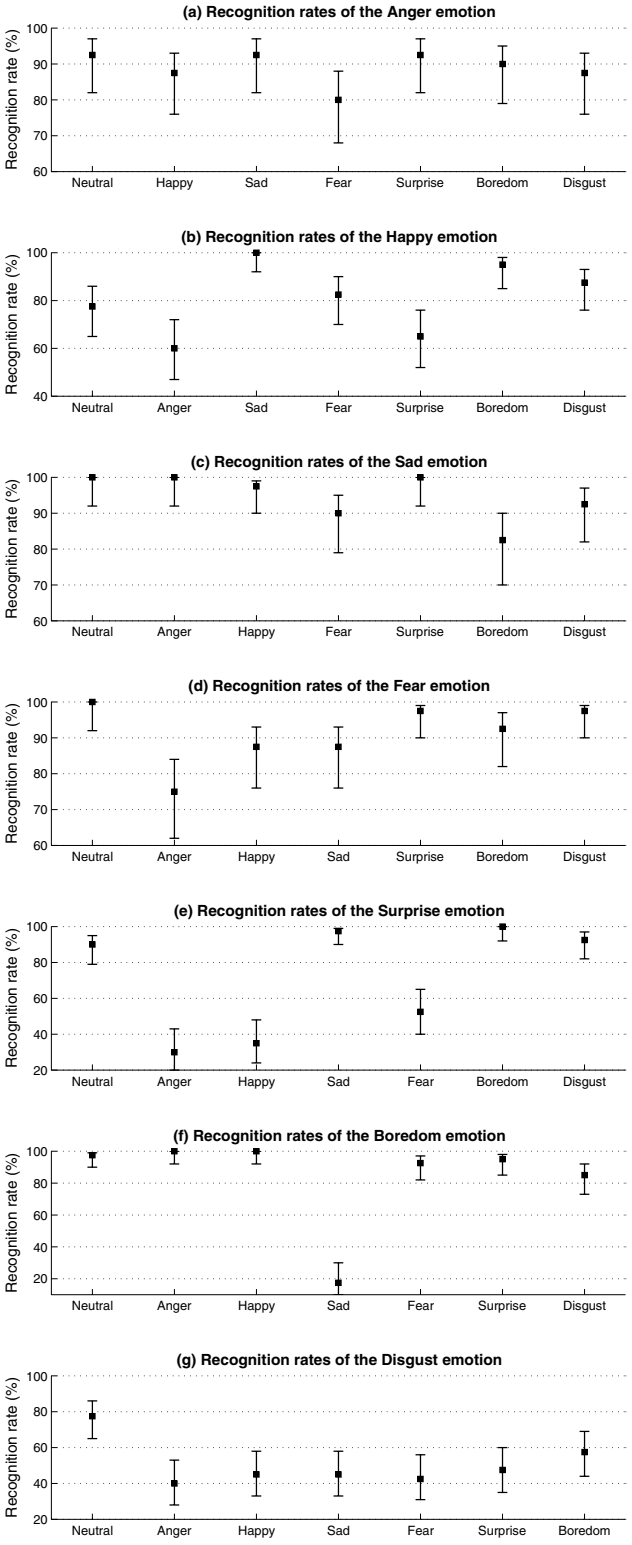

Figure 2: Recognition rates obtained for each emotion when compared with the other emotions, in the A-B listening test.

Even so, we think that for some emotions the acoustic transformations used were not sufficient. This is the case of surprise: when it was compared with anger and happiness, it obtained very low recognition rates. There are studies which indicate that the upward direction of the pitch contour [14] and the irregular pattern of pauses [15] are characteristics of the surprise emotion which differ from anger, happy and fear. Thus, the simulation of surprise by the EmoVoice system could be improved by including more prosodic transformations.

Also, the perceptual results show that the boredom emotion was almost not recognized when compared with sad, and that fear was mainly confused with anger. Since the prosodic features are very similar in the two pairs of emotions, more voice quality cues seem to be necessary to improve the accuracy rates of boredom and fear. For example, the simulation of the voice quality creaky for boredom and whispery for fear, as in [16], could improve the recognition rates obtained by these emotions.

There are other factors that could have undesirably influenced the subject's recognition of the simulated emotions. One is the distortion caused by the acoustic transformations when simulating intense vocal emotions. For example, the aspiration noise introduced to simulate happiness and surprise produces some degradation of the speech quality. Another limitation is the listener's difficulty to separate the sound from the semantic meaning of the utterance. Although, in the test, the utterances were chosen to convey any of the tested emotions, subjects may not be used to associate emotions like fear, anger or disgust to sentences without the appropriate semantic content.

7. Conclusions and Future Work

To study the correlation between the speech parameters and emotional states, a speech modification system was built, called EmoVoice. It uses the PSTS method to transform important prosodic and voice quality parameters of emotions, preserving the naturalness.

The values of the acoustic transformations for each emotion were derived from the literature and adjusted by listening to the modified speech. In case of the glottal source parameters, no results where found in the literature that correlate them directly with emotions. Thus, numerical rules were obtained by measuring the glottal parameters from an emotional speech database.

EmoVoice generates seven emotions: angry, happiness, sadness, fear, surprise, boredom and disgust. The system was evaluated by a formal A-B perceptual test. The results showed that the angry, happiness, sadness and fear obtained good recognition rates, similar to those reported by other speech researchers. Boredom was confused with sadness and surprise was confused with angry, happiness and fear. The results obtained by these two emotions were not as good as expected and suggest further study to identify the speech parameters that are more relevant to simulate their effects. Disgust was not recognized by the listeners which strengths the hypothesis that this emotion has weak vocal correlation.

The system uses constant factors to transform the speech parameters for each emotion. We expect to enhance its performance by applying non-uniform transformations along the utterance that are characteristic of emotions. Also, we intend to extend the transformations to other acoustic features, such as pauses, articulation and spectral parameters. We think that further tests must be conducted to evaluate the performance of the system to produce the basic emotions in synthetic speech.

8. Acknowledgements

This work was partially funded by the Portuguese Foundation for Science and Technology (FCT/FEDER/POSC).

9. References

[1] I. R. Murray and M. D. Edgington, "Rule-based emotion synthesis using concatenated speech," in *ISCA Workshop on Speech and Emotion*, Northern Ireland, 2000, pp. 173–177.

[2] R. Fernandez and R. Picard, "Classical and novel discriminant features for affect recognition from speech," in *Interspeech*, Lisbon, Portugal, 2005, pp. 473–476.

[3] C. Drioli, G. Tisato, P. Cosi, and F. Tesser, "Emotions and voice quality: Experiments with sinusoidal modeling," in *ITRW VOQUAL'03*, Switzerland, August 2003, pp. 127–132.

[4] E. Zovato, A. Pacchiotti, S. Quazza, and S. Sandri, "Towards emotional speech synthesis: a rule based approach," in *ISCA SSW5*, Pittsburgh, USA, June 2004, pp. 219–220.

[5] S.P. Whiteside, "Simulated emotions: an acoustic study of voice and perturbation measures," in *ICSLP*, Sydney, Australia, 1998, pp. 699–703.

[6] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," in *IEEE Signal Proc. Mag.*, 2001, vol. 18, pp. 32–80.

[7] J. P. Cabral and L. C. Oliveira, "Pitch-synchronous time-scaling for prosodic and voice quality transformations," in *Interspeech*, Lisbon, Portugal, 2005, pp. 1137–1140.

[8] D.G. Childers, "Glottal source modeling for voice conversion," *Speech Comm.*, vol. 16, no. 2, pp. 127–138, 1995.

[9] P. Alku, "Parameterisation methods of the glottal flow estimated by inverse filtering," in *ITRW (VOQUAL'03)*, Geneva, Switzerland, August 2003, pp. 81–88.

[10] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlemeier, and B. Weiss, "A database of german emotional speech," in *Interspeech*, Lisbon, Portugal, 2005, pp. 1517–1520.

[11] M. Airas, H. Pulakka, T. Bäckström, and P. Alku, "A toolkit for voice inverse filtering and parametrisation," in *Interspeech*, Lisbon, Portugal, 2005, pp. 2145–2148.

[12] B. Doval and C. d'Alessandro, "Spectral correlates of glottal waveform models: an analytic study," in *ICASSP*, Munich, Germany, 1997, pp. 1295–1299.

[13] K. R. Scherer, "Vocal communication of emotions: a review of research paradigms," *Speech Comm.*, vol. 40, pp. 227–256, 2003.

[14] R. Aishah, M. Izani, K. Komiya, and K. Ryoichi, "Emotion pitch variation analysis in malay and english voice samples," in *9th Asia Pacific Conference on Comm.*, September 2003.

[15] C. Alm and R. Sproat, "Perceptions of emotions in expressive storytelling," in *Interspeech*, Lisbon, 2005, pp. 533–536.

[16] I. Yanushevskaya, C. Gobl, and A. Chasaide, "Voice quality and f_0 cues for affect expression: implications for synthesis," in *Interspeech*, Portugal, 2005, pp. 1849–1852.