Automatic Phonetic Segmentation of Spanish Emotional Speech

A. Gallardo-Antolín [1], *R. Barra* [2], *M. Schröder* [3], *S. Krstulovic* [3], *J.M. Montero* [2]

[1] Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Spain
[2] Speech Technology Group, Universidad Politécnica de Madrid, Spain
[3] DFKI GmbH, Saarbrücken, Germany

gallardo@tsc.uc3m.es, barra@die.upm.es, {schroed, sacha}@dfki.de, juancho@die.upm.es

10.21437/Interspeech.2007-547

Abstract

To achieve high quality synthetic emotional speech, unit-selection is the state-of-the-art technique. Nevertheless, a large expensive phonetically-segmented corpus is needed, and cost-effective automatic techniques should be studied. According to the HMM experiments in this paper: segmentation performance can depend heavily on the segmental or prosodic nature of the intended emotion (segmental emotions are more difficult to segment than prosodic ones), several emotions should be combined to obtain a larger training set (especially when prosodic emotions are involved; this is especially true for small training sets) and a combination of emphatic and non-emphatic emotional recordings (short sentences vs. long paragraphs) can degrade overall performance.

Index Terms: expressive speech, automatic phonetic segmentation, emotional speech synthesis

1. Introduction

One of the most important trends on Speech Technologies is the synthesis of emotional speech, which can provide naturalness and variability to synthetic speech. The use of concatenation-based unit-selection strategies provides high-quality synthetic speech, although they are not cost-effective techniques without automatic support tools. As a large segmented and labelled corpus must be available and hand-labelling and segmenting are labour-intensive tasks, the use of unit-selection in emotional synthesis has to be based on conversion techniques (in order to minimize the effort to create emotional voices from a neutral one) or the use of efficient automatic tools to minimize costs.

The main tasks involved in the development of a new voice in this kind of systems are: revision of the phonetic transcription (this is a relatively low-cost task which can be carried out in parallel with the recording sessions), pitch-epoch extraction (easy to obtain from the EGG signal when recorded in parallel with the speech signal) and phonetic labelling (definitively, the most expensive task, especially when consistency is necessary; more powerful phonetic segmentation tools should be adapted to the needs of emotional speech).

Although many papers have addressed automatic phonetic segmentation, no paper has been devoted to how emotional speech affects the segmentation process. However, emotions severely alter speech characteristics and could have a great influence on the performance of phonetic segmentation systems.

In recent years, several methods have been proposed for tackling the automatic phonetic segmentation problem when the phonetic transcription is available (also called the linguistically-constrained approach). The most common strategies are based on Dynamic Time Warping (DTW) or Hidden Markov Models (HMMs) techniques. The DTW algorithm is used for carrying out a temporal alignment between the sentence to be segmented and an already segmented version (usually, synthetic speech) for which time marks between phones are known [1]. In HMM-based segmentation systems, the automatic segmentation is generated through a forced alignment between known phonetic transcriptions and recorded speech data [1], [2]. Other approaches consider hybrid techniques such as HMM/ANN-based systems [2] or a refinement of HMM segmentation [1], [3]. In this paper, we have chosen the HMM-based system because, nowadays, it is one of most widely employed.

This paper is organized as follows: Section 2 describes how segmentation can be affected by expressive speech. In Section 3, the Spanish emotional database is introduced. Section 4 describes the segmentation system and preliminary results. Section 5 fully describes emotional speech segmentation experiments and, finally, conclusions are commented in Section 6.

2. Issues on the automatic segmentation of expressive speech

This paper focuses on automatic phonetic segmentation of emotional or expressive speech oriented to speech synthesis systems. To the best of our knowledge, there is no study about this task in the literature. The aim of this paper is to address the following three issues.

Firstly, we will study the behavior of the segmentation process when processing each type of the emotion considered, in order to determine which emotions can be best segmented.

Secondly, in this paper we will study the behavior of the system when emotions in the training and evaluation stages are not the same. It is well known that speech recognition system performance dramatically decreases when there is a mismatch between training and evaluation conditions due to environmental noise or other types of distortion (for example, speech under stress or different speaking styles [4]). As the HMM-based segmentation system shares the main principles with conventional ASR systems, we may hypothesize that this drawback may also be present in the segmentation system. However, it will be shown that this is not generally true on emotional speech processing.

Finally, another main challenge related to expressive speech segmentation is the small amount of available emotional data for properly training the segmentation system. In this context, we will study the influence of the size and variability of the training set (containing different emotions or speaking styles) on the accuracy of the segmentation system for expressive speech.

3. Emotional speech and database description

In this work, we have used the Spanish Emotional Speech corpus (SES) [5]. 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 extraction program was used, but the outcome was manually revised using a graphical audio-editor program, also used for locating and labeling phonemes boundaries.

The assessment of the emotional voice was aimed at judging the appropriateness of SES recordings as a model for recognizable emotional speech [6]. Perceptual copy-synthesis experiments [5], [7], where durations and stylized F0 contours were mixed with emotional or neutral diphones, showed the different speech characteristics of each emotion. Table 1 shows that *cold anger* is the most identifiable emotion using just segmental information (95.6%) and *surprise* does not present any clear segmental pattern (9.5%).

EMOTION	Emotion diphones + Neutral prosody	Neutral diphones + emotion prosody
Happiness	52.4%	19%
Cold anger	95.6%	7.1%
Surprise	9.5%	76.2%
Sadness	45.2%	66.5%

Table 1. *A mixed-emotion perceptual test.*

Emotional patterns were also evaluated by means of automatic identification experiments [8]. Emotional information was analyzed using segmental (MFCC) and prosodic information (F0-related statistics). When both sources of information were combined, better classification rates were obtained. Table 2 shows the identification results obtained with both segmental and prosodic features. It can be observed that they are correlated with the perceptual experiments in Table 1. In particular, results confirm the segmental and prosodic nature of *anger* and *surprise,* respectively. Nevertheless, *sadness* is fully identifiable from the MFCC vector without the support of any prosodic feature. Therefore, all the emotions present certain discriminative segmental characteristics that allow them to be classified by an automatic system, although sometimes these segmental features are identified by human listeners.

EMOTION	Based on MFCC	Based on F0 statistics
Happiness	91.1%	44.4%
Anger	97.8%	48.9%
Surprise	66.9%	95.6%
Sadness	100%	75.6%
Neutral	73.3%	66.7%

Table 2. *Automatic identification experiments.*

These emotional rubrics and the different results between perceptual and automatic experiments, suggest that segmental differences could affect the segmentation performance, so that the optimal segmentation strategy for each emotion should be emotion-dependent.

4. Automatic segmentation system

4.1. System description

The segmentation system is based on Hidden Markov Models (HMMs) and it has been developed using the HTK toolkit [9]. Automatic segmentation is generated by carrying out a forced alignment between speech data and the corresponding phonetic transcription by means of the Viterbi algorithm. Boundaries between phonemes are placed at the time instants in which transitions between the corresponding HMM models occur [1].

We have considered a repertory of 29 or 50 Spanish phonemes, each of them represented by a left-to-right context-independent continuous density HMM (CI HMM) model with three states. More complex models could be used (for example, triphones); however, the limited amount of training data makes more adequate to use simpler models with a few gaussians per state [10]. Models are trained using the conventional Baum-Welch algorithm on the phonetic transcription of the sentences in the training set, but not the manual time marks. It is worth noting that, in this sense, we can consider this process to be an unsupervised one because no information about the manual segmentation of the training database is used.

As feature vectors, the system uses 12 MFCCs, 1 log-energy and their corresponding first and second derivatives. Parameters are extracted with a 25 ms analysis window and a 5 ms delay between frames.

4.2. Preliminary experiments

In order to guarantee the correct performance of the system, we have carried out a preliminary experimentation with the NatVox database. This database was intended for restricted-domain synthesis and it was recorded by the Speech Technology Group at Universidad Politécnica de Madrid. It comprises 922 sentences read by a female speaker in neutral-style Castilian Spanish. We have used a small part of the database (about 20 minutes of speech), for both training and evaluation.

The system is evaluated by comparing the time marks produced by the automatic system and the manual segmentation generated by a human expert. The segmentation error is the percentage of boundaries which are incorrectly placed when compared to the reference. Usually, we must allow a small deviation (called tolerance) between the automatic and manual marks, in order to take into account possible inconsistencies in the manual segmentation data.

Results obtained with the HMM-based system with several gaussians per state (from 1 to 5) and for different tolerances (from 5 ms to 25 ms) are presented in Table 3. As can be observed, the best performance is obtained with a mixture of two gaussians. In order to check whether the models are adequately trained or not, we carried out a set of automatic speech recognition experiments. The results obtained are shown in the column labelled as "PHON. ERROR (%)"in Table 3. Note that, in these experiments, the vocabulary of the task was only composed by the repertory of phonemes previously mentioned, so the recognized phoneme sequence could contain substitution, deletion and insertion errors. In this case, as the number of gaussians increases, the

phoneme recognition error decreases, so we can conclude that models are correctly trained. From these experiments, we can draw the conclusion that a good acoustic modeling for supervised speech recognition is not necessarily a good acoustic modeling for unsupervised phonetic segmentation.

# of gaussians	SEGMENTATION ERROR (%)					PHON. ERROR (%)
	5 ms	*10 ms*	*15 ms*	*20 ms*	*25 ms*	
1 g.	68.63%	42.31%	23.49%	13.08%	8.23%	30.83%
2 g.	67.88%	40.13%	22.01%	12.65%	8.05%	22.09%
3 g.	70.95%	45.29%	26.27%	15.04%	9.06%	17.21%
4 g.	72.24%	47.01%	27.52%	16.18%	10.01%	13.43%
5 g.	72.60%	47.90%	28.30%	16.65%	10.45%	12.14%

Table 3. *Segmentation error rate (%) for different tolerances and phoneme recognition error (%) on the NatVox database.*

We have computed segmentation error statistics per transition between phonemes grouped in broad classes. This information can be useful for determining whether HMMs perform better or worse for some transitions than for others and whether temporal mark shifts of certain transitions presents the same bias. From these statistics, we have observed that almost 33% of boundary errors occur in vowel-vowel, vowel-nasal and vowel-silence transitions. This fact confirms the habitual discrepancies with respect to the manual transcriptions. These transitions are also difficult to segment for human experts.

5. Experiments on the SES database

In this section, we describe the experiments carried out on the SES database. The segmentation system used is the HMM-based one described in subsection 4.1 with a set of 50 phonemes modelled by CI HMMs with mixtures of three gaussians per state. All the experiments described below correspond to 20 ms tolerance. For each emotion, the test set comprises three paragraphs, while the training set (except for the experiment in subsection 5.1) comprised all of them.

5.1. Influence of the training set

In Table 4, we show results on three sets of training material (short segmented sentences, a small set of segmented paragraphs and a larger set of non-segmented paragraphs) with only set for training (segmented paragraphs). Using the larger and smaller set of paragraphs resulted in a better performance when compared to just using the small set (because of the shortage of training data). However, the combined training set that comprises both paragraphs and sentences did not significantly improve the paragraphs score, because of the different way of emphasizing the same emotion in paragraphs and in short sentences. The improvement in surprise is due to the prosodic nature of this emotion: the prosody of surprised paragraphs and sentences are rather different, but the segmental components (which most influence segmentation results) are quite similar.

As can be observed, in all the emotions, the increase of training data with the same speaking style ("small set" vs. "all") produces an improvement in the performance of the system. This improvement is especially high with *surprise* speech, in which the segmentation error rate decreases from 14.46% to 9.75%. On the contrary, the lower improvement is achieved with *sadness*.

However, when adding training data with a different speaking style (short sentences) to the paragraphs, the segmentation system performance only improves slightly for *happiness*, *sadness* and *neutral* speech. Even the segmentation error rate slightly increases in several cases (for example, with *neutral* speech, the error increases from 8.60 % to 9.04 % when mixing all the paragraphs and the sentences). *Surprise* is the only emotion that gets profit from all the training data available.

From these experiments we can conclude that the increase of the training data with material of the same speaking style is beneficial for the segmentation system, while using data of different styles does not help to improve the performance of the system, with the exception of *surprise*.

TRAIN SET	TEST EMOTION			
	Happiness	*Surprise*	*Sadness*	*Neutral*
Paragraph (small set)	9.90%	14.46%	15.29%	10.71%
Paragraph (small set) + Sentences	10.15%	10.78%	15.21%	10.10%
Paragraph (all)	8.87%	9.75%	14.96%	8.60%
Paragraph (all) + Sentences	8.70%	8.21%	14.45%	9.04%

Table 4. *Segmentation error rate (%) on SES database for 20 ms tolerance and several training sets*

5.2. Influence of the emotion type

In the experiments described in this section, we have considered three types of emotional speech: happiness, surprise and sadness, as well as a neutral speaking style.

Table 5 shows the confusion matrix obtained when trying to segment each emotion with a segmentation system that was trained with another type of emotional speech.

TRAIN EMOTION	TEST EMOTION			
	Happiness	*Surprise*	*Sadness*	*Neutral*
Happiness	8.87%	8.73%	12.18%	6.58%
Surprise	8.79%	9.75%	15.13%	7.81%
Sadness	7.42%	12.15%	14.96%	7.55%
Neutral	9.22%	14.37%	15.71%	8.60%
All	7.94%	7.78%	12.77%	7.55%

Table 5. *Segmentation error rate (%) for several train-test configurations (20 ms tolerance).*

From Table 5, it can be observed that, surprisingly, no emotion is best segmented when using the system trained with the same emotional speech. In fact, the best training data are *happy* recordings which produce the best performance for segmenting *surprise*, *sadness* and *neutral* speech.

Sadness was the most difficult to segment emotion. In the subjective evaluation sessions, *happiness* was the most difficult to identify emotion. On the contrary, neutral-style recordings are the best segmented data, followed by happiness and surprise. Neutral and happiness are relatively insensitive to the training emotion. However, surprise performs better with happiness and surprise data and poorly with sadness and with neutral speech.

Sadness was the worst segmented emotion, obtaining the best results when segmented with happiness models. Negative emotions (sadness and cold-anger) proved to be the most segmental emotions in the listening and automatic tests. Emotions are called segmental when they are mainly identified because of their non-prosodic characteristics. As angry paragraphs were not manually segmented, we carried out an evaluation experiment with the segmented short sentences; resulting in angry and sad recordings to be poorly segmented (when using models from all the emotional paragraphs, sad and angry scores under 20 ms were 16.55% and 16.28%, respectively, while the remaining emotions scored from 12.07% to 13.10%).

The fastest emotion, happiness, seems to be the best emotion for training: fast-speech models can better recognize slow-speech than the other way around. Similar conclusions can be drawn from other experimental configurations: different number of gaussians or segmentation tolerance.

5.3. Relationship between speech recognition and segmentation accuracies

Table 6 shows the phone recognition error rate obtained when trying to recognize phonemes of each emotion with a system that was trained with another type of emotional speech.

	TEST EMOTION			
TRAIN EMOTION	***Happiness***	***Surprise***	***Sadness***	***Neutral***
Happiness	18.30%	39.22%	44.48%	46.76%
Surprise	51.77%	16.71%	87.31%	56.10%
Sadness	42.70%	51.15%	18.95%	56.76%
Neutral	47.98%	46.80%	68.53%	17.77%

Table 6. *Phone recognition error rate (%) on the SES database for different train-test configurations.*

As phonetic segmentation is an unsupervised task (not supervised as ASR), segmentation performance cannot be predicted from recognition results even on the same database because segmentation is a side effect of the recognition process. Segmentation is not severely affected by train-test mismatches because it cannot be over-trained and it has not been ML-optimized for segmentation, but for phonetic decoding.

6. Discussion and conclusions

The emotional speech segmentation experiments show that it is worthwhile combining all similar emotional material in a large training set, to reduce the shortage of data (9.97% is the mean average score when there are as many training sets as available emotions; and we get 9.16 when a 4-emotion training set is used). When we used just the small segmented paragraph set, the improvement by combination is greater.

However, it is not worthwhile combining these recordings to make a larger training set, because there are significant differences between the way of simulating the same emotions in short sentences and long paragraphs.

In this unsupervised task, a small number of gaussians (1-3) results in better performance, even when a greater training set is used.

Regarding cross-emotion segmentation, experiments suggest that neutral and happy recordings are the best training material for the other emotions, and surprise and sadness perform poorer.

From the experiments and discussion, we can conclude that:

- Segmentation performance depends on the segmental or prosodic nature of the intended emotion: segmental emotions are more difficult to segment than prosodic ones.
- Several emotions should be combined to obtain a larger training set, especially when prosodic emotions are involved. This is especially true for smaller training sets.
- A combination of emphatic and non-emphatic emotional recordings (short sentences vs. long paragraphs) can degrade overall performance.

Some future research lines are:

- To carry out experiments on a larger multi-speaker database.
- To use MMC-based unsupervised techniques to improve HMM segmentation performance.
- To try adaptation techniques to make use of available multi-speaker neutral speech databases.

7. Acknowledgement

This work has been partially funded by UC3M-CAM under contract CCG06-UC3M/TIC-0812, by the Spanish Ministry of Education and Science under contract DPI2004-07908-C02-02 (ROBINT) and by UPM-CAM under contract CCG06-UPM/CAM-516 (TINA).

8. References

[1] Adell, J., Bonafonte, A., Gómez, J. A. and Castro, M. J., "Comparative Study of Automatic Phone Segmentation Methods for TTS", Proc. of ICASSP'05, vol. 1, pp. 309-312, Philadelphia, USA, 2005.

[2] Malfrère, F., Deroo, O., Dutoit, T. and Ris, C., "Phonetic Alignment: Speech Synthesis-Based vs. Viterbi-Based", Speech Communication, vol. 40 (4), pp. 503-515, 2003.

[3] Zhao, Y., Wang, L., Chu, M., Soong, F. K. and Cao, Z., "Refining Phoneme Segmentations Using Speaker-Adaptive Context Dependent Boundary Models", Proc. of INTERSPEECH'05, pp. 2557-2560, 2005.

[4] Bou-Ghazale, S.E. and Hansen, J. H. L., "A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress", IEEE Trans. on Speech and Audio Proc. vol. 8 (4), pp. 429-442, 2000.

[5] Montero, J. M., Gutiérrez-Arriola, J., Palazuelos, S., Enríquez, E., Aguilera, S., and Pardo, J. M., "Emotional Speech Synthesis: From Speech Database to TTS", Proc. of ICSLP'98, pp. 923-925, Sydney, Australia, 1998.

[6] Montero, J. M., Gutiérrez-Arriola, J., Colás, J., Enríquez, E., Pardo J. M., "Analysis and Modelling of Emotional Speech in Spanish", Proc. of ICPhs'99, vol. II pp. 957-960, San Francisco, EEUU, 1999.

[7] Montero, J. M., Gutiérrez-Arriola, J., Córdoba, R., Enríquez, E., Pardo, J. M. "The Role of Pitch and Tempo in Emotional Speech", in Improvements in Speech Synthesis, pp. 246-251. Ed. Wiley & Sons, 2002.

[8] Barra, R., Montero, J. M., Macías-Guarasa, J., D'Haro, L. F., San-Segundo, R. and Córdoba, R., "Prosodic and Segmental Rubrics in Emotion Identification", Proc. of ICASSP'06, vol. 1, pp. 1085-1088, 2006.

[9] Young, S., HTK-Hidden Markov Model toolkit (ver 2.1), Cambridge University (1995).

[10] Romsdorfer H. and Pfister, B., "Phonetic Labeling and Segmentation of Mixed-Lingual Prosody Databases", Proc. of INTERSPEECH'05, pp. 3281-3284, 2005.