Corpus Design, Recording and Phonetic Analysis of Greek Emotional Database

Nikos Fakotakis

Wire Communications Laboratory
Department of Electrical and Computer Engineering
University of Patras
26500 Rion, Patras, Greece
fakotaki@wcl.ee.upatras.gr

Abstract

A Greek Emotional Speech Database has been recorded and analyzed. The scope of this work is to study through this database the prosodical phenomena as a function of each emotional state and the application to a text to speech synthesis system. Our database contains recordings of a female professional actress. We used an actress for this task because in order to faithfully simulate a number of emotions. The simulated emotions for our database were sadness, anger, fear, joy and a neutral session. The recordings consisted of ten single words, twenty short sentences, twenty five long sentences and twelve passages of fluent speech (ranging from three to five sentences each). Following the recordings, a listening test was performed to test whether normal listeners could identify the type of emotion that characterized the recorded utterances. Six qualified listeners were used, both men and women, of different ages, from several social environments.

1. Introduction

When compared to human speech, synthesized speech is distinguished by insufficient intelligibility, inappropriate prosody and inadequate expressiveness. These are serious drawbacks for conversational human-machine interfaces. Prosody-intonation (melody) and rhythm, clarifies syntactic structures, disambiguates meaning and helps in discourse flow control. Moreover aspects like expressiveness or affect, provide information about the speaker's mental state and intentions beyond what is revealed by word content.

The quality of synthetic speech has been greatly improved by the continuous research of the speech scientists. Nevertheless, most of these improvements were aimed at simulating natural speech as that uttered by a professional announcer reading natural text in a neutral speaking style. Because of mimicking this style, the synthetic voice results to be rather monotonous, suitable for some man-machine applications, but not for a vocal prosthesis device such as the communicators used by disabled people.

Synthesized speech is mainly distinguished by a lower intelligibility, a not natural prosody and lack of expressiveness. These are important drawbacks for computer human speech communication.

Our work comprises a systematic study of speech with emotional expression to model the effects of emotion on signal level. The scope of this research is to improve the naturalness of voice in text to speech systems.

Emotions are marked by three main operations:

- they reflect the result of concrete stimulus in relation to the needs and the preferences of individuals.
- they prepare bodily and psychologically the organism for concrete energies and
- they transmit the person's psychological situation in the surrounding environment

The major obstacle in the research of human emotions is the difficulty to describe them with a strict way (i.e. there is a degree of subjectiveness).

When it comes to assembling data for the study of expressive content, researchers are torn between a number of different methods. First approach is to ask actors or non actors to speak spontaneously using particular modes of expression. A second approach is to ask actors or non actors to read out utterances and depending on their content causing the reader to express a particular emotion. Third, bring an emotion in a speaking while talking and last to start conversations hoping that particular modes of expression will be captured.

Each of these techniques has its advantages and disadvantages. Regarding the last two techniques ethical drawbacks arise. For our study we have chosen the first approach.

Greek emotional speech database has been recorded under laboratory conditions, the speech corpora were declaimed by a professional Greek actress following a standard data recording procedure. This was necessary in order to systematically record the same utterance with different emotional contents. It is shown in (Montero etal 1998) that recordings with actors are good approximations to true emotional speech.

In this work we give the detailed description and the composition of an emotional speech database for Greek language.

2. Emotion Categories

Emotions have been categorized as 'basic' or 'non-basic'. Regarding non-basic emotions, they are classified variously as 'blends', 'combinations', 'mixed', or 'secondary'. States of emotions may be ranked as dimensions 'on the basis of within-category strength' as, for example, when irritation as a mild anger is different in perceived intensity from rage (Tatham & Morton, 2004).

The main task for emotional speech research is to spot the differing types. Emotional state identification has been compressed to the grouping of four to eight basic emotions. But the difficulty is how to characterize what appears to be overlapping emotion categories such as reports of an internal state.

In his work Tomkins (1962) suggested eight emotions, based on deriving universals of modes of expression. A graphical representation of eight primary emotions and the relation between them in terms of blends and intensity was presented by Plutchik (1984), however in a later study (Plutchik 1994) they were reduced to six 'basic emotions'. Johnson-Laird and Oatley (1992) examined the kind of words we have for emotions, and arrived at five for basic emotions-similar list to Ekman's, minus 'surprise'-but later revised this to number to four: happiness, anger, sadness, and fear (Oatley and Johnson-Laird 1998). For the choice of the emotional states that are included in our data base we have tagged on the work of Oatley et al. (1998). Therefore in our recordings we tried to capture the emotions of happiness, anger, sadness and fear. A neutral session was also recorded.

Although it can be claimed that it is not possible to speak without expression - even the so-called 'deadpan' speech, (Tatham & Morton, 2004) has expression, suggesting for example that no particular expression is intended, the concept is used as a kind of baseline to begin a characterization of expression. In this approach, expression is seen as a modification of baseline neutral speech, or some kind of overlay imposed on this baseline. We think of neutrality as some kind of carrier which gets modulated to reveal other expressions. This is a useful way of modeling the relationship between the different expressions, since in such a model they would be characterized as modulating the carrier differently.

3. Databases of Emotional Speech

The main reasons of research in emotional speech data is speech synthesis and recognition. A set of 32 databases of emotional speech were reviewed by Ververidis (2003) the following results, regarding the purpose of construction of the databases, were obtained.

Research Area	# databases
ASR application	17
Emotional TTS	10
Medical applications	4
Emotion perception	8
Speech under stress	2
Teaching applications	1

Table 1: Emotional database research area application.

Regarding the language used in recordings, as expected, English is the most frequent, followed by German. The number of databases recorded in each language is tabulated in table 2.

Language	# Databases
English	11
German	7
Japanese	3
Spanish	3
Dutch	2
French	1
Greek	1
Sweden	1
Hebrew	1
Danish	1
Slovenian	1
Chinese	1
Russian	1
English and German	1
Multi – Language	1
Artificial Language	1

Table 2: Language used in recording emotional speech.

The most common emotions that can be found in the existing emotional databases are tabulated in table 3.

Emotion	# Databases
Anger	26
Sadness	22
Happiness	13
Fear	13
Disgust	10
Joy	9
Surprise	6
Boredom	5
Stress	3
Contempt	2
Dissatisfaction	2
Shame, pride, worry, …	1

Table 3: Emotional recorded in databases.

4. Database design

The emotional speech database for Greek was recorded in a professional studio in Athens, following a standard data recording procedure. The studio was acoustically damped and the operators could get in contact with the actress via an intercommunication system. The recordings were monophonic with 44.1 kHz sampling frequency and a 16 bit resolution.

4.1 Speakers

As mentioned in introduction a professional actress familiar with radio theater was employed for the enunciation of the text corpus. The thirty year old speaker that was recorded for the database has the standard Greek accent as spoken in Athens and has been a professional actress for almost ten years.

To avoid the interference of a listener's decision on the emotional contents due to semantically meaning, we attempted to construct semantically neutral sentences. The use of identical utterances spoken with different expressive content is designed to normalize out the effects of non-expressive meaning in the utterances.

The actress was asked to use her every day way of emotional expression and not an exaggerated theatrical approach. She was instructed to read all the utterances with one emotion then change it and start over again. In

that way we wanted to assure that the speaker did not have to change emotion more than five times (expressing sadness, anger, fear, joy and neutral).

4.2 Data Resources

For the study and analysis of prosody, first we chose a number of sentences that composed our corpus. The corpus was designed in a way that each phoneme resides in various positions in a word (initial, medial, final) in that way the extraction of them is possible and can be used as a structural element in a text-to-speech system (TTS) inventory.

Sentences were extracted from passages and newspapers or were set up by a professional linguist. Finally the corpus was completed by adding ten single words, twenty short sentences, twenty five long sentences and twelve passages of fluent speech (ranging from three to five sentences each). All sentences were emotionally neutral, meaning that they do not convey any emotional charge through lexical, syntactical or semantical means.

A session of nonsense words was also recorded uttered in a neutral way so as to be used as carriers for concatenative TTS inventory units.

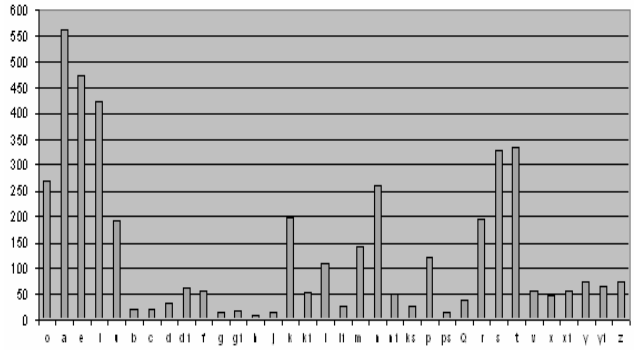

Figure 1: Database Phonemes Frequencies

5. Evaluation of the Natural Voice

Following the recordings, a listening test was performed to test whether normal listeners could identify the type of emotion that characterized the recorded utterances. Six qualified listeners were used both men and women, of different ages, from several social environments.

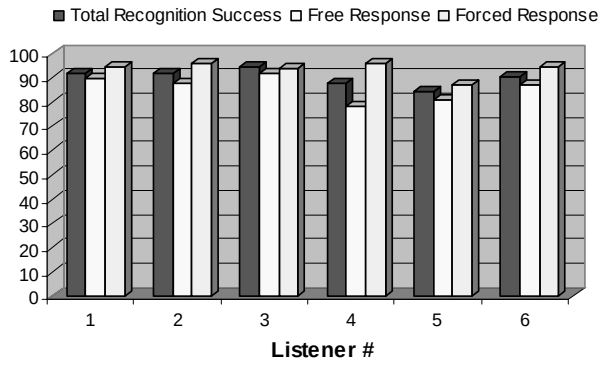

Figure 2: Free and forced response test results per listener

As a stimuli for the evaluation, the whole database recordings were played randomly to the listeners without knowing the actual number of utterances for each emotion.

Evaluation of the database was the result of a two-part procedure. In the first place a free response test was held. Each one of the listeners was labeling each utterance with whatever emotion had found appropriate. In the second part of the evaluation they were forced to choose between the four emotions that where included in our database. The results which are depicted in figure 2 describe each listener's total score. The overall results for both test for all the listeners, were 88,2%, the free response overall score was 86,88% and the forced response test was 89,63%.

The results of the evaluation regarding each emotion category are depicted in figure 3. We can see as expected that the emotion of anger and anger had the highest success score.

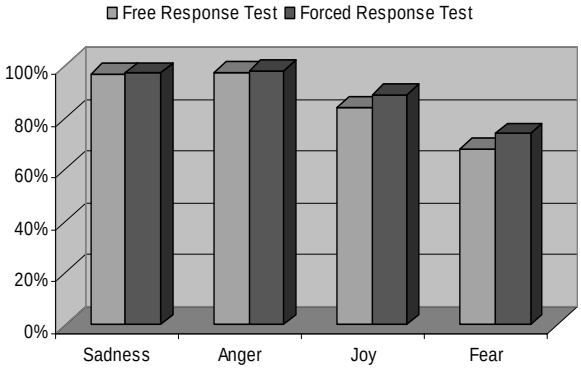

Figure 3: Free and forced response test results per emotion.

6. Future Work

Our next step is the parameterization of each emotion in order the task of creating them in a synthetic speech to be feasible. In view of finding such a description of phonetic operations under the effect of concrete sentimental situations, contemporary researchers have studied various parameter estimation techniques (effect on F0 contour, variation in number of pauses, length of pauses, ratio of pause duration to total phonation time and speech rate, fundamental frequency-its median value, the average pitch range, the rate of F0 change) (Murray & Arnott, 1995).

Taking into account all the above we concluded in a set of features for the description of each emotional state composed of the:

- Fundamental frequency F0
- Speech intensity
- Speech duration in various levels (sentence, word, phoneme)

The above parameters were adopted as the most efficient and most important factors for the recognition and variation of the emotions that were recorded in our database.

Extracting such information can be applied to a speech synthesis system prosody module to create emotions in the synthetic speech.

7. Conclusion

Demand for natural sounding synthetic speech has formed the need to model and synthesize emotions. For this purpose we created an emotional speech database for the Greek language. The recorded database represents a good base for emotional speech analysis and is also usable for emotional speech synthesis. Some improvements we could apply include "undercover" recording of real emotions in natural environments, automation of the post-processing phase (labeling, segmentation) and additional recordings of amateur speakers for emotional consistency analysis.

References

Arvaniti, A., Baltazani, M., GREEK ToBI: A System for the Annotation of Greek Speech Corpora, VOL. II, 555-562, LREC 2000.

Banse, R and Scherer, K. R., Acoustic Profiles in Vocal Emotion Expression, Journal of Personality and Social Psychology, 70(3):614-636, 1996.

Cahn J.E., Generating Expressions in Synthesized Speech, Master's Thesis, MIT, 1989.

Cornelius R. R., Theoretical Approaches to Emotion, Proc. Of ISCA Workshop on Speech and Emotion, Belfast, September 2000.

Heuft B., Portele T. and Rauth M., Emotions in Time Domain synthesis, Proc. Of ICSLP, Philadelphia, USA, October 1996.

Hillenbrand J., "Perception of aperiodicities in synthetically generated voices", JASA, 83:2361-70, June 1988.

Kienast, M. and Paeschke, A. and Sendlmeier, W. F. Articulatory Reduction in Emotional Speech, Proc Eurospeech, Budapest, 1:117-120, 1999.

Klatt, D. H. and Klatt, L. C. Analysis, Synthesis and Perception of Voice Quality Variations among Female and Male Talkers, JASA, 87 (2):820-856, 1990.

Montero L.M., Gutiérrez-Arriola J., Palazuelos S., Enríquez E., Aguilera S., Pardo J.M., Emotional Speech Synthesis: From Speech Database to TTS, ICSLP 1998.

Murray, I. R. and Arnott, J. L. Implementation and testing of a system for producing emotion-by-rule in synthetic speech, Speech Communication 16 (1995) 369-390

Murray, I. R. and Arnott, J. L. Towards the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion, JASA, 93(2):1097-1108, 1993.

Picard R., Affecting Computing, The MIT Press, 1997.

Rank, E. and Pirker, H. Generating Emotional Speech with a Concatenative Synthesizer, Proc ICSLP, Sidney, 975-978, 1998.

Tatham M., Morton K., Expression in Speech: Analysis & Synthesis, Oxford Linguistics, 2004

Ververidis D., Kotropoulos C., A Review of Emotional Speech Databases, Proc. PCI, Thessalonica, Greece, 2003.

Vroomen J., Collier R., Mozziconacci S., "Duration and intonation in emotional speech", Institute for Perception Research, Eindhoven.

8. Acknowledgments

The presented work has been supported by GEMINI (IST-2001-32343) EC project.