GREEK EMOTIONAL DATABASE: CONSTRUCTION AND LINGUISTIC ANALYSIS

Panagiotis Zervas	UNIVERSITY OF PATRAS
Nikos Fakotakis	UNIVERSITY OF PATRAS
Irini Geourga	UNIVERSITY OF PATRAS
George Kokkinakis	UNIVERSITY OF PATRAS

Περίληψη

Η παραγωγή της ανθρώπινη ομιλίας δεν συνίσταται μόνο στην άρθρωση διαδοχικών φθόγγων. Παράμετροι όπως η ένταση, ο ρυθμός και ο τόνος της φωνής συμβάλλουν στη διαμόρφωση του επικοινωνιακού χαρακτήρα της, χωρίς να αλλοιώνουν τη φωνητική ταυτότητα των επιμέρους ηχητικών μονάδων. Ελέγχοντας δυναμικά αυτές τις παραμέτρους ο ομιλητής προσθέτει στο εκφώνημα πληροφορίες που δεν προκύπτουν από την επεξεργασία των συντακτικών και σημασιολογικών δομών. Στο άρθρο περιγράφεται η δημιουργία και η ανάλυση μιας συναισθηματικής βάσης για την ελληνική γλώσσα. Ο σχεδιασμός της βάσης αυτής έγινε με σκοπό τα αποτελέσματα της ανάλυσης να χρησιμοποιηθούν σε σύστημα σύνθεσης ομιλίας. Η ηχογραφημένη βάση ομιλίας αποτελείται από 69 προτάσεις για κάθε συναίσθημα. Οι προτάσεις αυτές είναι ουδέτερες, ερωτηματικές, επιφωνηματικές ώστε να είναι εφικτή η εξαγωγή πληροφορίας όσον αφορά την προσωδία και τον επιτονισμό. Το σώμα κειμένου εκφωνήθηκε πέντε φορές ώστε να εκφράζει θυμό, φόβο, χαρά, λύπη καθώς και μια ουδέτερη κατάσταση η οποία χρησιμοποιήθηκε σαν μέτρο σύγκρισης για την αναγνώριση των συναισθημάτων.

Key words

emotional speech, speech synthesis, prosody, fundamental frequency, pitch contour, declination phenomenon, duration, speech intensity.

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 expressiveness, or affect, provides 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 remainder 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)

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 (**M**ontero etal 1998) that recordings with actors are good approximations to true emotional speech.

To avoid the interference of a listener's decision on the emotional contents due to semantically meaning, we attempted to construct semantically neutral sentences. In this work we give the detailed description and the composition of an emotional speech database for Greek.

2. Database Construction

For the study and analysis of prosody, first we choose a number of sentences that will compose 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, newspapers or were set up by a professional linguist. Finally the corpus was compromised by 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. The thirt**y** 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. 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).

3. 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 and none of them was used to synthetic speech. The stimuli for the evaluation was five neutral-content sentences (twenty recordings per listener), randomly played.

The whole evaluation process took place in two parts. First a free response test was held where the listeners were labeling each utterance with whatever emotion found appropriate and second they were forced to choose between the four emotions that where included in our database. The results are tabulated on table 1.

Emotion	Free Response Test	Forced Response Test
Sadness	97,1%	97,5%
Anger	97,8%	98,2%
Joy	84%	89%
Fear	68%	74%

Table 1: Free and forced response test results.

4. Parameters for emotional speech description

In view of finding 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 and 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.
In the next pages a detailed description and statistical analysis regarding the results on measured variations is given.

4.1 Fundamental Frequency Parameter

As far as it concerns the addition of emotional characteristics in synthetic speech is essential the analysis, modeling and finally the generation of pitch contour. The fundamental frequency (F0) contour for each sentence in our corpus was extracted. First we started with the analysis of neutral session's F0 and then we proceeded to the analysis of each emotional counterpart. The F0 contour of each emotional session was compared with the neutral part. Quantitative definition of F0 contours for each emotional state is contacted by the utilization of declination phenomenon. The values of B_{start}, B_{end} and B_{slope} of neutral sessions were compared with their emotional versions. The above values are characteristics of an F0 contours baseline.

Emotion	**B_{start} Variation (Raise)**	**B_{end} Variation (Raise)**	**B_{slope} Variation**
Sadness	17,83%	18,56%	2,2%(raise)
Joy	54,70%	11,67%	20% (decrement)
Anger	33,43%	11,20%	11,1% (decrement)
Fear	20,12%	18,5%	2,3% (decrement)

Table 2: Emotional / Neutral speech fundamental frequency parameters variation.

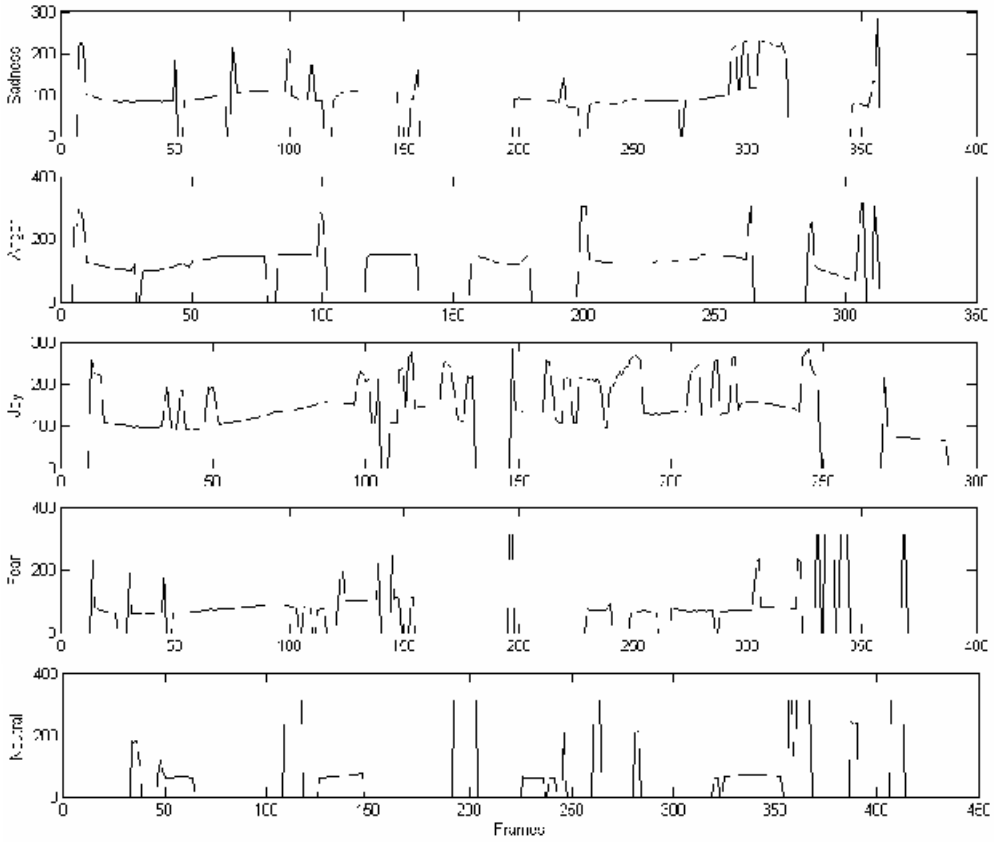
Comparison of the B_{start}, B_{end} and B_{slope} values showed that,

- B_{start} rises for all emotional states in regard of its neutral equivalent.
- B_{end} also seems to rise in emotional version of the utterances.
- As regards B_{slope} there was not a clear tension regarding each of the emotional state.

4.1.1 Comparing F0 Contours

Inspection of F0 contours of neutral utterances and their emotional versions led us to the conclusion that,

- Emotional version of each utterance had a contour similar to its neutral counterpart but shifted to higher frequencies.
- Pitch accent phenomena were still there but in a higher degree.
- Emotional versions (anger, joy mostly) seem to have a higher speech rate. In example pitch accent phenomena such as L*+H (Arvaniti and Baltazani 2000) were transformed, because of higher speech rate to H*.

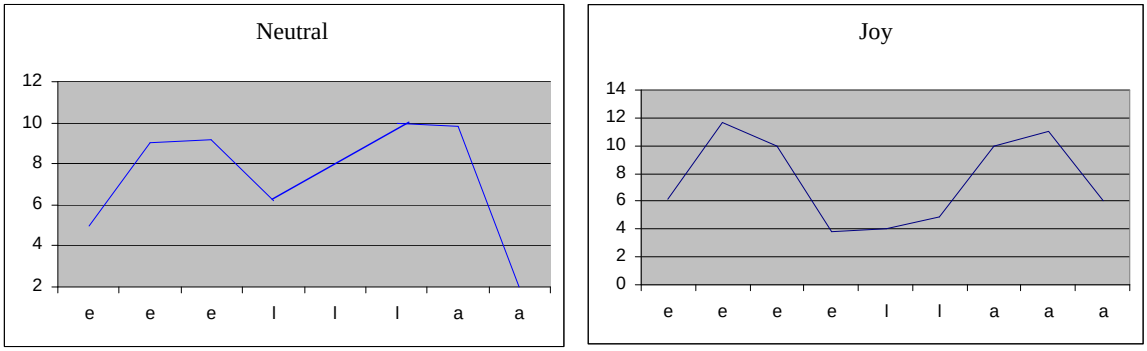


Picture 1: Emotional / Neutral speech pitch contour.

4.3 Speech Intensity Parameter

In order to verify if there are non random differences, as far as, it concerns the intensity of emotional speech, we calculated the energy per window (256 samples). We calculated the change of energy of each window against the mean value of the energy of the corresponding utterance. By inspection of the resulting graphs we came to the conclusion that the distribution of the intensity to the mean energy of the utterance is the same for the emotional and neutral speech.

For the interpretation of the intensity behaviour in each emotional state, we probe into phoneme energy. A category of phonemes (fricatives, explosives) showed an unbalanced behaviour (in some cases having almost zero energy and in other having exaggerated values). The main reason was that the behaviour of these phonemes was a function of the recording conditions.

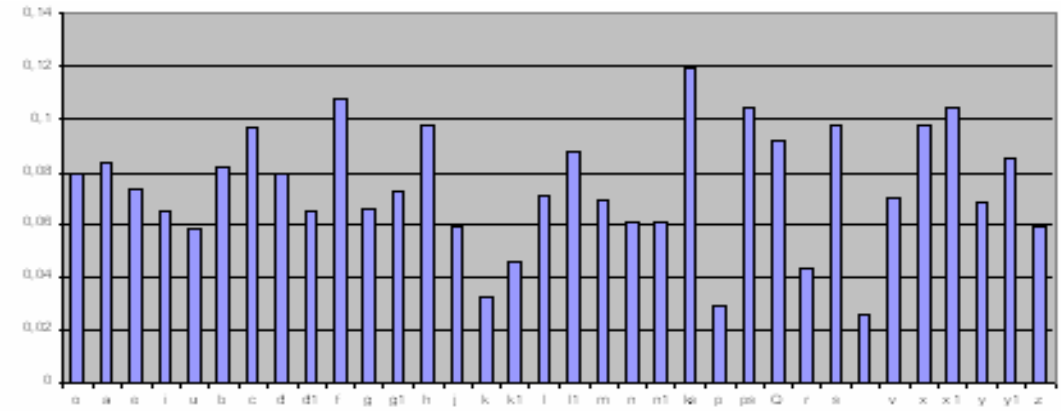


Picture 2: Neutral/Joy Intensity Distribution Examples

4.4 Speech Rate Parameter

Speech rate is known to be a variable affecting timing in a speech signal, but one that is difficult to quantify. Absolute measures of duration in text tell little about the relative lengths of segments, and account must be taken of all other factors involved if relative values as 'long', 'short', 'fast' or 'slow' are to be applied.

In picture 3 is depicted the mean duration of the phonemes of our database for the neutral session.

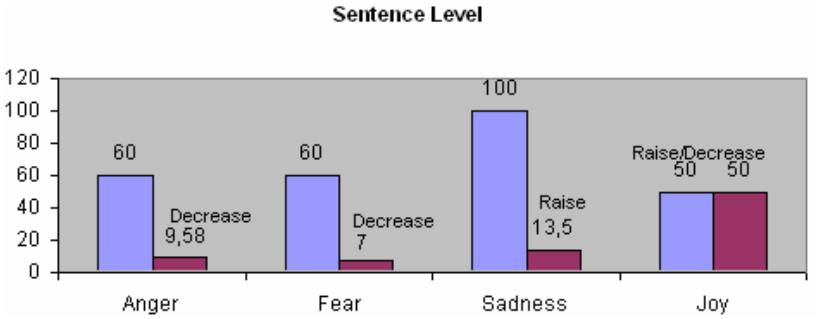


Picture 3: Neutral session phonemes mean duration.

For the measurement of the duration in sentence level we took the following results,

- Regarding anger we had a 60% decrease of sentence duration with a 9.58%.
- In fear we had a 90% decrease with a 7%
- For the sadness session there was a 100% raise of duration with a 13.5%
- And in joy there wasn't a clear tension for raise or decrease of duration.

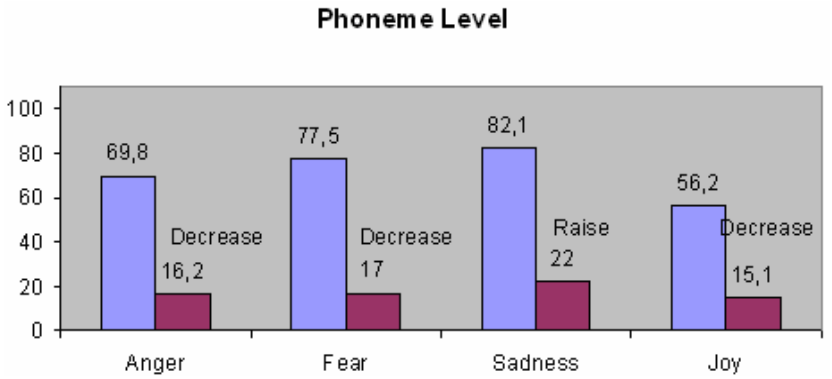
In the following picture the aforementioned observations are depicted.



Picture 4 Sentence level emotional sessions duration.

A further analysis of emotional speech duration was conveyed by measuring it in phonemic level. From this analysis of our data we took the following results,

- Regarding anger the 69,8% percent of the phonemes showed a decrease of duration by a 16,2% against the neutral counterpart.
- The 77,5% of phonemes in fear session showed a decrease of 17%.
- 82,1% had a raise in duration for the emotion of fear with a 22%.
- And in joy we had the 56,2% percentage of phonemes to show a decrease of duration in a percentage of 15,1% as regards the duration for its neutral equals.



Picture 5 Phoneme level emotional sessions duration

5. Conclusion

The recorded emotional speech database represents a good base for emotional speech analysis and is also usable for emotional speech synthesis. Some improvements we could apply consists of "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.

With a close inspection to the results of our research we can value our first hypothesis that emotional variation of speech can be achieved up to a level by slight manipulation of the three fundamental parameters we analyzed which are pitch, speech rate and speech intensity (Murray and Arnott, 1995).

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.

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., Gutierrez-Arriola J., Palazuelos S., Enriquez 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.

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

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