Employing Intonational Events Parameterization for Emotion Recognition

Panagiotis Zervas, Iosif Mporas, Nikolaos Fakotakis

Electrical and Computer Engineering Dept., University of Patras, Greece
{pzervas, imporas, fakotaki}@wcl.ee.upatras.gr

Abstract

In this work we introduce the utilization of Fujisaki's modeling of pitch contour for the task of emotion recognition. For the evaluation of the proposed features we have employed a decision tree as well as an instance based learning algorithm. The datasets utilized for training the classification models, were extracted from two emotional speech databases. Results showed that knowledge extracted from Fujisaki's modeling of intonation benefited all resulted emotion recognition models. Thus, an average raise of 9,52% in the total accuracy of all approaches was achieved.

1. Introduction

An extensive number of experiments to explore which particular aspects of speech would manifest saliently the emotional condition of a speaker, have been conducted lately. Results showed that those related to prosody [1], [2] (pitch contour, intensity, timing) are considered an important indicator to emotional states. Furthermore, voice quality [3] as well as certain co-articulatory phenomena [4] are also high correlated with some emotional conditions of a speaker.

In this study we cope with the task of recognizing emotional states, based on knowledge extracted only from speech signals. We present the evaluation of features carrying information regarding intonation and speaking style of a spoken utterance for the task of emotion classification.

For this purpose, prosodic knowledge was extracted from the Fujisaki's proposed model for F0 contour quantification. This model is based on the fundamental assumption that intonation curves, although continuous in time and frequency, originate in discrete events triggered by the speaker that appear as a continuum given physiological mechanisms related to fundamental frequency control. The model has been applied to several languages and good approximations of F0 contours have been presented. Fujisaki's representation of F0 is realized as the superposition of phrase effects with accent effects.

The outline of the paper is organized as follows. Initially we present a brief description of Fujisaki's model of intonation. In section 2 we present the methodology followed for the feature extraction from the utilized databases as well as the construction of the classification framework. Finally, we present and discuss the results of our evaluation framework.

2. Fujisaki's Intonational Model

Fujisaki's model, is the continuation of Ohman' s work [6] on the prosody of words. It is based on the fundamental assumption that intonation curves, although continuous in both time and frequency, originate in discrete events triggered by the reader and are the cause of physiological mechanisms related to F0 control. Fujisaki's model aims at modeling the generation process of F0 by giving an explanation to the physical and physiological properties behind it. The logarithm of the fundamental frequency contour is modeled superposing the output of two second order critically damped filters and a constant base frequency, figure 1.

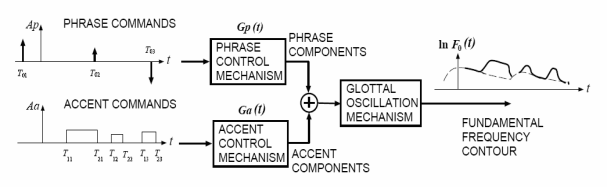

Fig. 1. Fujisaki's model for F0 contour generation.

One filter is excited with deltas (*phrase commands*), and the other with pulses (*accent commands*). With the technique of Analysis-by-Synthesis a given F0 contour is decomposed into its constituents (phrase and accent commands) and estimate the magnitude timing of their underlying commands by deconvolution. Equation 1 describes this relationship mathematically,

$$\ln F_0(t) = \ln F_b + \sum_{i=1}^{I} A_{p_i} G_{pi}(t-T_{0i}) + \sum_{j=1}^{J} A_{a_j}\{G_{a_j}(t-T_{1j}) - G_{aj}(t-T_{2j})\} \quad \textbf{(1)}$$

F_b is the baseline value of fundamental frequency, I the number of phrase commands, J the number of accent commands, A_{pi} the magnitude of the i_{th} phrase command, A_{aj} the amplitude of the j_{th} accent command, T_{0i} the timing of the i_{th} phrase command, T_{1j} the onset of the j_{th} accent command and T_{2j} the end of the j_{th} accent command. The output of the second order filters, described in equations 2 and 3, will provide the accent and phrase components of the pitch contour representation.

$$Gp_i(t) = \begin{cases} a^2 t e^{-at}, & t \geq 0 \\ 0, & t < 0 \end{cases} \quad \textbf{(2)}$$

$$Ga_j(t) = \begin{cases} \min[1-(1+\beta_j t)e^{-\beta t}, \gamma], & t \geq 0 \\ 0, & t < 0 \end{cases} \quad \textbf{(3)}$$

$G_{pi}(t)$ represents the impulse response function of the phrase control mechanism and $G_{aj}(t)$ the step response function of the accent control mechanism, α the natural angular frequency of the phrase control mechanism, β the natural angular frequency of the accent control mechanism and γ the relative ceiling level of accent components. Timing of commands, their amplitudes as well as phrase and accent components henceforth will be referred to as "*Fujisaki-parameters*". For the analysis of our data we have selected a value for *Fb* equal to the corpus-mean value yielded in the analysis. Concerning the time constants α and β they were chosen equal to 1.7 and 20 respectively. Furthermore, the parameter γ was set equal to 0.9.

3. Emotion Recognition Framework

The emotion recognition framework constructed for our evaluation consists of a feature extraction and a classification stage. Evaluation of the proposed prosodic features was conducted with the exploitation of two emotional speech databases which were holding recordings of four basic emotions and a neutral session.

3.1. Speech Material

The first corpus used for our evaluation was a Greek emotional speech (GrES) database [7] constructed in our lab. For the choice of the emotional states that were included in the database we tagged on the work of Oatley [8]. Therefore in our recordings we tried to capture the emotions of happiness, anger, sadness, fear and a neutral session. A professional actress familiar with radio theater was employed for the enunciation of the text corpus. 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 assisted to the normalization of 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 anger, joy, neutral, sadness, fear and neutral).

Conclusively, a Danish emotional speech (DES) [9] corpus which has been used previously for the task of emotion recognition was employed; in this database speech is expressed in 5 emotional states, such as anger, happiness, neutral, sadness, and surprise by four speakers, two male and two female. This particular database has been used in previous works in emotion recognition and it is easily accessible as well as annotated.

3.2. Features and Datasets

Selecting the appropriate acoustic feature vector is an important step in emotion recognition. For our exertion we initially calculated eighteen basic acoustic features and subsequently we extracted features from Fujisaki's model of intonation. Previous research has shown that emotional reactions are strongly related to the pitch and energy of the speech. For example, the pitch of speech associated with anger or happiness is always higher than that associated with sadness or fear, and the energy associated with anger is greater than that associated with fear.

As in similar research, this study adopts pitch and energy features. Thus, F0 values in the logF domain *(logF0)*, the difference of the pitch contour *(dlogF0)*, the first *(F1)* and second *(F2)* formant of the signal, the thirteen first Mel frequency cepstral coefficients *(MFCCs)* and the difference of energy *(dEnrg)* of each frame were calculated.

Pitch contour and formant frequency estimation was conducted with Praat [10] software. A 256 sample window and 128 sample frame shift was employed; pitch frequencies were assumed to be limited to the range of 60-320 Hz for both male and female data. The calculation of the thirteen MFCC parameters was carried out from [11]; a total of 40 filters and a 512 samples FFT size was applied for the calculation.

The second part of the feature set is consisted of four, novel to emotion recognition task, features derived from Fujisaki's Analysis-by-Synthesis parameterization of pitch. The features extracted from Fujisaki's model are the phrase component (PhrComp), the accent component (AccComp) per frame as well as the pitch resulted from Fujisaki's synthesis (FujLogF0) and the difference of resulted pitch contour (dFujLogF0). We have used the FujLogF0 as well as the dFujLogF0 in order to take advantage of the smoothed (no discontinuities) contour resulted from synthesis step.

With the application of Fujisaki's model we try to benefit from the fact that phrase commands are related to the slow varying component of intonation while the accent commands are related to fast changes. In that way, changes in prosody due to different emotional state could be exposed more straightforward. For the Fujisaki's parameters extraction we have utilized the freely available implementation of [12].
The datasets extracted from both databases contain only feature vectors corresponding to the voiced parts of the signal.

3.3. Classification Stage

The classification stage of our framework is consisted of two classifiers, the C4.5 tree inducer and the instance base learning approach. Both algorithms were acquired from the WEKA machine learning library [13] with a configuration resulted after a broad number of experiments with the GrES and DES datasets.

3.3.1. C4.5 Algorithm

In C4.5 [14], binary decision is carried out in the nodes of a decision tree producing a set of logical rules. Therefore, every path starting from the root of a decision tree and leading to a leaf represents a rule. The number of rules embodied to a given tree is equal to the number of its leaf nodes. The premise of each rule is the conjunction of the decisions leading from the root node, through the tree, to that leaf, and the conclusion of that rule is just the category that the leaf node belongs to.

For the growth of C4.5, the basic algorithm used, was a greedy method, constructing the tree in top-down recursive divide and conquer manner. In C4.5 tree algorithm the procedure of pruning is performed. Pruning is a process that is not included in some of its antecedent, such as the ID3 tree [14]. Unlike the stop splitting strategy, pruning is performed when a tree is grown fully and all the leaf nodes have minimum impurity. C4.5 selects a working set of examples at random from the training data and the tree growing/pruning process is repeated several times to ensure that the most promising tree has been selected. In our implementation of the algorithm only pre-prunning was applied and a confidence value of 25% was selected.

3.3.2. Instance Based Learning Algorithm

The Instance-Based (IBk) learning algorithm [15], represents the learned knowledge simply as a collection of training cases or instances. It is a form of supervised learning from instances; it keeps a full memory of training occurrences and classifying new cases using the most similar training instances. A new case is then classified by finding the instance with the highest similarity and using its class as prediction. IBk algorithm is characterized by a very low training effort. This leads to high storage demands caused by the need of keeping all training cases in memory. Furthermore, one has to compare new cases

with all existing instances, which results in a high computation cost for classification.

Each new instance is compared with existing ones using a similarity function, and the closest existing instance is used to assign the class to the new one. The similarity function used in IBk for k instances is given by equation 4,

$$Similarity(x, y) = -\sqrt{\sum_{i=1}^{n} f(x_i, y_i)} \quad (4)$$

where the instances are described by n attributes. As regards numeric valued attributes the f function equals to,

$$f(x_i, y_i) = (x_i - y_i)^2 \quad (5)$$

IBk is identical to the nearest neighbor algorithm except that it normalizes its attributes' ranges, processes instances incrementally, and has a simple policy for tolerating missing values. It saves only misclassified instances and employs a "wait and see" evidence-gathering method to determine which of the saved instances are expected to perform well during classification.

In our implementation of instance based learning, we adopted a number of 5 neighbors (IB5) as it provided the best classification results.

4. Evaluation

In this section we examine the resulted behavior of Fujisaki's features for the task of emotion prediction. To increase the evaluation's validity we applied two well established machine learning approaches on two emotional databases. As a result, four frameworks were built. Evaluation of the derived emotion recognition models was conducted with the application of the 10-fold cross validation [16] technique. Performance was estimated by calculating the F-measure metric per emotion category which is defined as the harmonic mean of precision and recall and is calculated as shown in equation 6:

$$F = 1 \Big/ \left(\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R} \right) \quad (6)$$

where α is a factor which determines the weighting of precision and recall. A value of α=0,5 is often chosen for equal weighting of precision and recall.

4.1. Results

In table 1 the accuracy of the derived emotion recognition models is presented. It shows that total accuracy increases with the addition of Fujisaki's features with an average of 9,52%. Due to the nature of the spoken data, models derived from GrES datasets presented higher accuracy for all scenarios. This is due to the fact that GrES database, is considered as text and speaker dependent (same text corpus across all emotions uttered by one speaker).

Table 1. *Emotion Recognition Models Total Accuracy*

Database	ML Method	Features Set	Total Accuracy (%)
GrES	C4.5	Basic	73,85
		Basic/Fujisaki	82,87
	IB5	Basic	86,16
		Basic/Fujisaki	90.47
DES	C4.5	Basic	50
		Basic/Fujisaki	66,01
	IB5	Basic	64,16
		Basic/Fujisaki	72,93

In Figure 2 the F-measure for GrES models for C4.5 and IB5 is illustrated; it shows that the Fujisaki's set of features improved the recognition of all emotion categories, especially the ones with the lower prior F-Measure score such as happiness, sadness and surprise. We can argue that this is a result of the fact that such emotions are manifested with pitch contours obtained from intonational phenomena less sharp, thus more accurately represented from Fujisaki's model.

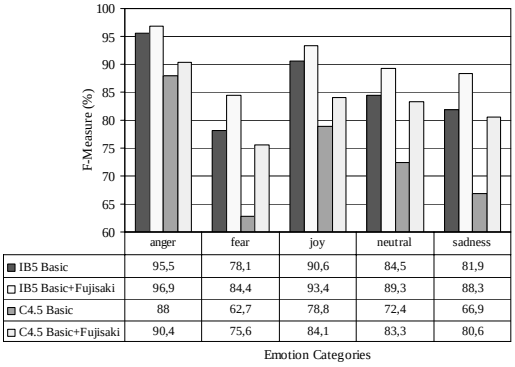

Figure 2: *F-Measure for IB5 and C4.5 models trained with GrES datasets*

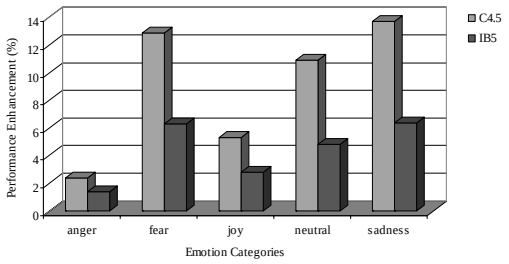

Figure 3: *Enhancement of F-measure for GrES datasets with the Fujisaki's set addition*

Figure 3 illustrates the difference in F-measure between the models trained with basic and basic+Fujisaki's set of features. High correlated emotion categories such as anger/joy and fear/sadness that share alike acoustical properties, were more accurately predicted with the addition of the Fujisaki's feature set. Higher recognition accuracy was achieved for the neutral category when Fujisaki's set of features was added; it resulted reduction in the confusion between emotion categories of fear and sadness.

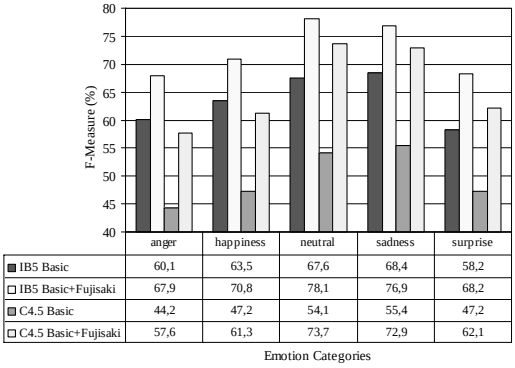

Figure. 4: *F-Measure for IB5 and C4.5 models trained with DES datasets*

Recognition of sadness with the basic set of features, showed confusion with the emotions of fear, joy and neutral; however, adding Fujisaki's features resulted segregation of the misclassified categories. Figure 4 depicts the emotion recognition results obtained from C4.5 and IB5 models trained with the full DES database; 5 emotions, 4 speakers, as a consequence the resulted datasets are considered text , speaker and gender independent [9]. Additionally, figure 5 depicts the prediction improvement achieved in each emotion category with the inclusion of Fujisaki's set of features to both algorithms.

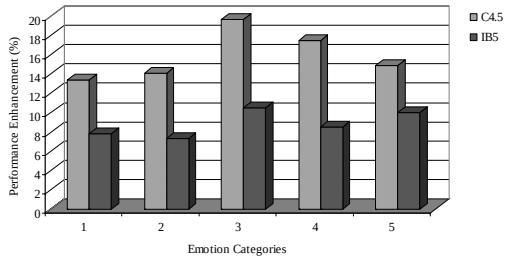

Figure 5:. *Enhancement of F-measure for DES datasets with the Fujisaki's set addition*

As in GrES experiments, the classification of certain emotion categories with acoustical similarities, such as neutral/sadness or surprise/happiness, presented higher results. It worth's mentioning that in the case of DES database, neutral category had the highest prediction F-measure for both classification algorithms with an equal performance to sadness for C4.5. Since neutrality is considered as some kind of carrier which gets modulated to reveal other expressions [17], it shares similar acoustical properties with all speakers such as flat pitch and energy contour etc. Tables 2 and 3 tabulate the classification results of human evaluation for GrES and DES [9] databases respectively.

Table 2. *GrES Human Classification Rates*

Stimuli	Response (%)				
	Anger	**Joy**	**Sadness**	**Fear**	**Neutral**
Anger	98.0	0.0	0.0	2.0	98.0
Joy	0.0	98.4	0.0	1.6	0.0
Sadness	0.0	1.7	72.9	25.4	0.0
Fear	0.0	0.0	11.3	64.5	0.0
Neutral	0.0	0.0	6.8	0.0	0.0

Table 3. *DES Human Classification Rates*

Stimuli	Response (%)				
	Neutral	**Surprise**	**Happiness**	**Sadness**	**Anger**
Neutral	60.8	2.6	0.1	31.7	4.8
Surprise	10	59.1	28.7	1.0	1.3
Happiness	8.3	29.8	56.4	1.7	3.8
Sadness	12.6	1.8	0.1	85.2	0.3
Anger	10.2	8.5	4.5	1.7	75.1

5. Conclusion

In this work, we evaluated the employment of features extracted from Fujisaki's intonational model for the task of emotion recognition from speech. These features were extracted from two emotional speech databases describing five basic emotional states each. The contribution of the introduced features was evaluated compared to a basic set of attributes previously employed for similar tasks. Extracted datasets were used for training C4.5 and IB5 classification algorithms. Their evaluation revealed the effectiveness of Fujisaki's features to the task of basic emotion classification.

6. References

[1] Murray I.R., Arnott J.L., Towards a simulation of emotion in synthetic speech: a review of the literature on human vocal emotion, JASA 93(2), pp. 1097-1108, 1993

[2] Cowie R. and Douglas-Cowie E., "Automatic statistical analysis of the signal and prosodic signs of emotion in speech," in Proc. of ICSLP, Philadelphia, pp. 1989–1992, 1998.

[3] Gobl C., Chasaide A.N. Testing Affective Correlates of Voice Quality through Analysis and Resynthesis, in Proc. Of the ISCA Workshop on Emotion and Speech, 2000.

[4] Kienast M., Sendlmeier W. Acoustical Analysis of Spectral and Temporal Changes in Emotional Speech, in Oric id the ISCA Workshop on Emotion and Speech, 2000.

[5] Fujisaki, H. and Hirose, K. "Analysis of voice fundamental frequency contours for declarative sentences of Japanese". In Journal of the Acoustical Society of Japan (E), 5(4): pp. 233-241, 1984.

[6] Ohman, S., "Word and sentence intonation, a quantitative model," Tech. Rep., Department of Speech Communication, Royal Institute of Technology (KTH), 1967.

[7] Zervas, P., Geourga, I., Fakotakis, N., Kokkinakis, G., Greek Emotional Database: Construction and Linguistic Analysis, 6th International Conference of Greek Linguistics, Rethymno, 2003.

[8] Oatley, K. & Gholamain, M., Emotions and identification: Connections between readers and fiction. In M. Hjort & S. Laver (Eds.) Emotion and the arts. (pp. 263-281). New York: Oxford University Press, 1997.

[9] Engberg, I., S., Hansen, A., V., Documentation of the Danish Emotional Speech Database (DES), Internal AAU report, Center for Person Kommunication, Denmark, 1996.

[10] Boersma, P., & Weenink, D. (2005). Praat: doing phonetics by computer (Version 4.3.01) [Computer program]. Retrieved from http://www.praat.org/

[11] Slaney M. (1998). Auditory Toolbox. Version 2. Technical Report #1998-010, Interval Research Corporation.

[12] Mixdorff, H., "A novel approach to the fully automatic extraction of Fujisaki model parameters," in Proceedings of ICASSP, 2000, vol. 3, pp. 1281–1284, Istanbul, Turkey.

[13] Witten, I., H., Frank, E., Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[14] Quinlan, R., C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., 1993.

[15] Aha, D., Kibler, D., Albert, M., Instance based learning algorithms. Machine Learning, 6:37 -- 66, 1991.

[16] Stone, M., Cross-validation choice and assessment of statistical predictions. Journal of the Royal Statistical Society, 36, 111-147, 1974.

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