

Speech Emotion Recognition: Comparison of Speech Segmentation Approaches

Muharram Mansoorizadeh

Electrical and Computer Engineering Department
Tarbiat Modarres University
Tehran, Iran
mansoorm@modares.ac.ir

Nasrollah.M. Charkari

Electrical and Computer Engineering Department
Tarbiat Modarres University
Tehran, Iran
charkari@modares.ac.ir

Abstract— Recognition of emotional states carried in speech, is of a great interest in modern human computer interaction developments. To reliably detect the aroused emotion, a sufficiently long continuous speech segment is required. This research aims to analyze different segmentation approaches of speech signals. Berlin emotional speech database is used for data set generation. Time frame based and voiced segmentation approaches are applied and compared. The experimental results show that accurate emotion recognition is obtained when the speech segments are longer than a second or are composed of 10 to 12 voiced segments. Based on the findings of this research, voiced based segmentation generates more accurate results than the other methods.

I. INTRODUCTION

The goal of active Human-Computer Interaction (HCI) research is to develop interaction methods similar to human-human communication methods. Because, Humans are already skilled in these methods and no much new knowledge is required to use computer applications. Speech, Body and Gesture Languages are familiar methods for inter-human communications. Besides textual content, speech may carry emotional states which have a great role in perception and interpretation in the other side. A fixed text may have quite different meanings; if it is spoken with different emotional articulations. To get actual meaning of a speech segment; it is required to recognize its affective content.

Many works has been done for speech emotion recognition but most of them use speech segments ranging from a word to a couple of sentences. Here a question arises: “How long a speech segment should be to know its emotional content correctly?” Physiological and psychological studies show that expressing an emotion in speech (and also in much other contexts) has a beginning, a raising side a peak and a falling side[1]. The speech segment should be long enough to capture the most possible information. Very small segments may lack informative peak area. In the other side, subsequently expressed emotions may affect each other; if the selected segment is too long.

It is desirable to segment the speech signal at least in two different ways: frame-based and voiced segmentation. In frame based approach the speech signal is fragmented into

segments of constant lengths (say 10 ms) and a couple of consecutive frames make an independent speech sample. Voiced speech segmentation, initially extracts voiced parts; and then fragments the speech into parts containing a predefined number of voiced segments.

The main contribution of this paper is to compare these approaches with different frame lengths or voiced segments. Results show that segmentation based on voiced segments performs better than the frame based approach. Also, in the case of frame based segmentation, segments longer than a second are required. For voiced segmentation, each block must have 10 -12 voiced segments.

The rest of the paper is organized as follows. In the next section we briefly review related works. In section 3, current study on segmentation approaches is explained. Section 4 contains experimental results of the approaches and their comparison and section 5 concludes the paper.

II. PREVIOUS WORK

Recently emotion recognition from speech has been extensively studied[2]. The main body of the existing research addresses selecting features or generalization issues. For summary, most speech emotion recognition systems use statistical features from fundamental frequency (Pitch) of speech, signal energy of special bands, formants and some global features like speech rate. An exhaustive search by Oudeyer[3] showed that a few statistical moments of pitch, energy and formant contours contain most of the discriminating information and other features like moments of their first and second derivatives are highly correlated with the features of the original contours .

In most studies data sets are composed of utterances ranging from less than a second to a couple seconds. As listed by Cowie, et. al. [4] the existing databases use sentences with different lengths to embody an emotion. As an example, the German emotional speech database is composed of utterances varying from 1 second to 8 seconds[5]. Obviously, global features like mean of the pitch or energy contours are dependant to signal duration. Furthermore,

speech rate ,which is known to be useful for recognition of highly active emotions(e.g. Anger) ,is calculated as density of pitch or energy peaks over time[6, 7]. Therefore, a fragment of speech should be long enough to make computation of these features reliable. The importance of time sensitiveness of some features is mentioned by cowie ET. Al. however none of the surveyed databases supports it[4].

A few researches have used segments of an utterance for emotion recognition. Shami [8] extracts voiced segments of speech and classifies the emotion based on each segment independently. Then, the results are combined for the whole utterance at decision level. Datcu [9] tried to divide the signal into N equal length frames and selected different combinations of these frames as samples. For example, after fragmenting the signal into 10 frames, he took samples composed of 1st, 3rd and 4th frames or 5th, 6th, 8th and 10th frames and so on. Datcu used mean and standard deviation of pitch and energy contours along with first four formants' energy and bandwidths. Since frames are taken in any combination, features like plateau of the pitch contour or speech rate cannot be estimated from these samples. The segmentation approach should preserve initial order and subsequence of the frames to be useful for extracting required features.

III. CURRENT STUDY

The goal of the paper is to demonstrate the importance of segmenting speech signals for emotion recognition and compare classification performances of different approaches. We briefly describe two segmentation approaches and apply them to the Berlin[5] emotional speech database. Then we compare recognition results of the approaches.

Human speech is produced through the air pressure transmission in the vocal tract. Resonance of vocal cords filters out the white noise signal originated from lungs and makes a periodic signal. The resulting sound is further filtered by the other parts of vocal tract acting as pipes with different lengths and shapes. The periodic speech signal created this way is called voiced speech. The resonance pattern determines the fundamental frequency of speech. On the other hand, unvoiced speech are white noise like sounds generated by lips or teeth[10]. Unvoiced parts act as delimiters for voiced parts and enable humans to form phonemes and words. The fundamental frequency of voiced speech is called Pitch. If, pitch is determined with low error for a segment of speech; then the segment is voiced, otherwise it is noise like and unvoiced. It is well known that voiced part of speech carries emotional content[11]. Figure 1 shows a speech signal with its energy and pitch contours. Pitch contour is a discrete set of sequences which are voiced segments.

A. frame-based segmentation

In this approach speech signal is divided into frames of short length (say 10 ms); then the signal is fragmented into consecutive blocks of n frames each. Small number of frames preserves features stability but may lack of emotional information. Also, acoustic features of a block with large number of frames maybe too variant. To select the proper number of frames, here we exhaustively test a broad range of selections; i.e. from 10 to 250 frames per block.

B. Voiced Segments

Pitch frequency is determined by analyzing periodicity of the signal in short time durations. Several approaches have been proposed for pith determination. Recent methods (e.g. one which Praat uses) are more or less based on the super resolution pitch determination by Medan et. al. [12].

Since voiced segments contain information about emotions; we are motivated to select blocks of speech containing a few voiced segments. After extracting voiced segments; samples of speech with at least 1,2,...,18 voiced segments are generated. Note that blocks of an utterance are no more equal in length. Because duration of the individual voiced segments are not the same.

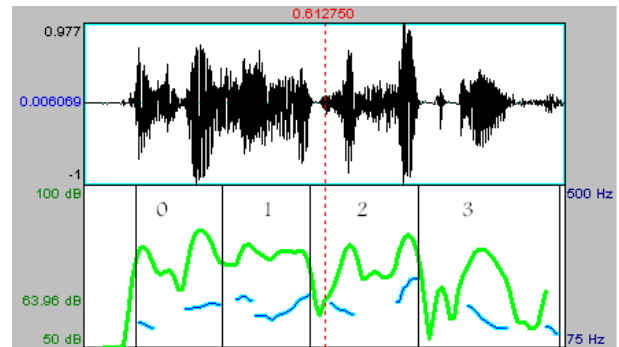


Figure 1 Speech Signal(up),its energy contour(down, continues) and Pitch contour of voiced Segments(Down, Discrete). Blocks containing two voiced segments are denoted by the vertical lines of the lower part. There are 4 blocks in this picture.

IV. RESULTS

These two approaches applied to the Berlin database and compared the results with each other also with results from Datcu et. Al.[9]. This database contains spoken sentences of 10 actors (5 male and 5 female) in 7 emotional contexts: Anger, Boredom, Disgust, Fear, Joy, Neutral and Sadness. The actors produced 10 predefined sentences with each emotion category. A set of 535 samples of this DB is available over their web site[5].

Using Praat [13], we have extracted total energy of the signal, energy of lower 250 Hz, 10 Hz smoothed pitch contour, first four formants and their related bandwidths. Then, as discussed below, the extracted sequences are segmented using both above approaches and data sets are

generated. Matlab® 7.4 is used for feature extraction and classification purposes.

Statistical moments (quartiles, standard deviation...) of the sequences are extracted as features. Plus, mean duration of the voiced segments is calculated as speech rate. The well known *sequential forward feature selection algorithm*[14] is used for selecting most informative features. The emotion classes are modeled by a multivariate Gaussian distribution of the selected features. Finally, the results are generated using 10 folds cross validation scheme [14] and a simple linear classifier. Accuracy of a class is the ratio of correctly labeled samples to the total samples of the class. In terms of confusion matrix, it is the well-known true-positive-rate quantity. Results of the both approaches are summarized in TABLE I. subsequent figures present the details of each experiment. To get the global trend of the results, for each curve, its moving average of 5-samples period is also over plotted.

A. Frame-based segmentation

In this approach, several tests have been done with 10 to 250 frames per sample. For the sake of clarity, we present the classification results of the emotion classes in three graphs (figures 2-3). As it is shown in Figure 2; “Sadness” has the highest classification accuracy. This is similar to findings multiple studies[15] and verifies the general statement that “lowly active emotions(e.g. sadness) are better recognized from speech signals”. It is clear from that Happiness requires segments longer than 150 frames (1.5 seconds) to be classified reliably. Anger’s recognition performance is also improves slightly with segments longer than 100 frames (1 second); but Sadness could be classified accurately (70% +) with even a short segment (50 frames or 0.5 seconds). Several studies[15] have shown that sadness and other lowly active emotions are categorized by stability in features(e.g. pitch contour) and lower energy. For short utterances, these features do not undergo subtle changes, and the related emotions are classified accurately. Figure 3 Denotes that classification accuracy of “Fear” could be slightly more than 50% if the samples are at least 1.5 seconds. Also recognition of ‘Neutral’ improves more with longer segments. Figure 4 shows that the classification accuracy of ‘Boredom’ and ‘Disgust’ are similar and bounce around 50%. There is a peak for ‘Disgust’ at 130 frames but it is not stable in subsequent tests. It may be a result of diversity of sample in k-folds cross validation[14]. By selecting segments around 1.5 seconds, we expect to achieve classification results around 50 % for ‘Boredom’ and ‘Disgust’.

To conclude this section, we can say that for accurate recognition of emotions, speech segments of length 1.5 seconds or more required. Furthermore, ‘Happiness’ needs longer segments than the others.

B. Results for Samples generated by collecting voiced segments

Sample generated by selecting voiced segments are classified by the same settings as the previous test (e.g. 10 folds cross validation and Gaussian modeling). Results are depicted in figures 5 and 6. Again, trend lines are plotted to get a smooth idea of the behavior of the curves. As Figure 6shows, over all performance of the recognition is improved compared to the results from frame based segmentation. The interesting point in this approach is that the accuracies of the emotions are at least 50%. Also, for accurate classification, this figure suggests using 10 or more voiced segments.

Similar to the previous approach, ‘Sadness’ has the highest accuracy. ‘Disgust’ is the emotion with the highest difference of accuracy between these approaches. So, we can conclude that it is very dependent in the whole body of the voiced segments.

In summary, samples generated by selecting a number of voiced segments are more informative than frame based segmentation and generate more accurate results as listed in TABLE I.

TABLE I. HIGHEST ACCURACIES RESULTED FROM EACH APPROACH. THE RESULTS ARE ESTIMATED FROM TREND LINES(BOLD LINES IN THE FIGURES) , INSTEAD OF THE ORIGINAL CURVES.

Emotion	Frame Based		Voiced Segments	
	#Frames	Acc*(%)	# Segments	Acc(%)
Anger	180	0.65	10	0.70
Boredom	140	0.62	15	0.72
Disgust	140	0.58	15	0.78
Fear	150	0.52	10	0.63
Happiness	200	0.55	14	0.55
Sadness	100	0.75	11	0.82
Neutral	220	0.72	18	0.77

* Acc: Accuracy

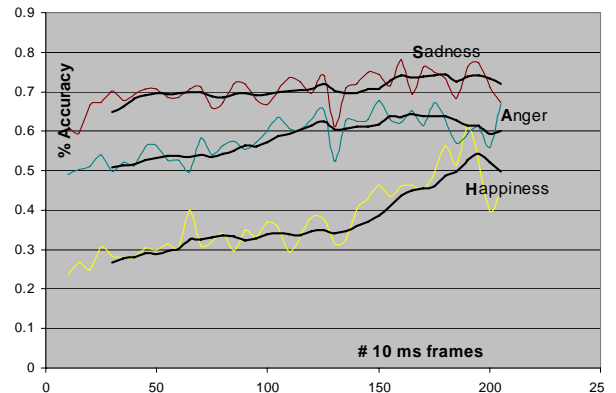


Figure 2 Accuracy of Sadness, Anger and Happiness using different number of frames per sample. Over each curve, the moving-average trend linewith period of 5 samples is also plotted (Bold Line).

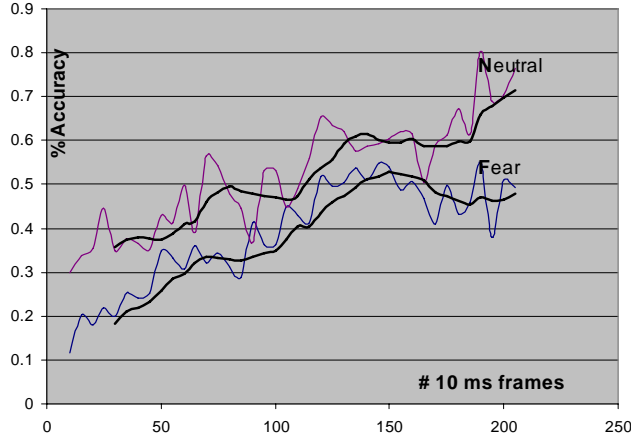


Figure 3 Accuracy of Neutral and Fear classes and their trendlines.

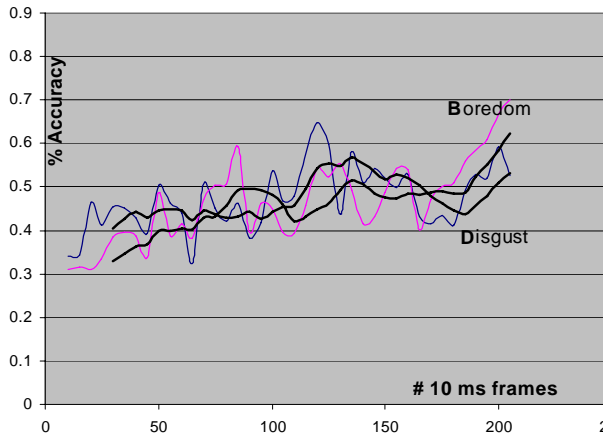


Figure 4 Accuracy of Neutral and Fear classes and their trendlines.

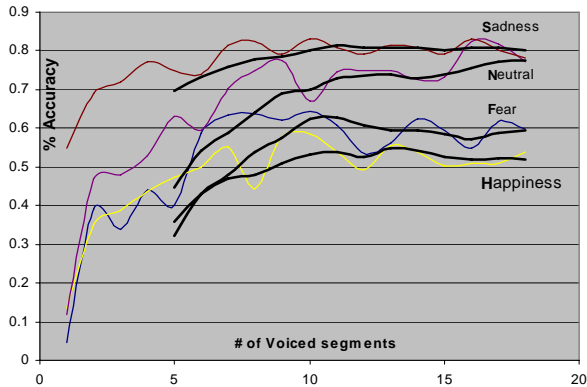


Figure 5 Classification Accuracy of Sadness, Neutral, Fear and Happiness using different number of voiced segments per sample. Over each curve, the moving-average trendline with period of 5 samples is also plotted

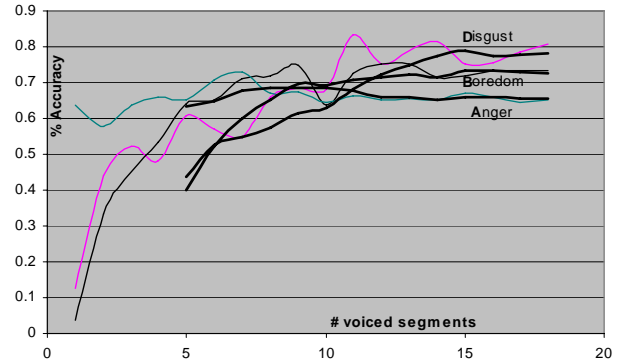


Figure 6 Classification Accuracy of Disgust, Boredom and Anger using different number of voiced segments per sample.

C. Comparison with Datcu's results

Results from Datcu [9] is summarized in TABLE II. the 'tpr' column (true positive rate) is the same as 'accuracy' which we have used in presenting the results. Compared to our experiments (see TABLE I.); results of sadness and happiness are similar. Datcu's accuracy for anger is higher than the voiced segmentation. Fear, disgust and boredom are better classified by voiced segmentation than the Datcu's method.

TABLE II. DATCU'S RESULTS[9].

emotion	tpr (%)	fpr (%)
anger	0.72±0.16	0.13±0.06
boredom	0.49±0.18	0.09±0.09
disgust	0.24±0.43	0.00±0.00
anxiety/fear	0.38±0.15	0.05±0.04
happiness	0.54±0.41	0.14±0.13
sadness	0.83±0.06	0.08±0.06

V. CONCLUSION AND FUTURE DIRECTIONS

Frame based and voiced segmentation of speech signals for classifying emotions are studied in this paper. Several tests have been done to assess the approaches. Results show that for frame based approach, segments longer than 1.5 seconds is required and for voiced segmentation, 10 or more voiced segments should be considered. Also comparison of the approaches show that voiced segmentation performs better than the frame based approach and Datcu's proposed segmentation method.

In future, we plan to test the approaches on more databases with more diverse samples. Also we will try to use more powerful classifiers instead of linear one used in this paper.

The Berlin database is available in public through <http://pascal.kgw.tu-berlin.de/emodb/>. Praat and Matlab scripts, together with the results which this paper is built upon, are available through the author's personal website at: <http://muharram.googlepages.com/index.htm>

REFERENCES

- [1] P. Ekman, *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*: Times Books, 2003.
- [2] O. Pierre-Yves, "The production and recognition of emotions in speech: features and algorithms," *Int. J. Human-Computer Studies* vol. 59, pp. 157-183, 2003.
- [3] O. Pierre-Yves, "The production and recognition of emotions in speech: features and algorithms," *Int. J. Human-Computer Studies*, vol. 59, 2003.
- [4] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech Communication*, vol. 40, pp. 33-60, 2003.
- [5] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," in *Interspeech* Lissabon, Portugal, 2005.
- [6] R. Cowie and E. Douglas-Cowie, "Automatic statistical analysis of the signal and prosodic signs of emotion in speech," 1996, pp. 1989-1992 vol.3.
- [7] B. Xie, L. Chen, G.-C. Chen, and C. Chen, "Statistical Feature Selection for Mandarin Speech Emotion Recognition," in *ICIC*, 2005, pp. 591 - 600.
- [8] M. T. Shami and M. S. Kamel, "Segment-based approach to the recognition of emotions in speech," 2005, p. 4 pp.
- [9] D. Datcu and L. J. M. Rothkrantz, "The recognition of emotions from speech using GentleBoost classifier. A comparison approach," in *International Conference on Computer Systems and Technologies*, 2006.
- [10] B. Gold and N. Morgan, *Speech and audio signal processing*: Wiley New York, 2000.
- [11] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, pp. 32-80, 2001.
- [12] Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech signals," *IEEE Transactions on Signal Processing*, vol. 39, pp. 40-48, 1991.
- [13] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (Version 4.6.12) [Computer program]." Retrieved Aug3, 2007, from <http://www.praat.org/>, 2006.
- [14] S. Theodoridis and K. Koutroumbas, *Pattern Recognition, 2nd Edition*: ELSEVIER ACADEMIC PRESS, 2003.
- [15] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, pp. 1162–1181, 2006.