# SEGMENT-BASED APPROACH TO THE RECOGNITION OF EMOTIONS IN SPEECH

*Mohammad T. Shami and Mohamed S. Kamel*

Pattern Analysis and Machine Intelligence Lab
Electrical and Computer Engineering, University of Waterloo
Waterloo, Ontario, Canada, N2L 3G1
{mshami,mkamel}@pami.uwaterloo.ca

## ABSTRACT

A new framework for the context and speaker independent recognition of emotions from voice, based on a richer and more natural representation of the speech signal, is proposed. The utterance is viewed as consisting of a series of voiced segments and not as a single object. The voiced segments are first identified and then described using statistical measures of spectral shape, intensity, and pitch contours, calculated at both the segment and the utterance level. Utterance classification is performed by combining the segment classification decisions using a fixed combination scheme. The performance of two learning algorithms, Support Vector Machines and K Nearest Neighbors, is compared. The proposed approach yields an overall classification accuracy of 87% for 5 emotions, outperforming previous results on a similar database.

## 1. INTRODUCTION

Adding emotional intelligence to computers is an interesting yet difficult challenge that promises to bring about a positive revolution in the existing relationship people have with automated systems. Research has started to focus lately on the modelling and interpretation of human emotions from a computational perspective. In [1], the author coins the term "*Affective Computing*" to describe a newly established research field that deals with the automatic sensing, recognition and possible synthesis of human emotions from any biological modality such as voice or facial expressions.

Most previous studies aiming at the automatic recognition of emotions in speech have made use of utterance level statistics of the pitch and intensity contours while ignoring the dynamic aspects of those contours as in [2]. Some studies make indirect or limited use of the segment level information such as [3] and [4]. Although utterance level approaches have resulted in considerable recognition success, there is mounting evidence that suggests that information, that more closely encodes the shape of the intensity and the pitch contours, is relevant and contributes information that is independent of utterance level information as in [5] and [6]. Features extracted at a lower temporal scale, such as the voiced segment scale, serve to retrieve this valuable but usually neglected information. In this study, we strive to increase the temporal complexity of the extracted features by incorporating features extracted at two different temporal scales of the speech signal, the utterance and the segment level scale.

In section 2 an overview of the proposed approach is provided. In subsections 2.2, 2.3, and 2.4 specific aspects of the approach are discussed. In section 3, experimental results using the KISMIT speech database from [3] are presented and compared to earlier results. Final conclusions are made in section 4.

## 2. PROPOSED APPROACH

### 2.1. Overview

As the flowchart in Fig. 1. shows, the speech sample as a whole is first summarized using statistical measures of spectral shape, intensity, and pitch contours. As a by-product of the pitch extraction process, the utterance is segmented into a sequence of $N$ voiced segments. Using that segmentation information, the same battery of statistical measures that was calculated at the whole utterance level is recalculated for each of the detected segments. Now a feature vector consisting of both utterance level information and information local to the voiced segment is formed for each of the voiced segments. At classification time, and since class labels are provided for utterances as a whole, it is assumed that each of the voiced segments contains an expression of the affective intent of the utterance it belongs to, and therefore it is given that same label. A segment classifier is trained using the segment feature vectors and the assumed segment labels. For the classification of whole

*0-7803-9332-5/05/$20.00 ©2005 IEEE*

utterances, the decisions made by the segment classifier for each of its voiced segments, expressed as a posteriori class probabilities, are aggregated to obtain a single utterance level classification decision.

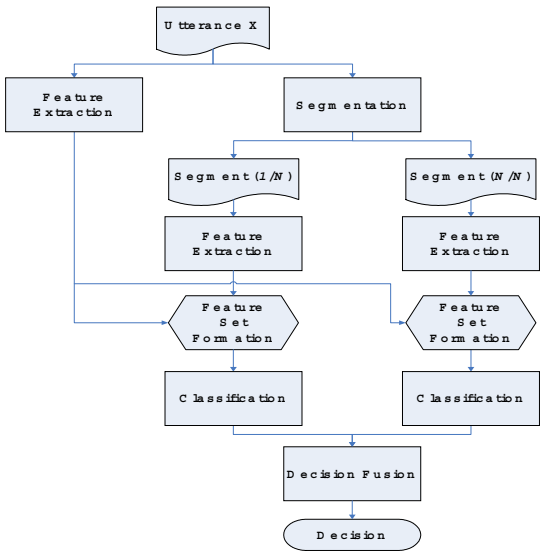


Fig. 1. The Flowchart of the Proposed Approach

### 2.2. Segment Based Representation of Speech

The segmentation of the speech sample will be based on its voicing content determined by the pitch extraction algorithm.

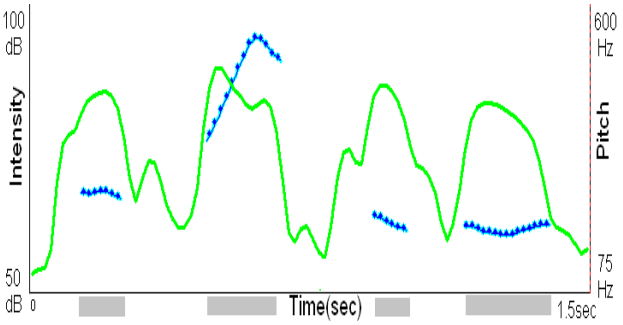


Fig. 2. The intensity contour (solid line) and pitch contour (dotted line) of a sample utterance. Note that this sample will be segmented into four segments

Since the speech samples in the database used contain utterances of different lengths and phonetic content, the segmentation process will result in a variable number $N$ of voiced segments and consequently, $N$ feature vectors, each of fixed length $M$=24. An example is in Fig. 2.

### 2.3. Features

The feature set consists of a battery of 12 statistical measures of spectral shape, intensity and pitch contours. These measures are derived from those commonly used in the literature, specifically, those in [3] and [4]. The feature vector describing a segment is made up of these 12 measures calculated twice, once at the segment level and once at the utterance level. Statistics such as max, minimum, and standard deviation make up the battery of measures.

In order to extract those measures, first the time series data (contours) of pitch, intensity, and Mel Frequency Ceptral Components (MFCCs) are extracted from the raw speech signal using the speech processing software in [7]. Subsequently, the statistical measures are calculated from those contours for the individual voiced segments and for the utterance as a whole. Table 1. describes the breakdown of the battery of measures across the different types of speech parameters.

TABLE 1. DISTRIBUTION OF MEASURES OVER ACOUSTIC PARAMETERS

| Voice Parameter | Number of Measures |
|---|---|
| Pitch Contour | 6 |
| Intensity Contour | 3 |
| Spectral Shape | 2 |
| Duration | 1 |
| Total | 12 |

### 2.4. Classification Scheme

For the segment classifier, the performance of two supervised learning algorithms is explored, K Nearest Neighbours and Support Vector Machines with 2nd degree polynomial kernel.

The output of the segment classifier consists of a vector of a posteriori class probabilities. Each of those probabilities describes how likely it is that a segment falls into a class $C_n$ knowing its feature vector $F_{Seg_X}$ The vector of a posteriori probabilities is in (1).

$$P\left(C_n \mid F_{Seg_x}\right), \qquad 1 \le n \le T \tag{1}$$

Where $T$ is the total number of existing classes and $n$ is the class index.

Obtaining the aposteriori class probabilities for the segments does not enable us to make a decision at the utterance level yet. Therefore some kind of postprocessing of the aposteriori probabilities obtained at the segment level is required. This postprocessing, which will be referred to as decision aggregation, should output the aposteriori class probabilities at the utterance level.

The contribution of a particular voiced segment to the utterance level decision is assumed to depend on the relative duration of that voiced segment. This leads to more weight being given to longer segments than to shorter ones.

Three decision aggregation methods are explored *weighted average*, *weighted product*, and *maximum* expressed in (2), (3), and (4), respectively.

$$P(C_n \mid F_{Utt_A}) = \sum_{x=1}^{NumSegsInUtterance} length(Seg_X) \times P(Cn \mid F_{Seg_X}) \quad (2)$$

$$P(C_n \mid F_{Utt_A}) = \left( \prod_{x=1}^{NumSegsInUtterance} length(Seg_X) \times P(Cn \mid F_{Seg_X}) \right) \quad (3)$$

$$P(C_n \mid F_{Utt_A}) = \max\left\{ length(Seg_X) \times P(Cn \mid F_{Seg_X}) \right\} \quad (4)$$

After applying the decision aggregation technique and obtaining the utterance a posteriori class probabilities, and to make a final classification decision concerning the utterance, the Maximum A Posteriori rule (*MAP*) is used as described in (5). An utterance represented by $F_{Utt_A}$ is classified as $C_w$ if:

$$P(C_w \mid F_{Utt_X}) = \arg\max\left\{ P(C_n \mid F_{Utt_X}) \right\}, \quad \boldsymbol{1 < n < T} \quad (5)$$

Where *T* is the number of existing classes and *n* is the class index.

## 3. EXPERIMENTAL RESULTS

### 3.1. Speech Database

The proposed approach is tested on a superset of the *KISMIT* speech database that has been previously used in [3]. The database used in this work contains a total of 1002 speech samples of varying verbal content produced by three speakers in five classes of affective communicative intents. The classes are *Approval*, *Attention*, *Prohibition Weak*, *Soothing*, and *Neutral* utterances. Recording is performed with 16-bit per sample at a sampling rate of 8 kHz under varying amounts of noise between recordings. The speech recordings are of variable length but were mostly in the range of 1.8 sec to 3.25 seconds.

### 3.2. Classification of Segments

The goal of this experiment is to validate the assumption that affective intents are really expressed at the voiced segment level and can, accordingly, be detected from the feature vectors describing those voiced segments. The use of three types of features is compared, utterance-level only features, segment-level only features, and both utterance and segment level features. In all of the experiments in this work, the performance measure is taken as the total accuracy resulting from 10-fold cross-validation runs. Furthermore, the split between training and testing set is made at the utterance and not at the segment level. Results of 5-way classification using *K-NN* are summarized in Table 2.

TABLE 2. RESULTS OF CLASSIFICATION OF INDIVIDUAL SEGMENTS

| Features | Accuracy |
|---|---|
| Utterance Level | 78% |
| Segment Level | 61% |
| Both | 81% |

These results show that a classification accuracy of 61%, higher than what can be obtained with random classification, is obtained with segment level features only. This means that, indeed, affective intents can be recognized from individual voiced segments and not just from whole utterances. Also, it is observed that the use of the features calculated from the two temporal scales outperforms the use of either alone. It remains to be seen if the last observation holds for the classification of whole utterances.

### 3.3. Classification of Utterances

In this section we compare three approaches to the classification of whole utterances using two learning algorithms, *K-NN* and *SVM* making use of a customised version of the Multi-Instance Learning Kit in [8] adapted to the proposed classification approach.

**Approach A, Utterance Level:** The utterance is represented as a single object, consequently, the features are calculated at the utterance level. This approach is representative of the standard utterance level approaches in the literature.

**Approach B, Segment-Based with Segment Level Features:** The proposed segment-based representation is used and the individual segments are described with segment level features only.

**Approach C, Segment-Based with Both Segment and Utterance Level Features:** It is essentially the same as approach B, except that the segments are described with both utterance and segment level features.

TABLE 3. RESULTS OF CLASSIFICATION OF WHOLE UTTERANCES

| | Aggregation Scheme | Learning Algorithm | |
|---|---|---|---|
| | | K-NN | SVM |
| Approach A | | 81% | 78% |
| Approach B | Average | 76% | 62% |
| | Product | 76% | 62% |
| | Max | 62% | 55% |
| Approach C | Average | 87% | 83% |
| | Product | 87% | 83% |
| | Max | 83% | 83% |

Analysing the results in Table 3. , it is evident that *K-NN* consistently outperforms *SVM* in all three approaches. The aggregation schemes *Average* and *Product* result in similar performance while *Max* consistently delivers the worst performance when either classifier is used. Best performance of 87% is achieved using approach C while it is 81% using approach A and 76% using approach B.

Examining the confusion matrices resulting from approaches A and B (not shown), it is seen that the two approaches A and B exhibit different misclassification tendencies. Since approach C makes use of the features utilised in both approaches A and B, it also combines the contributions of both kinds of features resulting in a higher overall performance.

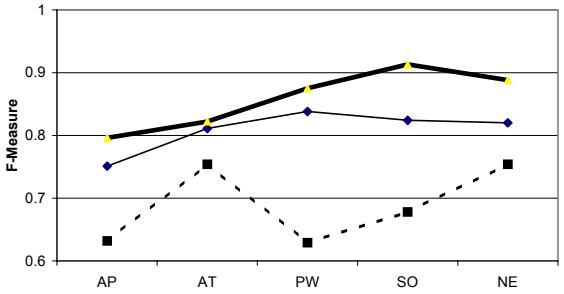


Fig. 3. Comparison of the Three Approaches: A (continuous), B (dashed), and C (bold) in One-Against-All Experiments for each of the five classes., Approval (AP), Attention (AT), Prohibition Weak (PW), and Neutral (NE)

To further bolster the evidence for the superiority of the proposed approach, i.e. approach C, one-against-all classification experiments are performed using the three previously outlined approaches. In a one-against-all classification scheme, we train five binary classifiers each of which is trained to recognise the utterances in one of the five classes in the database from the remaining four classes. Furthermore, to measure the performance of the individual classifiers generated we use the F-Measure, [9]. The F-Measure combines precision and recall into a single measure as shown in (6). The F-Measure is particularly useful when the data is unequally distributed over the classes.

$$F = 2 \times R \times P / (R + P) \quad (6)$$

Where F, R, and P correspond to F-Measure, Recall, and Precision, respectively.

Fig. 3. shows that approach C improves the classification performance over approaches A and B for all five classes.

Table 4. provides a comparison of results with those in claimed in [3]. The proposed approach outperforms the more optimized approach in [3].

TABLE 4. COMPARISON WITH PREVIOUSLY CLAIMED RESULTS

| | Approach | Accuracy |
|---|---|---|
| Previous Results in [3] | Utterance level features, feature selection, and *GMM* | 79% |
| | Serial hierarchical classifiers with utterance features and some segment features, feature selection and *GMM* | 82% |
| Implemented Approaches | Utterance level features and *K-NN* | 81% |
| | Segment level features and *K-NN* | 76% |
| | Both Utterance and Segment level features and *K-NN* | **87%** |

## 4. CONCLUSION

In this work a segment-based approach to the classification of affective intents in speech is proposed. The approach makes use of speech features calculated at two different temporal scales, the segment and the utterance level. For the classification of both individual segments and whole utterances, it was shown that the simultaneous use of the types of features yields better classification performance than using either type of features alone. Specifically, the performance was found to be superior in both 5-way and one-against-all classification experiments. Also, *K-NN* seems to be a more appropriate learning algorithm in our case than *SVM*. The segment-based approach presented here outperforms previous results in the literature, 87% to 82% 5-way classification accuracy.

## 5. REFERENCES


[1] R. Picard, "Affective Computing", MIT Press, Cambridge. 1997

[2] O. Pierre-Yves, "The production and recognition of emotions in speech: features and algorithms", International Journal of Human-Computer Studies, Volume 59, pp157-183 , July 2003

[3] C. Breazeal, L. Aryananda, "Recognition of Affective Communicative Intent in Robot-Directed Speech", Autonomous Robots, vol. 12, pp. 83-104, 2002

[4] M. Slaney, G. McRoberts, "A Recognition System for Affective Vocalization", Speech Communication, 39, pp. 367–384, 2003

[5] G. Katz, J. Cohn, C. Moore, "A combination of vocal f0, dynamic, and summary features discriminates between three pragmatic categories of infant-directed speech", Child Development 67, 205–217. 1996

[6] B. Schuller, G. Rigoll, M. Lang: "Hidden Markov Model-based Speech Emotion Recognition", Proc. 4th International Conference on Multimedia and Expo, Baltimore, MD, USA, Vol. I, pp. 401-404, 2003

[7] P. Boersma, D. Weenink, "PRAAT: a system for doing phonetics by computer", Report of the Institute for Phonetic Sciences of the University of Amsterdam 132, 1996, [http://www.praat.org]

[8] E. Frank and X. Xu. "Applying Propositional Learning Algorithms to Multi-instance data. Working Paper , Department of Computer Science, University of Waikato. 2003, [www.cs.waukato.nz/ml/milk]

[9] C. J. van Rijsbergen, "Information Retireval", Butterworths, London, 1979.