A Vowel Based Approach for Acted Emotion Recognition

Fabien Ringeval, *Mohamed Chetouani*

Institut des Systèmes Intelligents et de Robotique,
Université Pierre et Marie Curie – Paris 6, France
fabien.ringeval@isir.fr, mohamed.chetouani@upmc.fr

Abstract

This paper is devoted to the description of a new approach for emotion recognition. Our contribution is based on both the extraction and the characterization of phonemic units such as vowels and consonants, which are provided by a pseudo-phonetic speech segmentation phase combined with a vowel detector. Concerning the emotion recognition task, we explore acoustic and prosodic features from these pseudo-phonetic segments (vowels and consonants), and we compare this approach with traditional voiced and unvoiced segments. The classification is realized by the well-known k-nn classifier (k nearest neighbors) from two different emotional speech databases: Berlin (German) and Aholab (Basque).

Index Terms: emotion recognition, automatic speech segmentation, vowel detection

1. Introduction

The manifestation of emotions is a particularly complex domain of the human being communication, and concerns pluridisciplinary research areas such as affective computing, psychology, cognitive science, sociology and philosophy [1,2,3]. This pluridisciplinarity is due to the high variability of the human behaviour, which involves multifaceted emotions for both production and perception processes. Affect (feelings and the physical associated changes), cognition, personality, culture and ethics are the most important components of the emotion described in the literature [4]. Although that some studies list more than a hundred emotions terms [5], six primary emotions qualified as full-blown are widely accepted in the literature: fear, anger, happiness, boredom, sadness and disgust. Plutchik [6] postulates that all other emotions are mixed or derivate states, and occur as combinations, mixtures, or compounds of the primary ones.

Emotion-oriented computing aims at the automatic recognition and synthesis of emotions in speech, facial expression, or any other biological communication channel [7]. Concerning emotional speech classifications, one of the main difficulties resides in the determination of both feature sets and classifiers [8]. Other difficulties appear among them the definition of emotions and their annotation [9].

The commonly used feature extraction schemes are based on both acoustic and prosodic features resulting in a very large feature vector. The mainly used acoustic features are derived from speech processing (i.e. Mel Frequency Cepstrum Coding – MFCC), whereas the prosody is characterized by large statistics measures of pitch, energy and duration traditionally computed during the voiced segment. Since the manifestations of the emotions are particularly complex and concern different levels of communication, the identification and the extraction of optimally emotional speech descriptors are still open issues. The classification stage is usually based on machine learning methods such as distance based (k-nn), decision trees, Gaussian Mixture Models (GMM), Support Vector Machines and fusion of different methods [10].

2. A Vowel Based Approach

In this paper we propose a new feature extraction scheme based on a pseudo-phonetic approach (figure 1). The key idea of this work is to extract the features from different segments such as vowels and consonants. Some recent works have shown the relevance of a feature extraction at the phoneme level for the emotion recognition, especially for the vowels segments [11]. The vowels and consonants are identified by a segmentation of stationary segments (Divergence Forward Backward algorithm - DFB) combined with a vowel detector. This process is language independent and does not aim at the exact identification of phonemes as it could be done by a phonetic alignment. As a result, the obtained segments are termed pseudo-phonetic units.

This method has been introduced for automatic language identification [12]. Units similar to syllables termed "pseudo-syllable" have been defined by gathering the consonants preceding the detected vowels (C^nV structure) [13]. The study of these pseudo-syllables made possible the characterization of two main groups of language described in the literature: accentual (English, German and Mandarin) and syllabic (French and Spanish).

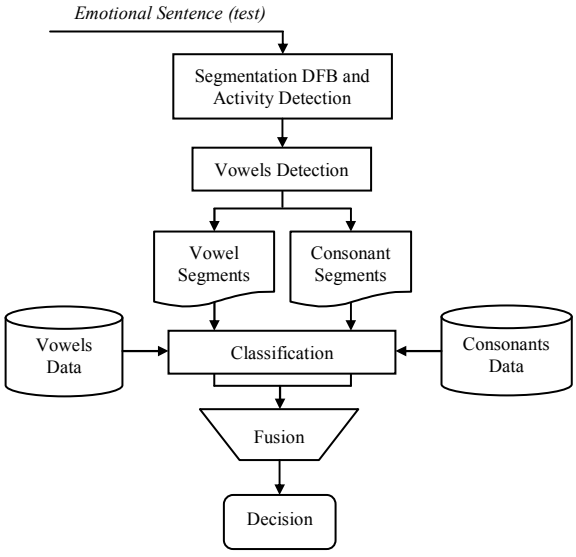

Figure 1: *Pseudo-phonetic approach: feature extraction and classification fusion.*

2.1. Databases Description

In this paper, two different emotional speech database are studied: Berlin [14], a German database, and Aholab [15] a Basque one. Both Berlin and Aholab contained acted speech from the six primary emotions plus a neutral style (Berlin contains the "Boredom" emotion).

The Berlin corpus is commonly used for emotion recognition [16, 17, 18]. 10 utterances (five short and five long) which could be used in everyday communication have been emotionally coloured by 10 gender equilibrated native German actors, with high quality recording equipment (anechoic chamber). 535 sentences marked as min. 60% natural and min. 80% recognisable by 20 listeners in a perception test have been kept and phonetically labelled in a narrow transcription. The Berlin corpus has a lexicon of 59 phonemes (24 vowels and 35 consonants).

The Aholab corpus [15] is composed of 702 sentences coming from a set of different sources: Basque newspapers, texts from several novels and others. From all these corpora (over 580,000 sentences), a reduced set of sentences have been extracted keeping the original frequency of the diphonemes as far as it was possible. A lexical balance has been then processed to get the 702 sentences. Concerning the emotions, two gender equilibrated professional speakers have acted the sentences in a semi-professional studio. The Aholab corpus has a lexicon of 35 phonemes (5 vowels and 30 consonants).

2.2. Pseudo-phonetic Unit Extraction

The automatic speech segmentation method reposes on the Divergence Forward Backward (DFB) algorithm [19]. Three kinds of segments are then identified: short or impulsive, transient and quasi-stationary zones. Once the segmentation is processed, a variance threshold is estimated and it enables speech activity detection. The vowels are identified by the spectral measure proposed by F. Pellegrino and al. [20]: "Reduced Energy Cumulating" (REC) (equation 1).

$$Rec(k) = \frac{E_{LF}(k)}{E_T(k)} \sum_{i=1}^{N} \left(E_i(k) - \overline{E}(k) \right)^+ \quad (1)$$

For a given sentence, peak detection on the REC curve allows vowel detection from the DFB segments. Speech segments that are not detected as vowels are considered as consonant segments. However these segments are not really consonants contrary to the vowels. This is mainly due to the fact that the DFB does not provide exact phonetic segments but rather stationary ones [13].

The vowel detector has been previously evaluated in [21] on three phonetically labelled corpuses: Berlin, TIMIT [22] and NTIMIT [23]. The TIMIT databases contain read speech in American English for respectively high quality and real telephone hand-set quality. 52 phonemes (20 vowels and 32 consonants) are composing their lexicon. Table 1 presents the results obtain in a vowel detection task. The Vowel Error Rate measure (VER) was used to this purpose. This measure sums the miss-detection and insertion ratios from the reference data contained in the phonetic transcription. It has been employed in many studies referenced in [12].

Table 1. *Performance of the vowel detector.*

Corpus	References (quantity)	Detections (in %)	Insertions (in %)	VER (in %)
Berlin	6,437	89.96	19.05	29.08
TIMIT	57,501	87.56	7.07	19.50
NTIMIT	57,493	81.06	5.13	24.07
Aholab	136,637	82.30	6.58	24.28
All	258,068	83.39	6.68	23.29

In terms of VER, the emotional corpuses produce the highest error rate contrary to the read speech databases (TIMIT and NTIMIT) where the detection is made easier. Indeed, the actors have emphasized their speech to produce the emotions, modifying thus the vowels characteristics. However, as the vowels from the Aholab corpus are represented by only five different phonemes, a low vowel insertion ratio is obtained.

3. Acoustic based Emotion Recognition

Two approaches are studied for the acoustic based emotion recognition: voiced and vowel based. Figure 1 illustrates the method employed for both features extraction and classification phases. During the voiced-based approach, speech is segmented by a sliding window of 32ms with a frame rate of 16ms. A voicing detector is then used to differentiate the voiced frames from the unvoiced ones. For the vowel-based approach, variable length frames according to the vowels and consonants segments are provided by the combination of both DFB segmentation and vowel detection as previously described (section 2). The feature extraction is performed by the computation of 22 MFCC parameters (Mel Frequency Cepstrum Coding). Concerning the classification phase, the MFCC features are labeled for each frame by a k-nn classifier (k nearest neighbors, k = 1). The scoring computation is based on a n-fold cross validation scheme where n is equal to 10. This process aims at the minimization of the empirical risk by creating different data configurations for both training and testing phases.

3.1. Voiced Fusion

Two emotion labels vectors are obtained by the k-nn classification from voiced and unvoiced MFCC features. In order to fuse them, we firstly compute their conditional posteriori probabilities $p(C_i \mid V)$ and $p(C_i \mid UV)$ according to the seven emotions classes from the data (C_1 to C_7). Secondly, two different approaches are employed to fuse them: static and dynamic. The used fusion method is a linear combination of the two conditional probabilities. For a given sentence, the emotion decision is taken by the following equation:

$$E = \arg\max_i \left(\lambda_V * p\left(C_i \mid V\right) + \lambda_{UV} * p\left(C_i \mid UV\right) \right) \quad (2)$$

The differentiation between static and dynamic fusion appears during the estimation of the weights from the voiced λ_V and unvoiced λ_{UV} classifiers. For the static fusion (equation 3), the weights are fixed during the entire phase test, and are estimated by their best combination from a set of chosen values (during the training phase). While the dynamic fusion is based on a voicing ratio computed for each test sentence. This method has been successfully used by Clavel and al. [24] for "fear" and "neutral" emotion discrimination. The voicing ratio r is defined as the proportion of voiced frames in speech. A power function is then applied to parameterize the decreasing velocity of the fusion weights: r^{α}.

$$E = \arg\max_i \left(r^{\alpha} * p\left(C_i \mid V\right) + \left(1 - r^{\alpha}\right) * p\left(C_i \mid UV\right) \right) \quad (3)$$

As we can expect, performance of the voiced frames classification from the two emotional databases are better than the unvoiced ones (table 2), which are not bad scores compared to their respective naïve classifier (classifying all the test utterances as the most common emotional class). The dynamic fusions achieve the best score with 75.00% for Berlin and 99.87% for Aholab, revealing the interest of the voicing ratio normalization during this phase. These results are in agreement with those from the literature [10,25].

3.2. Pseudo-Phonetic Fusion

Two labels vectors are obtained from the detected emotions by the k-nn classifications from both vowels and consonants. Since the studied databases provide a phonetic transcription of the emotional speech data, we therefore separately perform MFCC computations from the references and the detected segments (section 2). Posteriori probabilities estimated from vowels and consonants segments are combined similarly to the voiced and unvoiced fusion (cf. equation 4).

$$E = \arg\max_i \left(\lambda_{Vow} * p(C_i \mid Vow) + \lambda_{Csn} * p(C_i \mid Csn) \right) \quad (4)$$

Since the DFB segmentation trends to over-segment the speech signal for the consonant segment which are the most represented (mean detected consonant/vowel ratio from the emotional data is about 1.53 against 1.18 for the reference), dynamic fusion was not explored. Similar recognition rates are achieved for both references and detected segments. Though the references perform the best score, they appear to be less complementary than the detected ones. Indeed, the relative improvement on Berlin is 2.09% for the references against 6.28% for the detected segments (0.01% and 1.14% for Aholab).

Table 2. *Acoustic based emotion recognition rates.*

Approach	Recognition rate	
	Berlin	Aholab
Voiced	73.80%	99.08%
Unvoiced	49.00%	87.35%
Static Fusion	73.80%	99.83%
Dynamic Fusion	75.00%	99.87%
Vowels (reference)	76.90%	99.46%
Consonants (reference)	69.66%	97.60%
Static Fusion	78.51%	99.47%
Vowels (detected)	73.20%	98.47%
Consonants (detected)	65.60%	98.25%
Static Fusion	77.80%	99.59%
Naïve	22.43%	14.25%

4. Prosodic based Emotion Recognition

The voiced segments and those identified as "vowels" are exploited for the prosodic features extraction phase. The same features set are extracted for both approaches. Extracted features (Table 3) are based on statistical measures from the main components of the prosody (pitch, energy and rhythm).

4.1. Extraction of Prosodic Features

Whereas many descriptors have been proposed to characterize the pitch and the energy for emotional data, a few can be found concerning the rhythm. A first attempt for rhythmic modelling can be obtained through the segmental durations. To this end, we have extracted both vowels and consonants duration from the transcription of the two studied databases. Figure 2 and 3 present statistics from these data according to the seven emotions from Berlin and Aholab.

These figures illustrate many interesting results. First of all, we can notice that the emotions from the Basque corpus are much better separated than those from the German one. This is certainly due to the small number of speakers contained in the Aholab database (2 against 10 in Berlin). Secondly, many matching can be done between the speaking rate related to the emotions [26] and the vowels duration: "Fear" is perceived with the faster speaking rate while "Disgust" with the slowest one. "Anger" and "Sadness" are the most closed emotion, and "Happiness" is one of the most changeable inside and between the two studied databases.

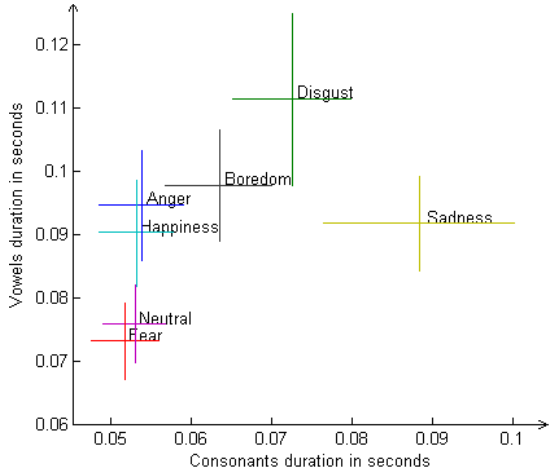

Figure 2: *Vowels and consonants duration from the Berlin database.*

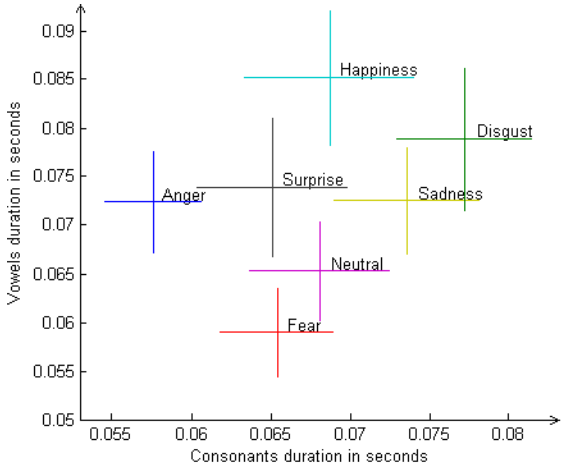

Figure 3: *Vowels and consonants duration from the Aholab database.*

Another way to extract rhythmic features from segmental durations can be realized by the measure proposed by Grabe and al.: "Pairwise Variability Indice" (PVI) [27]. This measure quantify the duration variability d_k from N vocalized intervals or successive intra vocalized (equation 5) and has been designed to characterize the dialects of the Britain English. This measure was implemented with the duration d_k between the voiced or "vowel" segments.

$$PVI = \frac{1}{N\text{-}1} \sum_{k=1}^{N-1} \left| d_k - d_{k+1} \right| \quad (5)$$

4.2. Emotion Recognition

A vector of conditional probability $p(C_i \mid Seg_x)$ is returned by the classifier for each tested segment Seg_x (voiced or vowel) according to the seven emotion class C_i. The decision of the emotion is obtained by an *argmax* function calculated on the sum of these probabilities. The results obtained by the prosodic models (table 4) confirm the interest of the vowels for the recognition of the emotions since they are higher again than those from the voiced segments. By way of comparison,

Shami and al. [10] obtain a score of 59% on the voiced segments from the Berlin corpus with an additional level of feature extraction (macro-prosody).

Table 3. *Extracted prosodic features.*

Features	Measures	Number
Pitch	Mean, Std, Normalised range, Max, Min, Std Δ	6
Energy	Mean, Max, Min, Std ΔΔ	4
Rhythm	Duration, PVI inter	2

Table 4. *Prosodic based emotion recognition rates.*

Approach	Recognition rate	
	Berlin	Aholab
Voiced	55.80%	80.82%
Vowels (reference)	56.40%	86.75%
Vowels (detected)	54.40%	83.96%
Naïve	22.43%	14.25%

5. Conclusion

A new feature extraction scheme for emotion recognition is presented in this paper: the vowel based approach. The automatic vowels detector is evaluated on four different databases with three different languages and one noisy environment representing more than 250,000 vowels. Obtained mean VER score from these data is 23.29%. Since performance decrease less than 4% for independent language detection, the employed method seems therefore appropriate for vowel extraction in multi-language context. Acoustic and prosodic emotion recognition systems are then processed on the emotional databases with two different approaches: voiced and the proposed pseudo-phonetic. For each classifier, the vowel based approach performs better than the voiced based one. A very high recognition rate of 78.51% is attained in acoustic recognition with the vowels and consonants fusion on Berlin, while the score of the voice based dynamic fusion is 75.00%.

Whereas an important information reduction takes place during the vowel-based approach, obtained performance from the pseudo-phonetic units for the two emotion recognizer are higher than for the voiced based approach. Those results are in agreement with the fact that among the speech structure units, vocalic nucleus has been proved to be the most perceptive one [28]. The pseudo-phonetic units can be thus considered as relevant to emotion recognition. We therefore propose to integrate the vowels and consonants units presented in this paper into emotions recognizer systems based on other machine learning techniques to improve their performance.

6. References

[1] Athanaselis, T., Bakamidis, S., Dologlou, I., Cowie, R., Douglas-Cowie, E., Cox, C., "ASR for emotional speech: clarifying the issues and enhancing performance", Neural Networks, 18(4):437-444, 2005.

[2] Plutchik, R., "The Psychology and Biology of Emotion", in Harper Collins [Ed], New York, 1994.

[3] Sherer, K. and al., "Acoustic correlates of task load and stress", in Proc. of ICSLP, 2002.

[4] Cowie, R., "Emotion-oriented computing: State of the art and key challenges", Humaine Network of Excellence, 2005.

[5] Appendix F, Labels describing affective states in five major languages, in K. Scherer [Ed]: Facets of emotion: recent research, Hillsdale, NJ: Erlbaum, [Version revised by the members of the Geneva Emotion Research Group], 241-243, 1988.

[6] Plutchik, R., "A general psychoevolutionary theory of emotion", in R. Plutchik & H. Kellerman [Ed], Emotion: Theory, research, and experience, New York: Academic, 1:3-33, 1980.

[7] Picard, R., "Affective Computing", The MIT Press, 1997.

[8] Ververidis, R., Kotropoulos, C., "Emotional speech recognition, features and method", Speech Communication, 48(9):1162-1181, 2006.

[9] Devillers, L., Vidrascu, L., Lamel, L., "Challenges in real-life emotion annotation and machine learning based detection", Journal of Neural Networks, 18(4):407-422, 2005.

[10] Shami, M., Verhelst, W., "An Evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech", Speech Communications, 49(3):201-212, 2007.

[11] Lee, C., M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., Narayanan, S., "Emotion recognition based on phoneme classes", in Proc. of ICSL, 2004.

[12] Rouas, J-L., Farinas, J., Pellegrino, F., André-Obrecht, R., "Rhythmic unit extraction and modelling for automatic language identification", Speech Communication, 47(4):436-456, 2005.

[13] Rouas, J-L., "Modelling long and short-term prosody for language identification", in Proc. of Interspeech, 2257-2260, 2005.

[14] F. Burkhardt and al., "A database of German emotional speech", in Proc. of Interspeech, 1517-1520, 2005.

[15] Saratxaga, I., Navas, E., Hernaez, I., Luengo, I., "Designing and recording an emotional speech database for corpus based synthesis in Basque", in Proc. of LREC, 2126-2129, 2006.

[16] Truong, K., Van Leeuwen, D., "An 'open-set' detection evaluation methodology for automatic emotion recognition in speech", Workshop on Paralinguistic Speech - between models and data, 5-10, 2007.

[17] Vogt, T., André, E., "Improving automatic emotion recognition from speech via gender differentiation", in Proc. of LREC, 2006.

[18] Datcu, D., Rothkrantz, L.J.M., "The recognition of emotions from speech using GentleBoost classifier. A comparison approach", CompSys-Tech, 5, 2006.

[19] André-Obrecht, R., "A new statistical approach for automatic speech segmentation", IEEE Transaction on ASSP, 36(1):29-40, 1988.

[20] Pellegrino F., André-Obrecht, R., "Automatic Language Identification: An alternative approach to phonetic modelling", Signal Processing, 80:1231-1244, 2000.

[21] Ringeval, F., Chetouani, M., "Exploiting a vowel based approach for acted emotion recognition", in A. Esposito and al. [Ed]: Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction, Springer-Verlag Publishers, 2008.

[22] Garofolo, J.-S. and al., "DARPA, TIMIT: Acoustic-Phonetic Continuous Speech Corpus", CDROM, NIST, 1993.

[23] Jankowski, C. and al., "NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database", ICASSP, 1:109-112, 1990.

[24] Clavel, C., Vasilescu, I., Richard, G., Devillers, L., "Voiced and unvoiced content of fear-type emotions in the SAFE corpus", in Proc. of Speech Prosody, 2006.

[25] Navas, E. and al., "Meaningful parameters in emotion characterisation", in A. Esposito and al. [Ed]: Verbal and Nonverbal Common Behaviours, Springer-Verlag Publishers, 4775:74-84, 2007.

[26] Murray, I. R., Arnott, J. L., "Towards the simulation of emotion in synthetic speech: A revue of the literature of human vocal emotion", Journal of Acoustics Society of America, 93(2): 1097-1198, 1993.

[27] Grabe, E., Nolan, F., Farrar, K., "IViE – A comparative transcription system for intonational variation in English", in Proc. of ICSLP, 1998.

[28] Pillot, C., Vaissière, J., "Vocal effectiveness in speech and singing: acoustical, physiological and perspective aspects. Applications in Speech Therapy", Laryngol Otol Rhinol Journal, 127(5):293–298, 2006.