Exploiting a Vowel Based Approach for Acted Emotion Recognition

Fabien Ringeval and Mohamed Chetouani

Université Pierre et Marie Curie – Paris 6,
Institut des Systèmes Intelligents et de Robotique, 3 rue Galilée,
94200 Ivry sur Seine, France
Fabien.Ringeval@isir.fr, Mohamed.Chetouani@upmc.fr

Abstract. This paper is dedicated to the description and the study of a new feature extraction approach for emotion recognition. Our contribution is based on 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. The segmentation algorithm is evaluated on both emotional (Berlin) and non-emotional (TIMIT, NTIMIT) databases. Concerning the emotion recognition task, we propose to extract MFCC acoustic features from these pseudo-phonetic segments (vowels, consonants) and we compare this approach with traditional voice and unvoiced segments. The classification is achieved by the well-known k-nn classifier (k nearest neighbors) on the Berlin corpus.

Keywords: Emotion recognition, automatic speech segmentation, vowel detection.

1 Introduction

The manifestation of emotions is a particularly complex field 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 process. Affect (feelings and the physical associated changes), cognition, personality, culture and ethics are the most important components described in the literature for the emotions [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, joy, 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

A. Esposito et al. (Eds.): HH and HM Interaction 2007, LNAI 5042, pp. 244–255, 2008.

feature extraction schemes are based on both acoustic and prosodic features resulting in a very large feature vector. The most discriminant ones are commonly determined by features selection algorithms [10]. This procedure was successful in many studies [11, 12, 13]. The mainly used acoustic features are derived from speech processing (e.g. Mel Frequency Cepstrum Coding - MFCC), whereas the prosody is characterized by large statistics measures of pitch, energy and duration 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 [14].

In this paper we propose a new feature extraction scheme based on a pseudo-phonetic approach (figure 1). Prosodic features are usually extracted from the voiced segments even if some approaches use also unvoiced segments for specific situations [15]. The key idea of this work is to extract the features from different segments such as vowels and consonants. These segments are identified by a segmentation of stationary segments (Divergence Forward Backward algorithm - DFB) combined with a vowel detector. The segmentation process is language independent and does not aim at the exact identification of phonemes as it can be done by a phonetic alignment. As a result, the obtained segments are termed pseudo-phonetic units.

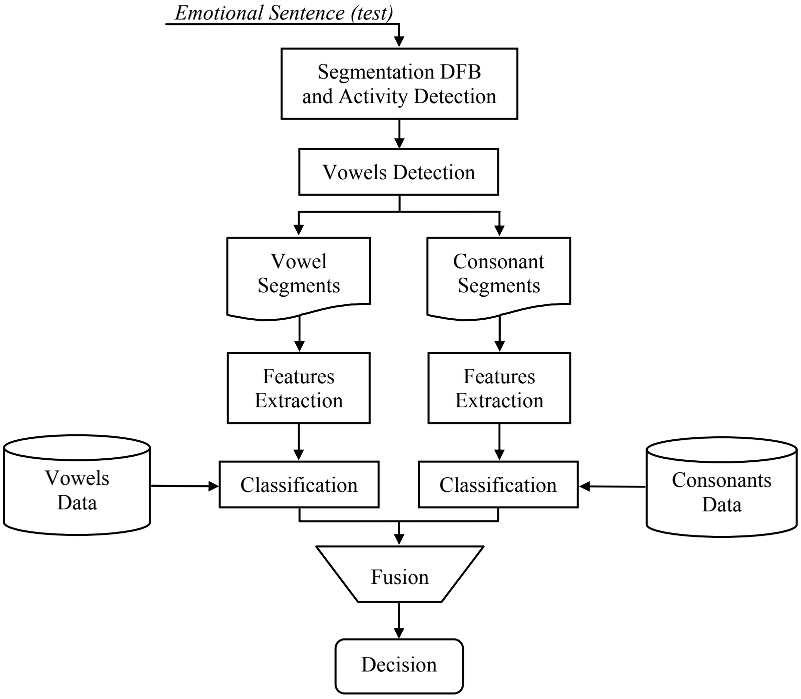

Fig. 1. Pseudo-phonetic approach: features extraction and classification fusion

The remainder of this paper is organized as follows: the segmentation DFB and the vowel detector are described and evaluated in section 2. Acoustic based emotion recognition results for both vowels-consonants units and voiced frames are presented in section 3.

2 Vowel-Based Approach

The following sections are dedicated to the study of the segmentation phase. For this purpose, we firstly present the used databases and secondly we give details about the segmentation algorithm. An evaluation of the vowel detection phase is also presented.

2.1 Description of the Databases

In order to evaluate our vowel based approach, we used three phonetically labelled databases: TIMIT [16], NTIMIT [17] and Berlin [18]. The TIMIT database contains 10 sentences pronounced by 630 speakers from 8 major dialect of American English. For each speaker, the sentences are grouped on three types. There are 5 phonetically-compact sentences (SX), 3 phonetically-diverse sentences (SI) and 2 dialect calibration sentences (SA). These sentences were labelled in a narrow transcription including word, orthographic, and phonetic time-alignment from a lexicon of 52 phonemes (20 vowels and 32 consonants). The NTIMIT database was created by transmitting all the 6.300 TIMIT recordings through a telephone handset and over both short and long-distance channels.

Concerning the Berlin corpus, it contains German emotional speech (six primary emotions plus the 'neutral' one) and is commonly used in the emotion recognition [19, 20, 21]. 10 utterances (five short and five long) which could be used in everyday communication have been emotionally coloured by ten 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 similar to the TIMIT database, excepted special diacritics and stress markers related to both emotional and articulatory characteristics. The Berlin corpus has a lexicon of 59 phonemes (24 vowels and 35 consonants).

2.2 Pseudo-phonetic Units Extraction

This section is dedicated to the description of the pseudo-phonetic segmentation method. The segmentation process used in this paper is based on the Divergence Forward Backward (DFB) algorithm [22]. For the DFB segmentation, the speech signal is hypothetically described by a sequence of stationary and transitory zones, each ones characterized by a statistical auto-regressive (AR) model. The method is based on a change detection criterion using prediction errors computed on two analysis windows M_0 and M_1 (figure 2). M_0 is a long-term and length fixed window sliding along the time axis, while M_1 is growing inside the long-term analysis window (M_0). The distance from the two AR models is performed by the use of a mutual entropy computation (Kullback divergence). Three segments sorts are then identified: shorts or impulsive, transitory and quasi-stationary. Once the segmentation is processed, a variance threshold from the resulting ones enables speech activity detection from these segments.

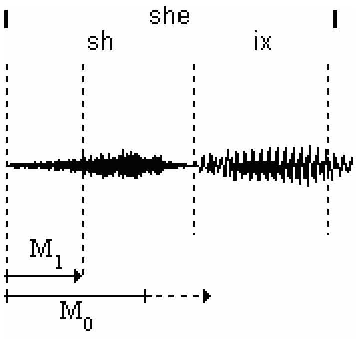

Fig. 2. Locations of the models in the DFB segmentation

2.3 Vowels Detection

According to the source-filter model of the vocal tract, the vowels are known to be characterized by a particular spectral envelope qualified as formantic. This spectral structure reveals the position of the formants (LPC model). The detection of vowel-like segments from the DFB ones is based on the characterization of the spectral envelope. To this purpose, Pellegrino et al. [23] proposed a spectral measure termed the 'Reduced Energy Cumulating' (REC) function (equation 1) for the vowel spectrum characterization. The key idea is to compare the energy computed from Mel bank filters. Firstly, the speech signal is segmented into overlapping frames. N energy values E_i are then extracted for each k frame, and those that are superior to their respective mean value $\bar{E}$ are cumulated and weighted by the energy ratio from low E_{LF} and total E_T frequency bands. The REC criterion is defined as follows:

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

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 classified 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. Figure 3 presents results of both DFB segmentation and vowel detection.

2.4 Evaluation of the Vowel Detector

We compare our automatic vowel detector based on the DFB segments with the reference ones provided by the phonetic transcriptions. The vowel error rate (VER) is used to evaluate the detector. The VER has been employed in many studies that are referenced in [24], and is expressed as follows:

$$\text{VER} = 100. \left[\frac{N_{del} + N_{ins}}{N_{vow}} \right] \% . \tag{2}$$

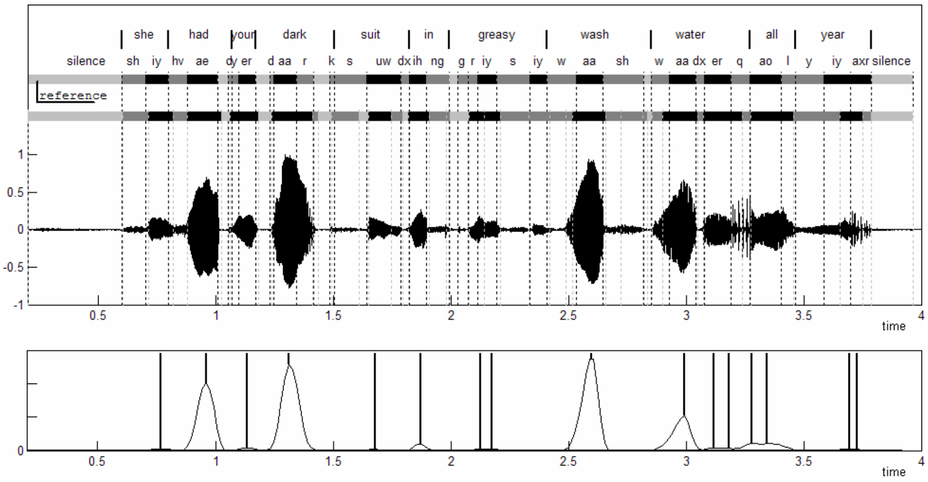

Fig. 3. (a) Comparison between the reference and the detected segments for the vowels (*black*) and the consonants (*dark gray*) from a TIMIT sentence. (b) Reduced Energy Cumulating (REC) function used for the vowels detection.

where N_{del} is the number of deletion (miss-detection), N_{ins} the number of insertion and N_{vow} the total number of vowels.

During the scoring process, a time-alignment between the detected vowels and the references ones is required to increment the number of detections, and one segment from our detector can not validate several reference ones. When no matching vowels from the N_{vow} references can be found, the number of insertions is then increased.

More than 120k vowels from the three databases are tested (table 1). In terms of VER, the best performances have been obtained on the TIMIT corpus 19.50%. This database contains read speech and due to the pronunciation, the detection is made easier. Though that NTIMIT is composed of the TIMIT data filtered at 8 kHz through real telephonic channel, implying an important and constraining spectral information reduction for vowels detection, the performances degradation are inferior to 5%. The vowels from the Berlin corpus are the best detected 90.06% as the most confused to 19.05%, producing the highest VER 29.08%. Berlin is an emotional database which contains full-blown acted emotions. Professional actors have therefore strongly emphasised their speech as the results show with high detection and insertion ratio from the vowel segments.

Table 1. Performances from the vowels detector on three speech corpus

Databases	Reference	Detection	Insertion	VER
Berlin	6437	5791 (89.96%)	1226 (19.05%)	29.08%
TIMIT	57501	50357 (87.56%)	4066 (7.07%)	19.50%
NTIMIT	57493	46601 (81.06%)	2948 (5.13%)	24.07%
All	121431	102755 (84.62%)	8240 (6.79%)	22.17%

Sensitivity to the duration from the vowel segments reference has been studied (figure 4). To this end, a percentage threshold on the duration ratio $d_{d.x}$ / $d_{r.x}$ from a detected vowel is introduced during the scoring computation.

Results show that more than 70% of the detected vowels from TIMIT and Berlin databases are included with a duration ratio superior than 50% from the reference ones (figure 5).

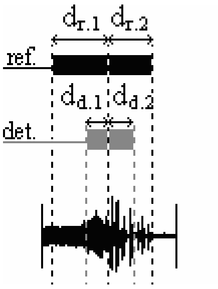

Fig. 4. Sensitivity to the duration from the detected vowel segments (*black*) according to the reference ones (*gray*)

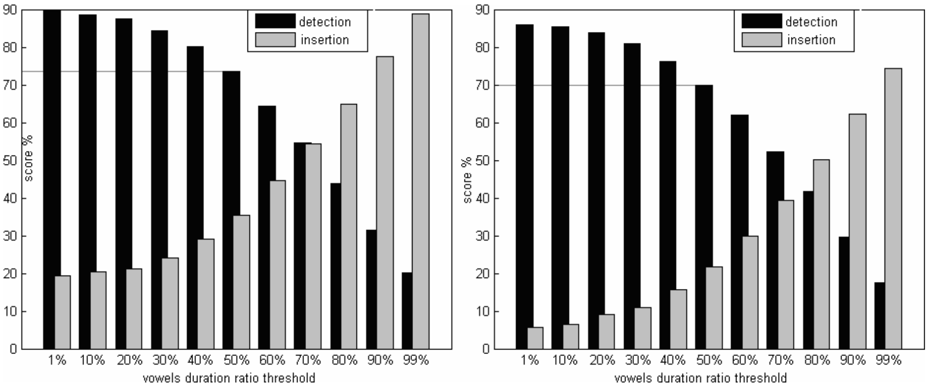

Fig. 5. Sensitivity of the detector to the duration from the vowel segments reference for Berlin and TIMIT databases respectively

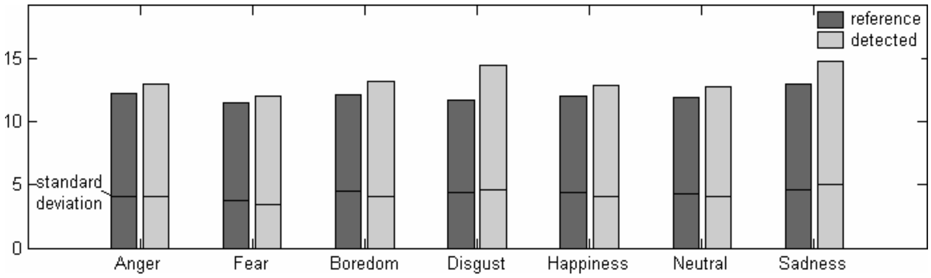

Fig. 6. Mean number of vowels per sentence (reference and detected) according to the seven emotions from Berlin

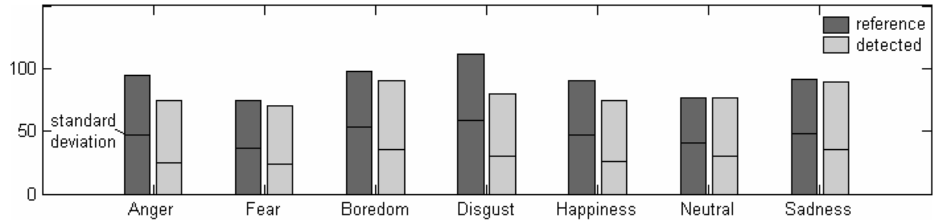

Fig. 7. Mean duration of vowels in *ms* (reference and detected) according to the seven emotions from Berlin

According to the seven emotions from Berlin, we extracted vowel characteristics such as mean number per sentence (ie. file) (figure 6) and mean duration (figure 7) from both reference and detected vowels. As the statistics show, the variability of the vowels number per sentence is better conserved than their mean values. On the whole, the detection is more less correctly done for all the emotion classes.

While three sorted groups can be defined from vowels duration statistics for both mean and standard-deviation: 1-Anger, Boredom and Disgust, 2-Happiness and Sadness and 3-Fear and Neutral with both lowest mean and distribution values. Since the neutral emotion is produced with no particular emphasis contrary to the others, few differences appear from reference and detected vowel characteristics.

Similar performance are obtained with a single configuration of the detector for the three databases (performance decrease only less than 4% for independent language detection), the employed approach seems therefore to be adequate for the vowels detection in speech.

3 Acoustic Based Emotion Recognition

This section presents methods and results for our acoustic-based emotion recognition approach for the Berlin corpus. Two approaches are studied: 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 segments according to the vowels and consonants segments are provided by the combination of both DFB segmentation and vowel detector as previously described (section 2). The feature extraction is performed by the computation of 24 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), which returns emotion labels corresponding to the nearest MFCC emotions data from the learning phase. The scoring computation is based on a n-fold cross validation scheme where n is equal to 10.

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 probabilities $p(C_i \mid V)$ and $p(C_i \mid UV)$ according to the seven emotions classes from Berlin (C_1 to C_7). Secondly, two different approaches are employed to fuse them: static and dynamic. The used fusion methods are a linear combination of the two conditional probabilities. For a given sentence, the emotion decision is taken by the following equation:

$$E = \operatorname{argmax}\left(\lambda_V * p\left(C_i \mid V\right) + \lambda_{UV} * p\left(C_i \mid UV\right)\right) \quad . \tag{3}$$

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), we estimate the optimal combination from the training data. Obtained values are then fixed for the all the test sentences. While the dynamic fusion is based on a voicing ratio computed for each test sentences. 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}.

$$E = \operatorname{argmax}\left(r^{\alpha} * p\left(C_i \mid V\right) + \left(1 - r^{\alpha}\right) * p\left(C_i \mid UV\right)\right) \quad . \tag{4}$$

This method has been successfully used by Clavel [15] for 'fear' and 'neutral' emotions differentiation.

As we can expect, performances of the voiced frames classification from Berlin 74.91% are better than the unvoiced ones 45.66%, which is not a low score compared with a naïve classifier 27.37% (classifying all the test utterances as the most common

Table 2. Performances from voiced and unvoiced classifiers with two different fusion approaches: linear combination (static) and voicing ratio combination (dynamic)

Approach	Recognition rate
Voiced	74.91%
Unvoiced	45.66%
Static Fusion	75.66%
Dynamic Fusion	76.04%
Naïve	27.37%

Table 3. Confusion matrix of the dynamic fusion based emotions recognition

	A.	F.	B.	D.	H.	N.	S.	Recognition
Anger	**130**	0	0	0	0	0	0	100%
Fear	24	**28**	4	0	0	1	3	47%
Boredom	0	1	**68**	0	0	7	4	85%
Disgust	9	1	2	**33**	0	5	0	66%
Happiness	**38**	0	0	5	27	0	0	39%
Neutral	0	1	16	0	0	**61**	2	76%
Sadness	0	0	4	0	0	0	**56**	93%
Confusion	47%	91%	69%	90%	100%	79%	88%	**76.04%**

emotional class) (table 2). Static fusion from these two classifiers reaches a very high recognition rate of 75.66% with the following weights: $\lambda_V = 0.8$ and $\lambda_{UV} = 0.1$. The dynamic fusion achieves the best score 76.04% with $\alpha = 0.5$, revealing the interest of the voicing ratio normalization during this phase.

In the case of the dynamic fusion, and according to the literature [12,14], ‘Anger’ and ‘Sadness’ are the best detected emotions with 100% and 93.33%, and ‘Fear’ and ‘Happiness’ are the worst ones: 46.67% and 38.57% respectively. The confusion matrix of this fusion can be found in table 3.

3.2 Vowels and Consonants Fusion

Unlike to the voiced fusion process, two labels vectors are obtained from the detected emotions by the k-nn classifications from both vowels and consonants. Since the Berlin corpus provides a phonetic transcription from the emotional speech data, we therefore separately perform MFCC computations from the references and detected segments (section 2). Similar classification process than for the static voiced fusion is applied to the pseudo-phonetic segments:

$$E = \operatorname{argmax}\left(\lambda_{Vow} * p\left(C_i \mid Vowels\right) + \lambda_{Csn} * p\left(C_i \mid Consonants\right)\right) . \quad (5)$$

Since the DFB segmentation trends to over-segment the speech signal for the consonant segment which are the most represented (detected consonant/vowel ratio is about 2.09 against 1.69 for the reference), dynamic fusion was not explored. Similar recognition rates are achieved for static fusion for both references and detected segments (table 4); the confusion matrices can be found in table 5. As the consonant segments are much more represented than the vowels, we can consider that the vowels MFCC features are more discriminant than the consonants ones in spite of a lower recognition rate: 46.42% (reference) and 52.56% (detected) against respectively 59.06% and 56.71% for the consonants. This consideration is agreed to the fact that among the speech structuring units, vocalic nucleus have been proved to be the most perceptive ones [25].

Table 4. Emotions recognition performances for the vowel-based approach

Approach	Reference	Detected
Vowels	46.42%	52.56%
Consonants	59.06%	56.71%
Fusion	60.57%	59.06%

Since a lot of information reduction takes place when only vowels and consonants segments are used for the emotions recognition, obtained performances are about 25% lower than the voiced-based fusion. During this approach, the speech is segmented by a sliding window with an overlap ratio of 50%, which involves numerous information extractions including some redundancy. Since a small quantity of information is provided by the MFCC extraction from the pseudo-phonetic approach unlike to the voice-based one, we can carefully suppose that the acoustic characteristics from the vowel segments are relevant for the emotion recognition.

Table 5. (a) Confusion matrix of the vowel-based emotions recognition (reference transcription). (b) Confusion matrix of the vowel-based emotions recognition (detected vowels).

(a)

	A.	F.	B.	D.	H.	N.	S.	Recognition
Anger	**126**	3	0	1	0	0	0	97%
Fear	**26**	22	5	0	1	2	4	37%
Boredom	8	5	**29**	0	1	18	19	36%
Disgust	18	0	4	**21**	2	2	3	66%
Happiness	**39**	3	2	0	19	6	1	42%
Neutral	3	3	10	0	2	**57**	5	71%
Sadness	0	0	5	0	0	8	**47**	78%
Confusion	57%	61%	53%	95%	76%	61%	59%	**60.57%**

(b)

	A.	F.	B.	D.	H.	N.	S.	Recognition
Anger	**130**	0	0	0	0	0	0	100%
Fear	**36**	12	3	1	1	4	3	20%
Boredom	7	3	**33**	0	1	20	16	41%
Disgust	18	3	1	**21**	1	3	3	42%
Happiness	**35**	0	0	4	14	12	5	20%
Neutral	5	0	9	0	0	**51**	15	64%
Sadness	3	0	2	1	0	2	**52**	87%
Confusion	52%	64%	68%	84%	60%	59%	53%	**59.06%**

4 Conclusion

A new features extraction scheme for the emotions recognition is presented in this paper: the vowel based approach. The automatic vowels detector is evaluated on three different databases with both language independent and noisy environment. Obtained mean VER score from these data is 22.17%. The best results are obtained on the read speech corpus 19.50% and the lowest on the emotional database 29.08%. The employed method seems therefore to be appropriate for vowel extraction. An emotion recognition system is processed on the Berlin corpus with two different approaches: voiced and the proposed pseudo-phonetic. The voiced-based approach reaches a very high recognition rate for dynamic fusion 76.04%, while the vowels and consonants fusion score is 60.57%. According to the important information reduction that takes place during the vowel-based approach, obtained results from the vowels segments can be carefully considered as relevant to the emotion recognition. Beyond the acoustic characterization of the pseudo-phonetic units, vowels and consonants segments may have a strong potential of interest in the emotion recognition as they convey a lot of prosodic information such as duration and rhythm. We therefore propose to integrate the vowels and consonants units presented in this paper into the emotions recognizer systems to improve their performances.

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, 437–444 (2005)
2. Plutchik, R.: The psychology and Biology of Emotion, Harper-Collins College, New York (1994)
3. Sherer, K., et al.: Acoustic correlates of task load and stress. In: Proceedings 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: Scherer, K. (ed.) Facets of emotion: Recent research, pp. 241–243. Lawrence Erlbaum, Hillsdale (1988) [Version revised by the members of the Geneva Emotion Research Group]
6. Plutchik, R.: A General Psychoevolutionary Theory of Emotion. In: Plutchik, R., Kellerman, H. (eds.) Emotion: theory, research, and experience, vol. 1, pp. 3–33. Academic press, New York (1980)
7. Picard, R.: Affective Computing. MIT Press, Cambridge (1997)
8. Ververidis, D., 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. Spence, C., Sajda, P.: The Role of Feature Selection in Building Patterns Recognizers for Computer-Aided Diagnosis. In: Kenneth, M.H. (ed.) Proceedings of SPIE, vol. 3338, pp. 1434–1441. Springer, Heildberg (1998)
11. Oudeyer, P.-Y.: The Production and Recognition of Emotions in Speech: Features and Algorithm. International Journal of Human-Computer Studies, special issue on Affective Computing 59(1-2), 157–183 (2002)
12. Xiao, Z., et al.: Hierarchical Classification of Emotional Speech. IEEE Transactions on Multimedia (submitted to, 2007)
13. Vogt, T., André, E.: Comparing Feature Sets for Acted and Spontaneous Speech in view of Automatic Emotion Recognition, pp. 474–477. ICME (2005)
14. Shami, M., Verhelst, W.: An Evaluation of the Robustness of Existing Supervised Machine Learning Approaches to the Classification of Emotions in Speech. Speech Communications 48(9), 201–212 (2007)
15. 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)
16. Garofolo, J.-S., et al.: DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM. In: NIST (1993)
17. Jankowski, C., et al.: NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database. ICASSP 1, 109–112 (1990)
18. Burkhardt, F., et al.: A Database of German Emotional Speech. In: Proc. of Interspeech (2005)
19. Truong, K., Van Leeuwen, D.: An open-set Detection Evaluation Methodology for Automatic Emotion Recognition in Speech. In: Workshop on Paralinguistic Speech - between models and data, pp. 5–10 (2007)
20. Vogt, T., André, E.: Improving Automatic Emotion Recognition from Speech via Gender Differentiation. In: Proc. of Language Resources and Evaluation Conference (2006)

21. Datcu, D., Rothkrantz, L., J.-M.: The Recognition of Emotions from Speech using GentleBoost Classifier. A Comparison Approach. CompSysTech Session V (2006)
22. André-Obrecht, R.: A New Statistical Approach for Automatic Speech Segmentation. IEEE Transaction on ASSP 36(1), 29–40 (1988)
23. Pellegrino, F., André-Obrecht, R.: Automatic Language Identification: an Alternative Approach to Phonetic Modelling. Signal Processing 80, 1231–1244 (2000)
24. Rouas, J.-L., Farinas, J., Pellegrino, F., André-Obrecht, R.: Rhythmic Unit Extraction and Modeling for Automatic Language Identification. Speech Communication 47(4), 436–456 (2005)
25. 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)