Cogn Comput (2009) 1:194–201
DOI 10.1007/s12559-009-9016-9

Time-Scale Feature Extractions for Emotional Speech Characterization

Applied to Human Centered Interaction Analysis

Mohamed Chetouani · Ammar Mahdhaoui · Fabien Ringeval

Published online: 22 April 2009

Abstract Emotional speech characterization is an important issue for the understanding of interaction. This article discusses the time-scale analysis problem in feature extraction for emotional speech processing. We describe a computational framework for combining segmental and supra-segmental features for emotional speech detection. The statistical fusion is based on the estimation of local a posteriori class probabilities and the overall decision employs weighting factors directly related to the duration of the individual speech segments. This strategy is applied to a real-world application: detection of Italian motherese in authentic and longitudinal parent–infant interaction at home. The results suggest that short- and long-term information, respectively, represented by the short-term spectrum and the prosody parameters (fundamental frequency and energy) provide a robust and efficient time-scale analysis. A similar fusion methodology is also investigated by the use of a phonetic-specific characterization process. This strategy is motivated by the fact that there are variations across emotional states at the phoneme level. A time-scale based on both vowels and consonants is proposed and it provides a relevant and discriminant feature space for acted emotion recognition. The experimental results on two different databases Berlin (German) and Aholab (Basque) show that the best performance are obtained by our phoneme-dependent approach. These findings demonstrate the relevance of taking into account phoneme dependency (vowels/consonants) for emotional speech characterization.

M. Chetouani (✉) · A. Mahdhaoui · F. Ringeval
Institut des Systèmes Intelligents et Robotique (ISIR), Université Pierre et Marie Curie-Paris6 (UPMC), 4 Place Jussieu, 75252 Paris Cedex, France
e-mail: mohamed.chetouani@upmc.fr

Keywords Emotional speech · Time-scales analysis · Feature extraction · Statistical fusion · Data-driven approach

Introduction

In the past few years, many attempts have been made to exploit computational models for human interaction analysis. This interaction can be directed towards other Human partners but also to machines (computers, virtual agents, or robots). Computational models aim to characterize signals emitted by human beings during interaction. Various frameworks are currently being used to analyze and to understand the interaction. One of them comes from cognitive psychology and focuses on emotion [1]. The key idea of this concept, also termed as affective computing, is that people perceive other's emotions through stereotyped signals (facial expressions, prosody, gestures, etc.). Another framework, coming from linguistic field, aims at understanding the meaning of these signals. Indeed, humans employ different strategies in order to convey the same message using multi-modal signals such as specific words, tone of voice, gesture, or more generally body language [2, 3]. Recently, a new framework has been introduced for the study of interaction termed as Social Signal Processing (SSP) [4] which focuses on the analysis of social signals by measuring the amplitude, frequency, and timing of prosody, facial movement, and gesture. SSP is different from the previously mentioned frameworks in the sense that it consists of non-linguistic and unconscious signals. More specifically, SSP aims to predict human behaviors or attitudes (agreement, interest, attention, etc.) by the analysis of non-verbal signals and it is considered as a separate channel of communication.

Most of the frameworks proposed in the literature for the understanding of interaction are based on the analysis of verbal and non-verbal signals [1, 3, 5]. The verbal component has been extensively investigated by the speech processing community. Non-verbal signals are expressed in a different way among the modalities. In [5], five different non-verbal behavioral cues have been defined: physical appearance, gestures and postures, face and eyes behaviors, vocal behavior, and space and environment behaviors. The combination of different codes make it possible to convey various information such as emotion, intention but are also useful for managing interaction, and/or sending relational messages (dominance, persuasion, embarrassment, etc.).

In this article, we focus on the analysis of a specific class of non-verbal behaviors which accompanies the verbal message termed as vocal behaviors in [5]. They allow to group empty speech pauses (silences), non-verbal vocalizations (i.e., filled pauses, laughters, cries, etc.), speaking styles (i.e., emotion, intention, etc.), and also turn-taking patterns. Even if these behaviors do not always have lexical meanings, they play a major role during natural interactions. Many efforts have been taken to extract features with no clear consensus on the most efficient ones [4, 6]. However, the prosody channel, characterized by the fundamental frequency (f0), the energy and the duration of sounds, has various functions in human communication since it serves to convey linguistic information, but also para-linguistic (e.g., speakers state), and non-linguistic information (e.g., age) [7, 8].

The remainder of this article presents various strategies for the fusion of time-scale features in order to study interactions. Section "Units for Emotional Speech Characterization" reports previous works in the literature associated with time-scale with a focus on unit selection problem for emotion recognition. Section "Combining Frame-Level and Segment-Based Approach for Intention Recognition in Infant-Directed Speech" describes the statistical framework for the fusion of frame and segment level features for infant-directed speech discrimination. Section "Data-Driven Approach for Time-Scale Feature Extraction" highlights the relevance of the pseudo-phonetic strategy for emotion recognition and provides results and discussion for time-scale analysis.

Units for Emotional Speech Characterization

The characterization scheme can be divided in two main steps: feature extraction and pattern classification. Regarding the first step, most methods are based on statistical measures of pitch, energy, and duration [6]. These statistical features (e.g., mean, range, max, min, etc.) have also been found to be related to human perception of emotions [9–11]. These features are usually termed as supra-segmental in contrast to segmental features (short-term) such as the Mel Frequency Cepstral Coefficients (MFCC) intensively used in speech processing. The classification step employs traditional machine learning and pattern recognition techniques such as distance based (nearest neighbor k-nn), decision trees, Gaussian Mixture Models (GMM), Support Vector Machines (SVM), and fusion of different methods [12].

One particular aspect of the speech emotion recognition process is the use of both static features (statistics) and static classifiers (e.g., k-nn or SVM). Indeed, the standard unit is the speaker turn level [12–14] which consists in the characterization of a whole sentence by a large number of features. This approach assumes that the emotional state is not changing during the speaker turn level. Even if the turn level approach has proven its efficiency, other units have been investigated for the exploitation of dynamical aspects of emotion. The methods can be divided into two groups: machine learning and data-driven methods.

Machine Learning Based Units

This approach employs machine learning techniques such as Hidden Markov Models [13]. Speech and speaker recognition techniques: short-term features and statistical modeling (GMM, HMM) have been successfully combined with a traditional turn based level approach [15]. In [16], a time-scale is identified by a the extraction of short-term feature extraction (25 ms windows, MFCC) and the use of statistical modeling (HMM). The time-scale is called by the authors chunk level. Once the HMM are trained (one for each emotion class), a Viterbi segmentation is applied resulting in specific sub-turn units that depend on emotion changes. Tested on emotion recognition tasks, the chunk level approach outperforms syllable based segmentation. This was mainly due to the fact that the proposed approach produces longer segments than the syllable segmentation method.

Data-Driven Units

The second approach aims at exploiting various knowledge about speech signals for the definition of units. For instance, voiced segments are known to convey more relevant information about emotion and focusing on these segments has been proven to be efficient [1, 12]. Various methods have been investigated for combining different levels [12, 14–17]. In [12], the Segment Based Approach (SBA) proposes to divide the whole utterance (turn level) on N voiced segments and then to characterize each voiced segments. The utterance based approach consists of the computation of statistical features (F0, energy, spectral shape) on the whole utterance while the SBA aims at describing more precisely each voiced segment. From this

local description an estimation of a posteriori class probabilities is done and the whole decision consists in merging the probabilities.

The SBA technique has been applied to emotion recognition for different well-known corpora and it outperforms the traditional utterance based feature extraction technique with k-nn classifiers (best classifier for these databases [12]): BabyEars 61.5% vs. 68.7% (SBA), Kismet 82.2% vs. 86.6% (SBA). However, with the same framework, different corpora (Berlin and Danish), and various classifiers (k-nn, SVM) different results have been achieved. For the Berlin corpus, SBA provides similar performance for both k-nn and SVM but it is outperformed by the traditional utterance level approach: k-nn 67.7% vs. 59.0% (SBA), SVM 75.5% vs. 65.5% (SBA). Once again the performance is correlated with the length of the utterance: SBA provides better results for short sentences (BabyEars, Kismet) while the turn level is more suited for longer ones (Berlin). Additionally, it should be noted that the performance also depends on the employed classifier as it has been found for the Danish corpus for instance: k-nn 49.7% vs. 55.6% (SBA), SVM 63.5% vs. 56.8%.

Data-Fusion Approach

The above experiments highlight the need of investigations into sub-units for emotional speech processing. In this article, we propose to address this problem by data-fusion of features extracted from different time-scales. The investigations are carried out in two phases:

- no assumption on the sub-unit (see. "Combining frame-level and segment-based Approach for Intention Recognition in infant-directed Speech") Section : the idea is to exploit speaker recognition techniques which are mainly based on frame-level modeling (all the frames are exploited for the characterization) as it is done in [16, 18].
- data-driven approach (see "Data-Driven Approach for Time-Scale Feature Extraction") : speech signals are characterized by prominent segments such as vowels which are then employed as sub-units.

The next sections present the two phases applied to different applications: motherese detection and traditional emotion recognition tasks.

Combining Frame-Level and Segment-Based Approach for Intention Recognition in Infant-Directed Speech

Expanded Intonation Contours

Communication of intentions is one of the major functions of interaction that uses both linguistic (syntax, semantic) and para-linguistic (prosody) elements. In the literature, communication of intentions with infants has received substantial attention [19, 20]. The main reason is that infants are not yet linguistically competent and the communication of intentions is done by prosody. More specifically, the communication is done by the parents by a specific register termed as infant-directed speech or motherese [21–23].

From an acoustic point of view, motherese has a clear signature (high pitch, exaggerated intonation contours). The phonemes, and especially the vowels, are more clearly articulated. Motherese has been shown to be preferred by infants over adult-directed speech and might assist infants in learning speech sounds. The exaggerated patterns facilitate the discrimination between the phonemes or sounds. Similarly to what happens with infants, several works have investigated modifications of speech registers when talking to animals [24], foreigners [20], or robots [25–27]. The important conclusion from this literature is the existence of common prosodic characteristics usually termed as expanded intonation contours (or Fernald's prototypical contours) [19, 22] due to their exaggerated contours: modulations of the fundamental frequency (F0) (mean, range).

Investigations on the characterization of these expanded contours have identified five categories [19]: rising, falling, flat, bell-shaped, and complex contours of the F0. These categories are used for the communication of intents such as attention, prohibition, approval, or comfort. For instance, rising contours aim at eliciting attention and encouraging a response while bell-shaped contours aim at maintaining attention. Consequently, adults convey intentional messages to infants by the use of these expanded contours. Among the most characterized speaker's intentions, one can cite: approval, attention, and prohibition. The classification of intention from speech signals offers an interesting application to the time-scale problem. Two approaches can be investigated: the use of only prosodic description of expanded intonation contours (voiced segments) or to also extract frame-level segments.

Motherese Detection

In order to study these intentional messages and more specifically the influence on engagement in an ecological environment, we followed a method usually employed for the study of infant development: home movies analysis. For more than 30 years, interest has been growing about family home movie of autistic infants. Typically developing infants gaze at people, turn toward voices and express interest in communication and especially to infant-directed speech. In contrast, infants who become autistic are characterized by the presence of abnormalities in reciprocal

social interactions and in patterns of communications [28]. Recently, researchers in autism pathology and researchers in early social interactions highlighted the importance of infant-directed speech for infants who will become autistic [29, 30]. First manual investigations [31] have shown a positive impact on the interaction and specially on the engagement: a response (vocalization, facial expression, gesture, etc.) by the infant to the production of infant-directed speech by the parents.

The study of home movies is very important for future research, but the use of this kind of database makes the study very difficult and long. The manual annotation of these films is very costly in time and including automatic detection of relevant events will be of great benefit to the longitudinal study. For the analysis of the role of infant-directed speech during interaction, we developed an automatic motherese detection system [30, 32]. The speech corpus used in these experiments is a collection of natural and spontaneous interactions usually used for child development research (home movies). The corpus consists of recordings in Italian of some mothers and fathers as they addressed their infants. The recordings are not carried out by professionals resulting in adverse conditions (noise, camera, microphones, etc.). We focus on one home video totaling 3 h of data describing the first year of an infant. Verbal interactions of the mother have been carefully annotated by two psycholinguists on two categories ($\kappa = 0.69$) : motherese and normal directed speech. From this manual annotation, we extracted 100 utterances for each class. The utterances are typically between 0.5 s and 4 s in length. For all the experiments in this paper a 10-fold cross-validation method is employed.

System Description

As a starting-point, and following the definition of motherese [21], we characterized the verbal interactions by the extraction of supra-segmental features (prosody). To evaluate the impact of frame-level feature extraction, segmental features are also employed. Consequently, the utterances are characterized by both segmental (short-time spectrum) and supra-segmental (statistics of fundamental frequency, energy) features. These features aim at representing the verbal information for the next classification stage based on machine learning techniques. Figure 1 shows a schematic overview of the final system [30, 32] which is described in more detail in the following paragraphs.

Supra-Segmental Characterization

The supra-segmental characterization follows the Segment Based Approach (see "Units for Emotional Speech Characterization"). Previous works on SBA [12] have shown to be more suited for short sentences as is usually the case in our corpus. The features consist of statistical measures (mean, variance and range) of both the fundamental frequency (F0) and the short-time energy estimated from voiced segments. An utterance U_x is segmented into N voiced segments (F_{xi}) obtained by F0 extraction. Local estimation of a posteriori probabilities is carried out for each segment. The utterance classification combines the N local estimations:

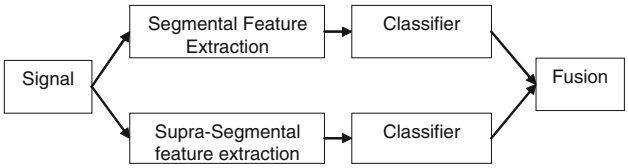

Fig. 1 Motherese classification system: fusion of features extracted from different time-scales

$$P(C_m|U_x) = \sum_{x_i=1}^{N} P(C_m|F_{xi}) \times \text{length}(F_{xi}) \quad (1)$$

where C_m represents the class membership.

The duration of the segments is introduced as weights of a posteriori probabilities: importance of the measured voiced segment (length(F_{xi})) with respect to the length of the utterance. The estimation has been carried out for various classifiers in [30, 32] and GMMs have been found to give good performance (number of parameters versus performance).

Segmental Characterization

For the computation of segmental features, a 20 ms window is used, and the overlapping between adjacent frames is 1/2. Mel Frequency Cepstrum Coefficients (MFCC) of order 16 were computed. We exploit traditional speaker recognition techniques [33]. For the whole utterance U_x, a posteriori probabilities are estimated resulting in the estimation of $P_{seg}(C_m|U_x)$. The estimation can be carried out for different time-scales: voiced, unvoiced, and whole-sentence.

To evaluate the system performance we used the receiver operating characteristic (ROC) methodology [34]. A ROC curve represents the tradeoff between the true positives (TPR = true positive rate) and false positives (FPR = false positive rate) as the classifier output threshold value is varied. A quantitative measure, the area under ROC curve (AUC), is computed and it represents the overall performance of the classifier over the entire range of thresholds. The results for different time-scales are presented in Table 1. As can be expected voiced segments provide better results than unvoiced ones. However, the

Table 1 Infant-directed speech discrimination performance of different time-scales for segmental features

Time-scale	Area under the ROC
Voiced	0.78
Unvoiced	0.55
Whole sentence	0.93

best results are obtained by using the whole-sentence as is usually done in speaker recognition showing that authentic emotional speech recognition is still an open issue compared to acted speech.

Fusion of Time-Scales

The segmental and supra-segmental characterizations provide different temporal information and a combination of them should improve the accuracy of the detector. Many decision techniques can be employed [35, 36] but we investigated a simple weighted sum of likelihoods from the different classifiers:

$$C_l = \lambda \cdot log\big(P_{\text{seg}}(C_{\text{m}}|U_{\text{x}})\big) + (1-\lambda) \cdot log\big(P_{\text{supra}}(C_{\text{m}}|U_{\text{x}})\big) \quad (2)$$

with $l = 1$ (motherese) or 2 (normal directed speech). λ denotes the weighting coefficient.

For the GMM classifier, the likelihoods can be easily computed from a posteriori probabilities ($P_{\text{seg}}(C_{\text{m}}|U_{\text{x}})$, $P_{\text{supra}}(C_{\text{m}}|U_{\text{x}})$)[37]. The weighting factor λ is automatically optimized in order to obtain the best results on the training part of the database. Since we employed a 10-fold cross-validation methodology, we present the means of the weighting factors.

Figure 2 presents the obtained ROC curves for segmental and supra-segmental features and the best combination ($\lambda = 0.6$). The weighting factor reveals a balance between the two different time-scales.

The above experiment results clearly show that even if motherese is defined as the modulation of supra-segmental features, using this basic definition does not produce efficient results (supra-segmental models). Real-world applications, such as analysis of home movies with authentic interactions and with a noisy environment, require the combination of the initial definition (supra-segmental features) with short-term features such as the MFCC as details of the short-term spectrum. Once again, for an efficient characterization, one should employ several features from different time-scales.

In this section we used short- and long-term features extracted from the short-term spectrum (MFCC) and from the evolution of supra-segmental features (statistics of F0, energy). By definition, the last set of features are extracted only from the voiced segments. Consequently all the voiced segments are processed identically even if very well-known distinctions exist between them (e.g., vowels versus consonants).

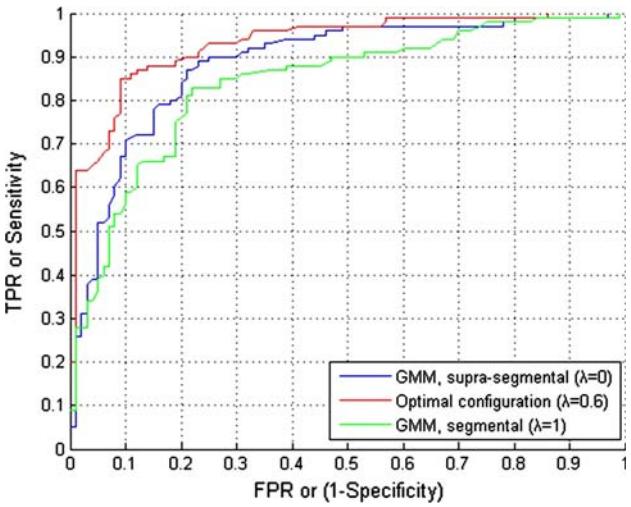

Fig. 2 ROC curve for segmental and supra-segmental systems

Data-Driven Approach for Time-Scale Feature Extraction

Nature of the Segments

The last section showed the relevance of combining frame and turn level approaches for emotional speech processing. One of the main limitations of this method relies on the fact that no sub-units are clearly identified: all the frames are exploited as it is usually done in speech and speaker recognition tasks. In this section, we propose to extract the frame levels on specific units defined here by taking into account the nature of the segments: vowel or consonant. Several investigations have been carried out on the relation of the nature of phonemes and emotional/affective states [17, 38–41]. All these works highlight the dependency between emotional states and the produced phonemes. In addition, vowel sounds seem to convey more emotional information than voiced consonant sounds [40]. These results motivate the need of different time-scale analysis for emotional speech processing.

We recently proposed a new feature extraction scheme aiming at exploiting the nature of phonemes [41]. The approach, described in Fig. 3, uses a first segmentation phase by the help of the Divergence Forward Backward (DFB) algorithm [42]. The resulting stationary segments are then classified as vowels by a criterion based on a spectral structure measure. This process is language independent and does not aim at the exact identification of

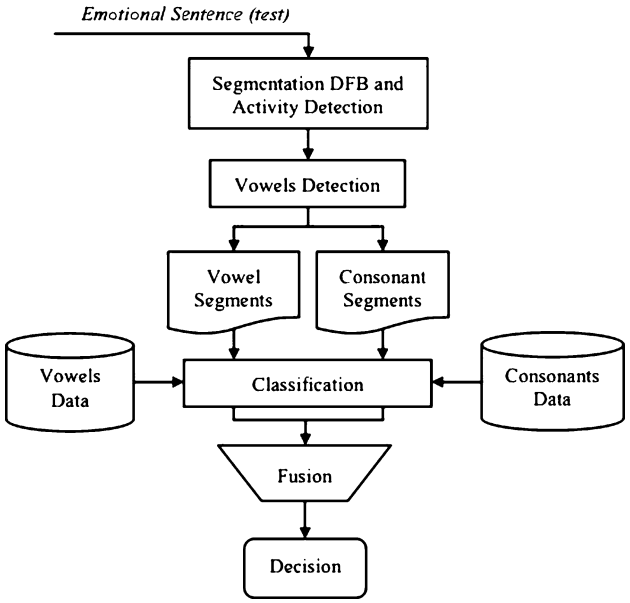

Fig. 3 Pseudo-phonetic approach: feature extraction, classification and fusion

phonemes as this could be done by a phonetic alignment. As a result, the obtained segments are termed as pseudo-phonetic units. This method has been introduced for automatic language identification [43] and consists in characterizing pseudo-syllables which have been defined by gathering the consonants preceding the detected vowels (C^nV structure). The study of these pseudo-syllables made possible the characterization of two main groups of language described in the literature: stressed (English, German) and syllabic (French and Spanish). We recently evaluated this segmentation system for both emotional and non-emotional speech with an average vowel error rate of 23.29% [41].

Corpora

We evaluate a time-scale analysis by using transcripted emotional databases: Berlin and Aholab. The Berlin corpus [44] is commonly used for emotion recognition. Ten utterances (five short and five long) that could be used in everyday communication have been emotionally colored by 10 gender equilibrated native German actors, with high quality recording equipment (anechoic chamber). A total of 535 sentences marked as minimum 60% natural and minimum 80% recognizable by 20 listeners in a perception test have been kept and phonetically labeled in a narrow transcription. The Berlin corpus has a lexicon of 59 phonemes (24 vowels and 35 consonants). The Aholab corpus [45] 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 was possible. Then, a lexical balance has been processed to get the 702 sentences. Concerning the emotions, two gender equilibrated professional speakers acted out the sentences in a semi-professional studio. The Aholab corpus has a lexicon of 35 phonemes (5 vowels and 30 consonants).

Classification With the Vowel–Consonant Time-Scale

The vowel–consonant time-scale is now exploited for emotion recognition problem by the use of the automatic pseudo-phonetic characterization (Fig. 3). We followed a segment-based approach (SBA) (equation 1) similar to what has been done for infant-directed speech discrimination (see "Combining Frame-Level and Segment-Based Approach for Intention Recognition in Infant-Directed Speech"). But here the segments are categorized as vowels and consonants. The utterance decision is made by the fusion of the local a posteriori class probabilities. This approach can be viewed as a segment dependent based approach:

$$E_{\mathrm{i}} = \arg \max_{\mathrm{i}} \{\lambda_{\mathrm{Vow}} P(C_{\mathrm{i}}|\mathrm{Vow}) + \lambda_{\mathrm{Cons}} P(C_{\mathrm{i}}|\mathrm{Cons})\} \quad (3)$$

where $P(C_{\mathrm{i}}|\mathrm{Vow})$ and $P(C_{\mathrm{i}}|\mathrm{Cons})$ denote the local a posteriori class probabilities respectively estimated from vowel and consonant segments. λ_{Vow} and λ_{Cons} represent the weighting factors for the fusion process. Different strategies have been employed for the estimation of the weighting factors [41]: static and adaptative (depending on the vowel–consonant duration ratio). Here, we report results for the static fusion process and the optimization is done on training data (as previously described in Section "Combining Frame-Level and Segment-Based Approach for Intention Recognition in Infant-Directed Speech").

The segment dependent approach has been used for classification [41] and we report the results for only segmental characterization (MFCC) and with a k-nn classifier for different times-scales. Table 2 presents the obtained classification scores for both Berlin and Aholab databases. Obviously, the extraction of segmental features from voiced segments gives better results than unvoiced ones and the fusion of them does not improve the performance. Similar results have been also found for the communicative intent classification (see "Combining Frame-Level and Segment-Based Approach for Intention Recognition in Infant-Directed Speech") but the main difference relies on the impact of taking all the frames (voiced and unvoiced) for authentic and noisy data as it is the case for the motherese application (see Table 1).

By using the transcription, we extracted the same features but from vowel and consonant segments. Promising

Table 2 Segmental based emotion recognition rates for different time-scales

Time-scale	Berlin (%)	Aholab (%)
Voiced	73.80	99.08
Unvoiced	49.00	87.35
Static fusion	73.80	99.83
Vowels (transcription)	76.90	99.46
Consonants (transcription)	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

results are obtained by the vowel time-scale for emotional speech processing: for the Berlin corpus, we obtained 76.90% for the vowel time-scale and 69.66% for the consonant time-scale. And by using the automatic and non perfect segmentation procedure (Fig. 3), we, respectively, obtain 73.20% for vowels and 65.60% for consonants. In addition, we also investigated the fusion of these dependent segment levels and the best results are still obtained by the transcription (78.51%) but the pseudo-phonetic approach (77.80%) is more efficient than the initial voiced segment (73.80%).

The classification results can be correlated to the number of speakers in the databases (Berlin: 10 versus Aholab: 2). The Aholab corpus presents less confusions between durations than the Berlin corpus and consequently the results are better.

Conclusion and Perspectives

This article presents a method for the combination of time-scale features: segmental (acoustic)/supra-segmental features (prosody) and also vowel/consonant phonemes. The cases studies provided (authentic and longitudinal interactions, acted corpus) illustrate the usefulness of combining different time-scale feature extractions for emotional speech classification. The advantages of this approach are the increase in robustness and also the integration of perceptual knowledge related to emotional sounds. The literature has shown the relative prominence of vowel sounds in the perception of emotions [9–11] and the reported framework makes it possible to employ this phenomenon.

Our future works will be devoted to the characterization of another important phenomenon such as the rhythm. The role of rhythm in the perception of sounds is very important [46] and it has been shown to be efficient for language identification [43, 47]. Most of the models proposed in the literature for the extraction of rhythmic features require the definition of a rhythmic unit (e.g., vowels, syllable) and a metric (inter, intra units)[48, 49]. A first application of these models to emotional speech processing reveals promising results [41].

References

1. Picard R. Affective computing. Cambridge, MA: MIT Press; 1997.
2. Argyle M. Bodily communication. 2nd edn. Madison: International Universities Press; 1988.
3. Kendon A, Harris RM, Key MR. Organization of behavior in face to face interactions. The Hague: Mouton; 1975.
4. Pentland A. Social signal processing. IEEE Signal Process Mag. 2007;24(4):108–11.
5. Vinciarelli A, Pantic M, Bourlard H, Pentland A. Social signals, their function, and automatic analysis: a survey. In: IEEE international conference on multimodal interfaces (ICMI'08). 2008. p. 61–8.
6. Schuller B, Batliner A, Seppi D, Steidl S, Vogt T, Wagner J, et al. The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In: Proceedings of interspeech; 2007. p. 2253–6.
7. Keller E. The Analysis of voice quality in speech processing. In: Chollet G, Esposito A, Faundez-Zanuy M, et al. editors. Lecture notes in computer science, vol. 3445/2005. New York: Springer; 2005. p. 54–73.
8. Campbell N. On the use of nonverbal speech sounds in human communication. In: Esposito A, et al. editors. Verbal and nonverbal communicational behaviours, LNAI 4775. Berlin, Heidelberg: Springer; 2007. p. 117–128.
9. Williams CE, Stevens KN. Emotions and speech: some acoustic correlates. J Acoust Soc Am. 1972;52:1238–50.
10. Sherer KR. Vocal affect expression: a review and a model for future research. Psychol Bull. 1986;99(2):143–65.
11. Murray IR, Amott JL. Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. J Acoust Soc Am. 1993;93(2):1097–108.
12. Shami M, Verhelst W. An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions, speech. Speech Commun. 2007;49(3):201–12.
13. Schuller B, Rigoll G, Lang M. Hidden Markov model-based speech emotion recognition. In: Proceedings of ICASSP'03, vol. 2. 2003. p. 1–4.
14. Lee Z, Zhao Y. Recognizing emotions in speech using short-term and long-term features. In: Proceedings ICSLP 98; 1998. p. 2255–58.
15. Vlasenko B, Schuller B, Wendemuth A, Rigoll G. Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing. Affect Comput Intell Interact. 2007;139–47.
16. Schuller B, Vlasenko B, Minguez R, Rigoll G, Wendemuth A. Comparing one and two-stage acoustic modeling in the recognition of emotion in speech. In: Proceedings of IEEE automatic speech recognition and understanding workshop (ASRU 2007), 9–13 Dec 2007, Kyoto, Japan; 2007. p. 596–600.
17. Jiang DN, Cai L-H. Speech emotion classification with the combination of statistic features and temporal features. In: Proceedings of ICME 2004 IEEE, Taipei, Taiwan; 2004. p. 1967–71.
18. Kim S, Georgiou P, Lee S, Narayanan S. Real-time emotion detection system using speech: multi-modal fusion of different timescale features. In: IEEE international workshop on multimedia signal processing; 2007.

19. Fernald A, Simon T. Expanded intonation contours in mother's speech to newborns. Dev Psychol.1987;20(1):104–13.
20. Uther M, Knoll MA, Burnham D. Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant directed speech. Speech Commun. 2007;49:2–7.
21. Fernald A, Kuhl P. Acoustic determinants of infant preference for Motherese speech. Infant Behav Dev. 1987;10:279–93.
22. Fernald A. Intonation and communication intent in mothers speech to infants: is the melody the message? Child Dev. 1989;60:1497–510.
23. Slaney M, McRoberts G. Baby ears: a recognition system for affective vocalizations. Speech Commun. 2003;39(3–4):367–84.
24. Burnham D, Kitamura C, Vollmer-Conna U. What's new, Pussycat? On talking to babies and animals. Science. 2002;296:1435.
25. Varchavskaia P, Fitzpatrick P, Breazeal C. Characterizing and processing robot-directed speech. In: Proceedings of the IEEE/RAS international conference on humanoid robots. Tokyo, Japan, 22–24 Nov 2001.
26. Batliner A, Biersack S, Steidl S. The prosody of pet robot directed speech: evidence from children. In: Proceedings of speech prosody; 2006. p. 1–4.
27. Breazeal C, Aryananda L. Recognition of affective communicative intent in robot-directed speech. Auton Robots. 2002;12:83–104.
28. Maestroa S, et al. Early behavioral development in autistic children: the first 2 years of life through home movies. Psychopathology. 2001;34:147–52.
29. Muratori F, Maestro S. Autism as a downstream effect of primary difficulties in intersubjectivity interacting with abnormal development of brain connectivity. Int J Dialog Sci Fall. 2007;2(1):93–118.
30. Mahdhaoui A, Chetouani M, Zong C, Cassel RS, Saint-Georges C, Laznik M-C, et al. Automatic Motherese detection for face-to-face interaction analysis. In: Anna Esposito, et al. editors. Multimodal signals: cognitive and algorithmic issues. Berlin: Springer; 2009. p. 248–55.
31. Laznik MC, Maestro S, Muratori F, Parlato E. Les interactions sonores entre les bebes devenus autistes et leur parents. In: Castarde MF, Konopczynski G, editors. Au commencement tait la voix. Ramonville Saint-Agne: Eres; 2005. p. 171–81.
32. Mahdhaoui A, Chetouani M, Zong C. Motherese detection based on segmental and supra-segmental features. In: IAPR international conference on pattern recognition, ICPR 2008; 2008.
33. Chetouani M, Faundez-Zanuy M, Gas B, Zarader JL. Investigation on LP-residual representations for speaker identification. Pattern Recogn. 2009;42(3):487–94.
34. Duda RO, Hart PE, Stork DG. Pattern classification. 2nd edn. New York: Wiley; 2000.
35. Kuncheva I. Combining pattern classifiers: methods and algorithms. Wiley-Interscience; 2004.
36. Monte-Moreno E, Chetouani M, Faundez-Zanuy M, Sole-Casals J. Maximum likelihood linear programming data fusion for speaker recognition. Speech Commun; 2009 (in press).
37. Reynolds D. Speaker identification and verification using Gaussian mixture speaker models. Speech Commun. 1995;17:91108.
38. Leinonen L, Hiltunen T, Linnankoski I, Laakso MJ. Expression or emotional–motivational connotations with a one-word utterance. J Acoust Soc Am. 1997;102(3):1853–63.
39. Pereira C, Watson C. Some acoustic characteristics of emotion. In: International conference on spoken language processing (ICSLP98); 1998. p. 927–30.
40. Lee CM, Yildirim S, Bulut M, Kazemzadeh A, Busso C, Deng Z, Lee S, Narayanan S. Effects of emotion on different phoneme classes. J Acoust Soc Am. 2004;116:2481.
41. Ringeval F, Chetouani M. A vowel based approach for acted emotion recognition. In: Proceedings of interspeech'08; 2008.
42. Andr-Obrecht R. A new statistical approach for automatic speech segmentation. IEEE Trans ASSP. 1988;36(1):29–40.
43. Rouas JL, Farinas J, Pellegrino F, Andr-Obrecht R. Rhythmic unit extraction and modelling for automatic language identification. Speech Commun. 2005;47(4):436–56.
44. Burkhardt F. et al. A database of German emotional speech. In: Proceedings of Interspeech; 2005. p. 1517–20.
45. Saratxaga I, Navas E, Hernaez I, Luengo I. Designing and recording an emotional speech database for corpus based synthesis in Basque. In: Proceedings of LREC; 2006. p. 2126–9.
46. Keller E, Port R. Speech timing: Approaches to speech rhythm. Special session on timing. In: Proceedings of the international congress of phonetic sciences; 2007. p. 327–29.
47. Tincoff R, Hauser M, Tsao F, Spaepen G, Ramus F, Mehler J. The role of speech rhythm in language discrimination: further tests with a nonhuman primate. Dev Sci. 2005;8(1):26–35.
48. Ramus F, Nespor M, Mehler J. Correlates of linguistic rhythm in the speech signal. Cognition. 1999;73(3):265–92.
49. Grabe E, Low EL. Durational variability in speech and the rhythm class hypothesis. Papers in Laboratory Phonology 7, Mouton; 2002.