Emotional Speech Characterization Based on Multi-Features Fusion for Face-to-Face Interaction

Ammar Mahdhaoui, Fabien Ringeval, Mohamed Chetouani
Université Pierre et Marie Curie, Paris VI
Institut des Systèmes Intelligents et de Robotique, CNRS UMR 7222
75252 Paris Cedex 05, France
Ammar.Mahdhaoui@isir.fr, Fabien.Ringeval@isir.fr, Mohamed.Chetouani@upmc.fr

Abstract*—Speech contains non verbal elements known as paralanguage, including voice quality, emotion and speaking style, as well as prosodic features such as rhythm, intonation and stress. The study of nonverbal communication has focused on face-to-face interaction since that the behaviors of communicators play a major role during social interaction and transport information between the different speakers. In this paper, we describe a computational framework for combining different 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-life application: detection of Italian motherese in authentic and longitudinal parent-infant interaction at home. The results suggest that short- and long-term information 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 discriminant feature space for acted emotion recognition.

***Keywords*—face-to-face interaction, emotional speech, time-scales analysis, feature extraction, statistical fusion, data-driven approach**

I. INTRODUCTION

Affective computing is a branch of the study and development of artificial intelligence that deals with the design of systems and devices that can recognize, interpret, and process human emotions. It is an interdisciplinary field spanning computer sciences, psychology, and cognitive science [1]. The aim of the affective computing is the automatic recognition and synthesis of emotions in speech, facial expressions, or any other biological communication channel [2]. Within the field of affective computing, this paper addresses the problem of the automatic recognition of emotions in speech for human interaction analysis. This interaction can be directed towards other Human partners but also to machines (computers, virtual agents or robots). Most of the frameworks proposed in the literature for the understanding of interaction are based on the analysis of verbal and non-verbal signals [2, 4]. While the verbal components were extensively investigated by the speech processing community, non-verbal signals are expressed in a different way among the modalities that imply specific approaches for their analysis.

In this paper, we focus on the analysis of a specific class of non-verbal behaviors which accompanies the verbal message termed as vocal behaviors in [4]. 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 [3, 5]. However, the prosody channel, which is 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 (e.g. emphase, modality), but also para-linguistic (e.g. speakers state according to the message) and non-linguistic information (e.g. affect, age) [6].

The remainder of this paper presents various strategies for the fusion of time-scale features in order to study interactions. Section 2 reports some works in the literature associated with time-scale with a focus on unit selection problem for emotion recognition. Section 3 describes the statistical framework for the fusion of frame and segment level features for infant-directed speech discrimination. While section 4 underlines the relevance of the pseudo-phonetic strategy for the recognition of emotions and provides results and discussion for multi-features fusion for both segmental/supra-segmental and time-scales.

II. EMOTION RECOGNITION

Emotion recognition can be divided into two major processing phases, namely the features extraction phase, and the classification on the features vectors. Regarding the first step, most methods are based on statistical measures of pitch, energy and duration [5]. These statistical features (e.g., mean, range, max, min) have also been found to be related to human perception of emotions [7, 8]. These features are usually termed as supra-segmental in contrast to segmental features (short-term) such as the MFCC (Mel Frequency Cepstral Coefficients) intensively used in the field of speech processing. In our study, we make a distinction between segmental and supra-segmental features that can be used in different modeling techniques to develop an emotion

recognition system. The classification phase 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 [9].

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 [9] 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 method used in our study calls data-driven method.

A. *Data-driven units*

This 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 [2, 9]. Various methods have been investigated for combining different levels [9, 10]. In [9], 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 [9]): 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%.

B. *Data-fusion approach*

The above experiments highlight the need of investigations into sub-units for emotional speech processing. In this paper, 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 (section 3): 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 [10, 12].
- data-driven approach (section 4): speech signals are characterized by prominent segments such as vowels which are then employed as sub-units.

III. Combining Frame-Level And Segment-Based Approach For Intention Recognition In Infant-Directed Speech

A. *Infant-directed speech characterisation*

There are different communicative signals in parent-infant communication. In our study, we focus on one type of signal which is infant-directed speech called also motherese. Infant-directed speech is a simplified language/dialect/register [14]. From an acoustic point of view, motherese has a clear signature (high pitch, exaggerated intonation contours) [13]. The phonemes, and especially the vowels, are more clearly articulated. The exaggerated patterns facilitate thus the discrimination between the phonemes or sounds. In addition, motherese plays a major role since it is a highly communicative and social event in parents-child communication that can elicit emotional reactions. Even if motherese is clearly defined in terms of acoustic properties, the modelling and the detection is expected to be difficult as it is the case for the majority of emotional speech recognition. Indeed, the characterization of spontaneous and affective speech in terms of features is still an open issue and several parameters have been proposed in the literature [5].

B. *Motherese detection*

Given that home movies offer a unique opportunity to follow infant development and parents-infants interactions, interest has been growing about family home movie of autistic infants. Infants who become autistic are characterized by the presence of abnormalities in reciprocal social interactions and in patterns of communications [15]. The study of home movies is very important for future research, but the use of this kind of database make the study very difficult and long because the manual annotation of these films is very costly and time consuming. From the manual analysis, we have found that 80% of positive sequences (vocalization, facial expression, gesture) have been induced by motherese. For the analysis of the role of infant-directed speech during interaction, we developed an automatic motherese detection system [16, 17]. 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 thus in adverse conditions (noise, camera, microphones). However, verbal interactions of the mother have been carefully annotated by two psycholinguists on two categories (kappa=0.69): motherese and adult-directed speech. From this manual annotation, we extracted 100

utterances for each class which are typically between 0.5s and 4s in length. A 10-fold cross-validation method is employed for the scoring computation.

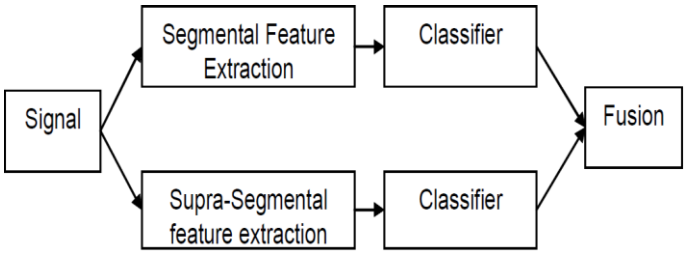

Figure 1. Motherese detection system: fusion of features exctracted from different time-scales

1) System description

Following the definition of motherese [14], 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. The utterances are therefore 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 [16, 17] which is described in more detail in the following paragraphs.

2) Supra-segmental characterization

The supra-segmental characterization follows the Segment Based Approach (see section 2). Previous works on SBA [9] 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:

$$P(C_m|U_x) = \sum_{xi=1}^{N} P(C_m|F_{xi}) * length(F_{xi}). \qquad (1)$$

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 [16, 17] and GMMs have been found to give good performance (number of parameters vs. performance). The best result using supra-segmental features is obtained with GMM classifier; 78% accuracy.

3) Segmental characterization

Traditional speaker recognition techniques are exploited for the computation of segmental features, a 20ms window is used, and the overlapping between adjacent frames is ½ [18]. Mel Frequency Cepstrum Coefficients (MFCC) of order 16 were computed. For the whole utterance U_x, a posteriori probabilities are estimated by the computation of $P_{seg}(C_m|U_x)$ and are carried out for different time-scales: voiced, unvoiced and whole-sentence. The best result using segmental features is obtained with GMM classifier; 82% accuracy. In order to evaluate the global system performance we used the Receiver Operating Characteristic (ROC) methodology [19]. A ROC curve represents the tradeoff between the true positives and false positives as the classifier output threshold value is varied. A quantitative measure, the area under the ROC is computed to represent the overall performance of the classifier over the entire range of thresholds. The results based on the ROC are given in table 1.

TABLE I. 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

As can be expected voiced segments provide better results than unvoiced ones. However, the best results are obtained by using the whole sentence as it is usually done in speaker recognition, showing that authentic emotional speech recognition is still an open issue compared to acted speech.

4) Fusion of time-scales

The segmental and supra-segmental characterizations provide different temporal information and a combination of them should improve the performance of the detector. Many decision techniques can be employed [20, 21] but we investigated a simple weighted sum of likelihoods from the different classifiers:

$$C_l = \lambda * \log\left(P_{seg}\left(C_m|U_x\right)\right) + (1-\lambda) * \log\left(P_{supra}\left(C_m|U_x\right)\right). \quad (2)$$

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 [22]. The weighting factor λ is automatically optimized in order to obtain the best results on the training part of the database. The best result is obtained by combining segmental and supra-segmental features using GMM classifier with a weighting factor equals to 0.5 revealing a balance between the two features; 87.5% accuracy.

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.

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 is 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 vs. consonants).

IV. Data-Driven Approach For Time-Scale Feature Exctraction

A. 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 [23, 24, 25, 11, 26]. All these works highlight the dependency between emotional states and the produced phonemes. In addition, vowels seem to convey more emotional information than voiced segments [25]. 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 [26]. The approach, described in figure 2, uses a first segmentation phase by the help of the Divergence Forward Backward (DFB) algorithm [28]. 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 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 [29] 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). This segmentation system was evaluated for both emotional and non-emotional speech with an average vowel error rate of 23.29% [27].

B. Corpus

We used a transcripted emotional database for the analysis. The Berlin corpus is commonly used for emotion recognition [30]. 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). 535 sentences marked then as min. 60% natural and min. 80% recognizable by 20 listeners in a perception test have been kept and phonetically labeled in a narrow transcription.

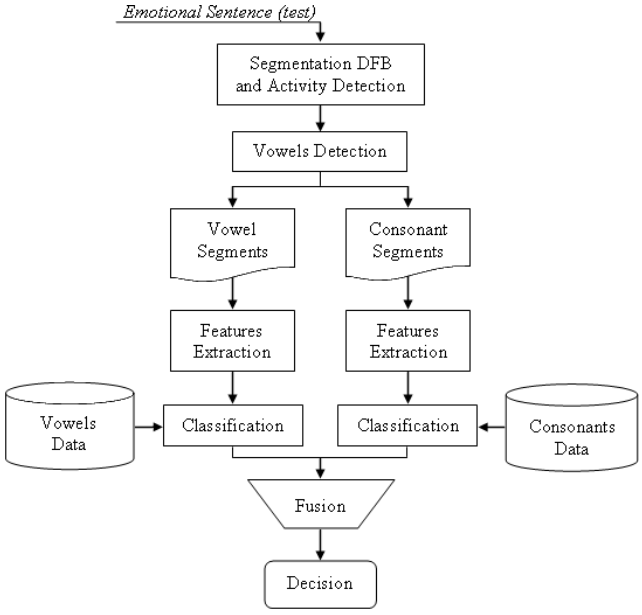

Figure 2. Pseudo-phonetic approach: feature extraction, classification and fusion

C. 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 (figure 2) and transcription data. We followed a segment based approach (SBA) (equation 1) similar to what has been done for infant-directed speech discrimination (see section 3). But here the segments are categorized as vowels and consonants. At the end of the process, the utterance decision is made by the fusion of the local a posteriori class probabilities. This approach can be seen as a segment-dependent based approach:

$$E_i = \arg\max_i \left\{ \lambda_{Vow} P\left(C_i \middle| Vow\right) + \lambda_{Cons} P\left(C_i \middle| Cons\right) \right\} . \quad (3)$$

Where $P(C_i|Vow)$ and $P(C_i|Cons)$ 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 used for the estimation of the weighting factors [27]: static and adaptative. In this paper, we report results only for the static fusion process and the optimization is done for all the training data (as previously described in section 3). Segmental (16 MFCC, 32 ms window, overlap ratio ½) and supra-segmental features (27 statistics on pitch, energy and formants F1, F2, F3 and F4) are extracted for different time-scales (e.g. voiced/vowels). These features are normalized by a Z-score where both mean and standard-deviation values (μ and σ) are extracted from the features for a given emotional class. Because physical characteristics of the speaker (e.g. speaking style) can be considered as a bias for the emotion recognition system, we used neutral style.

$$F_x = \frac{F_x - \mu}{\sigma}, x = \{Vow, Cons\} . \quad (4)$$

Different approaches are used for Z-score normalization depending of the use or not of a priori information. During "Z-score all" (ZA), features are normalized without using a priori information by both mean and standard-deviation values computed from neutral data (equation 4), while others approaches repeat this step for each speaker/gender (ZS/ZG), including thus dependency on the a priori information. Since prosodic features are numerous compared to MFCC, we selected one hundred of them by using the Relief-F algorithm [31].

A posteriori class probabilities provided by the classification of both acoustic and prosodic features are fused as it was accomplished in section 3 (equation 2) except that we did not use a *log* on the a posteriori probabilities. The segment dependent based approach (equation 3) is then employed for the fusion of the different time-scales. A k-nn classifier was used for the experiments with a stratified version of the 10-cross-fold validation scheme since the quantity of data differs along the emotions.

TABLE II. ACTED EMOTION RECOGNITION RESULTS FOR ACOUSTIC AND PROSODIC FEATURES FUSION ACCORDING TO DIFFERENT TIME-SCALES

TIME-SCALE	Normalization			
	Raw	ZA	ZG	ZS
Consonants (transcription)	$53.28_{1/0}$	$60.98_{7/3}$	$58.16_{8/2}$	$66.23_{7/3}$
Consonants (detected)	$54.41_{9/1}$	$56.85_{7/3}$	$64.54_{7/3}$	$65.48_{6/4}$
Unvoiced	$41.46_{7/3}$	$41.28_{7/3}$	$39.96_{8/2}$	$45.22_{8/2}$
Vowels (transcription)	$57.41_{9/1}$	$59.47_{8/2}$	$66.98_{7/3}$	$70.36_{8/2}$
Vowels (detected)	$58.91_{9/1}$	$61.54_{6/4}$	$66.23_{8/2}$	$71.11_{8/2}$
Voiced	$60.98_{9/1}$	$63.41_{9/1}$	$67.35_{8/2}$	$71.48_{9/1}$

TABLE III. ACTED EMOTION RECOGNITION RESULTS FOR DIFFERENT TIME-SCALES FUSION

TIME-SCALES FUSION	Normalization			
	Raw	ZA	ZG	ZS
Vowels – Consonants (transcription)	$60.60_{5/5}$	$64.73_{5/5}$	$70.36_{6/4}$	$75.42_{5/5}$
Vowels – Consonants (detected)	$61.91_{7/3}$	$64.35_{5/5}$	$67.73_{6/4}$	$74.67_{6/4}$
Voiced – Unvoiced	$62.29_{8/2}$	$64.35_{9/1}$	$68.11_{8/2}$	$71.67_{9/1}$

Obviously, the fusion of both segmental and supra-segmental features from voiced segments gives better results than unvoiced ones. Fusion of these two time-scales does not improve the performance in a significant way. Similar results have been found for the communicative intent classification (section 3) 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). While voiced segments outperform the five others studied time-scales when both acoustic and prosodic features are fused (tables 2), promising results are obtained by the vowel based approach for emotional speech processing when time-scales fusion is achieved. For the Berlin corpus, we obtained 75.42% for the vowel based approach and 71.67% for the voice based approach. In addition, when the automatic and non perfect segmentation procedure is used (figure 2), we obtain 74.67% which is more efficient than the initial voiced segments and close to the transcription data. Z-score normalizations highlight the fact that introducing a priori information during the features extraction phase clearly improves the emotion recognition performances, especially when the knowledge of the speaker is available.

V. CONCLUSION

This paper 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 [7, 8] 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 [32] and it has been shown to be efficient for language identification [29, 33]. 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 measure to perform on it [34, 35]. The first applications of these models to emotional speech processing reveal promising results [26, 36].

REFERENCES

[1] Jianhua, T. and Tan, T. "Affective Computing: A Review". Affective Computing and Intelligent Interaction LNCS 3784: 981995, Springer. doi:10.1007/11573548, (2005)

[2] Picard, R. "Affective Computing". MIT Press, (1997)

[3] Pentland, A. Social signal processing, IEEE Signal Processing Magazine, 24(4), 108-11, (2007)

[4] Vinciarelli, A., Pantic, M., Bourlard, H. and Pentland, A. Social signals, their function, and automatic analysis: a survey, IEEE International Conference on Multimodal Interfaces (ICMI'08), 61-68, (2008)

[5] Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L. and Aharonson, V. The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In Proc. of Interspeech, 2253-56, (2007)

[6] Campbell, N. On the Use of NonVerbal Speech Sounds in Human Communication. in A. Esposito et al. (Eds.): Verbal and Nonverbal Commun. Behaviours, LNAI 4775, 117-28, 2007; Springer-Verlag Berlin Heidelberg, (2007)

[7] Williams, C.-E. and Stevens, K.-N. Emotions and speech: some acoustic correlates. Journal of the Acoustical Society of America 52, 1238-50, (1972)

[8] Murray, I.-R. and Amott, J.-L. Toward the simulation of emotion in synthetic speech - A review of the literature on human vocal emotion. Journal of Acoustic Society of America, 93(2), 1097-1108, (1993)

[9] Shami, M. and Verhelst., W. An Evaluation of the Robustness of Existing Supervised Machine Learning Approaches to the Classification of Emotions, Speech.Speech Communication, 49(3), 201-12, (2007)

[10] Vlasenko, B., Schuller, B., Wendemuth, A. and Rigoll, G. Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing, Affective Computing and Intelligent Interaction, 139-47, (2007)

[11] Schuller, B., Vlasenko, B., Minguez, R., Rigoll, G. and Wendemuth, A., Comparing one and two-stage acoustic modeling in the

recognition of emotion in speech. In Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2007), 596-600, (2007).

[12] Kim, S., Georgiou, P., Lee, S. and Narayanan, S. Realtime emotion detection system using speech: Multimodal fusion of different timescale features. IEEE International Workshop on Multimedia Signal Processing, (2007)

[13] Fernald, A. and Simon, T. Expanded intonation contours in mother's speech to newborns. Developmental Psychology, 20(1), 104-13, (1987)

[14] Fernald, A. and Kuhl, P. Acoustic determinants of infant preference for Motherese speech. Infant Behavior and Development, 10, 279-93 (1987)

[15] Maestroa S. el al., Early Behavioral Development in Autistic Children: The First 2 Years of Life through Home Movies, Psychopathology, 34:147-52, (2001)

[16] Mahdhaoui, A., Chetouani, M., Zong, C., Cassel, R. S., Saint-Georges, C., Laznik, M-C., Maestro, S., Apicella, F., Muratori, F. and Cohen, D. Automatic Motherese Detection for Face-to-Face Interaction Analysis, In Anna Esposito et al. (ed.) Springer-Verlag, Multimodal Signals: Cognitive and Algorithmic Issues, 248-55, (2009)

[17] Mahdhaoui, A., Chetouani, M., Zong, C., Motherese Detection Based On Segmental and Supra-Segmental Features. International Conference on Pattern Recognition, ICPR 2008, (2008)

[18] Chetouani, M., Faundez-Zanuy, M., Gas, B. and Zarader, J.-L. Investigation on LP-residual representations for speaker identification, Pattern Recognition, 42(3), 487-94, (2009)

[19] Duda, R.-O., Hart, P.-E. and Stork, D.-G. "Pattern classification". 2nd edition. New York: Wiley, (2000)

[20] Kuncheva, I. "Combining pattern classifiers: Methods & algorithms". Wiley, (2004)

[21] Monte-Moreno, E., Chetouani, M., Faundez-Zanuy, M., and Sole-Casals, J. Maximum Likelihood Linear Programming Data Fusion for Speaker Recognition. Speech Communication, In press. (2009).

[22] Reynolds, D. Speaker identification and verification using Gaussian mixture speaker models, Speech Communication, 17, 91-108, (1995)

[23] Leinonen, L., Hiltunen, T., Linnankoski, I., and Laakso, M.-J. Expression or emotional-motivational connotations with a one-word utterance, Journal of Acoustic Society of America, 102(3):53-63, (1997)

[24] Pereira, C. and Watson, C. Some Acoustic Characteristics of Emotion. In International Conference on Spoken Language Processing (ICSLP98), 927-30, (1998).

[25] Lee, C. M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S. and Narayanan, S. Effects of emotion on different phoneme classes. Journal of Acoustic Society of America., 116(4), 2481, (2004).

[26] 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, LNAI 5042, Springer-Verlag Berlin Heidelberg, 244-245, (2008)

[27] Ringeval, F., Chetouani, M. A Vowel Based Approach for Acted Emotion Recognition, In Proc. of Interspeech'08, 2763-2766, (2008)

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

[29] 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)

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

[31] Robnik, M. and Konenko, I, Theoretical and empirical analysis of ReliefF and RReliefF, Machine Learning Journal, 53, 23-69, (2003)

[32] Keller, E. and Port, R. Speech timing: Approaches to speech rhythm. Special session on timing. In Proc. of the International congress of phonetic sciences, 327–329, (2007)

[33] Tincoff, R., Hauser, M., Tsao, F., Spaepen, G., Ramus, F. and Mehler, J. The role of speech rhythm in language discrimination: further tests with a nonhuman primate. Dev Sci. 8(1):26–35, (2005)

[34] Ramus F., Nespor M., Mehler J. Correlates of linguistic rhythm in the speech signal. Cognition. 73(3):265–92, (1999)

[35] Grabe, E., and Low, E. Durational variability in speech and the rhythm class hypothesis. Papers in Laboratory Phonology 7, Mouton, (2002)

[36] Ringeval, F., and Chetouani, M. Non-linearity modelling of speech rhythm using Hilbert-Huang Transform. In Proc. of NOn-LInear Speech Processing, NOLISP 2009, (2009)