Automatic Motherese Detection for Face-to-Face Interaction Analysis

A. Mahdhaoui[1], M. Chetouani[1], C. Zong[1], R.S. Cassel[2,3], C. Saint-Georges[2,3], M-C. Laznik[4], S. Maestro[5], F. Apicella[5], F. Muratori[5], and D. Cohen[2,3]

[1] Institut des Systèmes Intelligents et de Robotique, CNRS FRE 2507, Université Pierre et Marie Curie, Paris, France
[2] Department of Child and Adolescent Psychiatry, AP-HP, Groupe Hospitalier Pitié-Salpétrière, Université Pierre et Marie Curie, Paris, France
[3] Laboratoire Psychologie et Neurosciences Cognitives, CNRS UMR 8189, Paris, France
[4] Department of Child and Adolescent Psychiatry, Association Santé Mentale du 13ème, Paris, France
[5] Scientific Institute Stella Maris, University of Pisa, Italy

Ammar.Mahdhaoui@robot.jussieu.fr, Mohamed.Chetouani@upmc.fr

Abstract. This paper deals with emotional speech detection in home movies. In this study, we focus on infant-directed speech also called "motherese" which is characterized by higher pitch, slower tempo, and exaggerated intonation. In this work, we show the robustness of approaches to automatic discrimination between infant-directed speech and normal directed speech. Specifically, we estimate the generalization capability of two feature extraction schemes extracted from supra-segmental and segmental information. In addition, two machine learning approaches are considered: k-nearest neighbors (k-NN) and Gaussian mixture models (GMM). Evaluations are carried out on real-life databases: home movies of the first year of an infant.

Keywords: motherese detection, feature and classifier fusion.

1 Introduction

Since more than 30 years, interest has been growing about family home movies of infants who will become autistic. Typically developing infants gaze at people, turn toward voices and express interest for communication. In contrast, infants who became autistic will be characterized by the presence of abnormalities in reciprocal social interactions and in patterns of communications [1]. In this paper, we focus on a verbal information which has been recently shown to be crucial for engaging interaction between the parent and infant. This verbal information is called "motherese" (also termed infant-directed speech) and it is a simplified language/dialect/register [2] that parents use spontaneously when speaking to their young baby. From an acoustic point of view, motherese has a clear signature (high

A. Esposito et al. (Eds.): Multimodal Signals, LNAI 5398, pp. 248–255, 2009.

pitch, exaggerated intonation contours). The phonemes, and particularly the vowels, are more clearly articulated. Motherese has been shown to be preferred by infants over adult-directed speech and might assist infants during the language acquisition process [3]. The exaggerated patterns facilitate the discrimination between the phonemes or sounds. Motherese plays also a major role during social interactions. However, even if motherese is clearly defined in terms of acoustic properties, the modeling and the detection are expected to be difficult which is the case of the majority of emotional speech. Indeed, the characterization of spontaneous and affective speech in terms of features is still an open question and several parameters have been proposed in the literature [4].

As a starting-point and following the definition of motherese [2], we characterized the verbal interactions by the extraction of supra-segmental features (prosody). However, segmental features are often used in speech segmentation. Consequently, the utterances are characterized by both segmental (short-time spectrum) and supra-segmental (statistics of fundamental frequency, energy and duration) features. These features aim at representing the verbal information for the next classification stage based on machine learning techniques.

This paper presents a framework for the study of parent-infant interaction during the first year of age focusing on the engagement produced by motherese, in normal infants or infants who will become autistic. To this purpose, we use a longitudinal case study methodology based on the analysis of home movies. We focus on a basic and crucial task namely the classification of verbal information as motherese or adult directed speech which implies the design of robust motherese detector. Section 2 presents the longitudinal corpora used. The proposed method is described in section 3 which needs specific attentions to the different stages: feature extraction, classification and decision fusion. The experiments performed are discussed in section 4. Finally, the conclusions and suggestions for further works are detailed in section 5.

2 Database

The speech corpus used in this experiment is a collection of natural and spontaneous interactions usually used for children development research (home movies). The corpus consists of recordings in Italian of mother and father as they addressed their infants. In addition, the analysis of home movies makes it possible to set up a longitudinal study (months or years) and gives information about early behaviors of autistic infants, a long time before the diagnostic would be made by the clinicians. However, this large corpus makes it inconvenient for people to review it. Also, the recordings are not done by professionals resulting in adverse conditions (noise, camera, microphones). We focus on one home video totaling 3 hours of data describing the first year of an infant. Verbal interactions of the mother have been carefully annotated by a psycholinguist on two categories: motherese and normal directed speech. From this manual annotation, we extracted 100 utterances for each class. The utterances are typically between 0.5s and 4s in length.

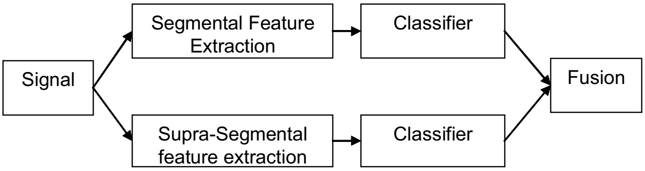
Signal
Segmental Feature Extraction
Supra-Segmental feature extraction
Classifier
Classifier
Fusion

Fig. 1. Motherese classification system

3 System Description

This work aims at designing an automatic detection system for the analysis of parent-infant interaction. This system will provide an independent classification of the utterances. In order to improve the system, two different approaches have been investigated individually: segmental and supra-segmental. Figure 1 shows a schematic overview of the final system which is described in more details in the following paragraphs.

3.1 Feature Extraction

In this paper, we evaluate two approaches respectively termed as segmental and supra-segmental features. The first ones are characterized by the Mel Frequency Cepstrum Coefficients (MFCC) while the second ones are characterized by statistical measures of both the fundamental frequency (F0) and the short-time energy, these statistical measures are calculated from the voiced segments.For the computation of segmental features, a 20ms window is used, and the overlapping between adjacent frames is 1/2. A parameterized vector of order 16 was computed. The supra-segmental features are characterized by 3 statistics (mean, variance and range) of both F0 and short-time energy resulting on a 6 dimensional vector. One should note that the duration of the acoustic events is not directly characterized as a feature but it is taken into account during the classification process by a weighting factor (eq. 5). The feature vectors are normalized (zero mean, unit standard deviation).

3.2 Classification

In this study, two different classifiers were investigated: k-NN and GMM. The k-NN classifier [6] is a distance based method while the GMM classifier [5] is a statistical model. For the fusion process, we adopted a common statistical framework for both the k-NN and the GMM by the estimation of a posteriori probabilities.

A Posteriori Probabilities Estimation. The Gaussian mixture model (GMM) [5] is adopted to represent the distribution of the features. Under the assumption that the feature vector sequence $x = \{x_1, x_2, ..., x_n\}$ is an independent identical distribution sequence, the estimated distribution of the d-dimensional feature vector x is a weighted sum of M component Gaussian densities $g_{(\mu,\Sigma)}$, each parameterized by a mean vector μ_i and covariance matrix Σ_i; the mixture density for the model C_m is defined as :

$$p(x|C_m) = \sum_{i=1}^{M} \omega_i g_{(\mu_i,\Sigma_i)}(x) \tag{1}$$

Each component density is a d-variate Gaussian function:

$$g_{(\mu,\Sigma)}(x) = \frac{1}{(2\pi)^{d/2}\sqrt{det(\Sigma)}} e^{-1/2(x-\mu)^T \Sigma^{-1}((x-\mu))} \tag{2}$$

The mixture weights ω_i satisfy the following constraint: $\sum_{i=1}^{M} \omega_i = 1$. The feature vector x is then modeled by the following posteriori probability:

$$p_{gmm}(C_m|x) = \frac{p(x|Cm)P(C_m)}{p(x)} = \frac{p(x|Cm)P(C_m)}{\sum_{j=1}^{2} p(x|C_j)P(C_j)} \tag{3}$$

where $P(Cm)$ is the prior probability for class Cm, we assume equal prior probabilities. We use the expectation maximization (EM) algorithm for the mixtures to get maximum likelihood.

The k-NN classifer [6] is a non-parametric technique which classifies the input vector with the label of the majority k-nearest neighbors (prototypes). In order to keep a common framework with the statistical classifier (GMM), we estimate the posteriori probability that a given feature vector x belongs to class C_m using k-NN estimation [6]:

$$p_{knn}(C_m|x) = \frac{k_m}{k} \tag{4}$$

where k_m denotes the number of prototypes which belong to the class C_m among the k nearest neighbors.

Segmental and Supra-Segmental Characterizations. Segmental features (i.e. MFCC) are extracted from all the frames of an utterance U_x independently of the voiced or unvoiced parts. Posteriori probabilities are then estimated by both GMM and k-NN classifiers and are respectively termed $P_{gmm,seg}(C_m|U_x)$ and $P_{knn,seg}(C_m|U_x)$.

The classification of supra-segmental features follows the segment-based approach (SBA)[7]. An utterance U_x is segmented into N voiced segments (F_{xi}) obtained by F0 extraction (cf. 3.1.). Local estimation of posteriori probabilities is carried out for each segment. The utterance classification combines the N local estimations.

$$P(C_m|U_x) = \sum_{x_i=1}^{N} P(C_m|F_{xi}) \times length(F_{xi}) \tag{5}$$

The duration of the segments is introduced as weights of the posteriori probabilities: importance of the voiced segment ($length(F_{xi})$). The estimation is also carried out by the two classifiers resulting on supra-segmental characterizations: $P_{gmm,supra}(C_m|U_x)$ and $P_{knn,supra}(C_m|U_x)$.

3.3 Fusion

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 [9] but we investigated a simple weighted sum of likelihoods from the different classifiers:

$$C_l = \lambda.log(P_{seg}(C_m|U_x)) + (1 - \lambda).log(P_{supra}(C_m|U_x)) \tag{6}$$

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 the posteriori probabilities ($P_{gmm,seg}(C_m|U_x)$, $P_{gmm,supra}(C_m|U_x)$) [5]. However, the k-NN estimation can produce a null posteriori probability (eq. 4) incompatible with the computation of the likelihood. We used a solution recently tested by Kim et al. [8], which consists in using the posteriori probability instead of the log probability of the k-NN:

$$C_l = \lambda.log(e^{P_{knn,seg}(C_m|U_x)}) + (1 - \lambda).log(P_{gmm,supra}(C_m|U_x)) \tag{7}$$

Table 1. Table of combinations

$Comb_1$	$P_{knn,seg}$	$P_{knn,supra}$
$Comb_2$	$P_{gmm,seg}$	$P_{gmm,supra}$
$Comb_3$	$P_{knn,seg}$	$P_{gmm,supra}$
$Cpmb_4$	$P_{gmm,seg}$	$P_{knn,supra}$
$Comb_5$	$P_{gmm,seg}$	$P_{knn,seg}$
$Comb_6$	$P_{gmm,supra}$	$P_{knn,supra}$

Consequently, for the k-NN classifier we used equation 7 while for the GMM the likelihood is conventionally computed. We investigated cross combinations listed in table 1.

4 Experimental Results

4.1 Classifier Configuration

To find the optimal structure of our classifiers, we adjust the different parameters: the number of Gaussian (M) for GMM classifier and the number of neighbors (k) for the k-NN classifier. Table 2 shows the best configuration for both GMM and k-NN classifiers with segmental and supra-segmental features, the

Table 2. Accuracy of optimal configurations

	segmental	supra-segmental
k-nn	72.5% (k=11)	61% (k=7)
GMM	79.5% (M=15)	82% (M=16)

same table shows that GMM classifier trained with prosody features outperforms the other classifiers confirming the definition of motherese [2].

4.2 Fusion of Best Systems

To evaluate the system performance we used the receiver operating characteristic (ROC) methodology [6]. A ROC curve (Fig. 2) represents the tradeoff between the true positives (TPR = true positive rate) and false positives (FPR = false positive rate) of as the classifier output threshold value is varied. Two quantitative measures of verification performance, the equal error rate (EER) and the area under ROC curve (AUC), were calculated. It should be noted that while EER represents the performance of a classifier at only one operating threshold, the AUC represents the overall performance of the classifier over the entire range of thresholds. Hence, we employed AUC and not EER for comparing the verification performance of two classifiers and their combination. However, the result obtained in table 2 motivates an investigation on fusion of both features and classifiers following the statistical approach described in section §3.2. The combination of features and classifiers is known to be efficient [9]. However, one

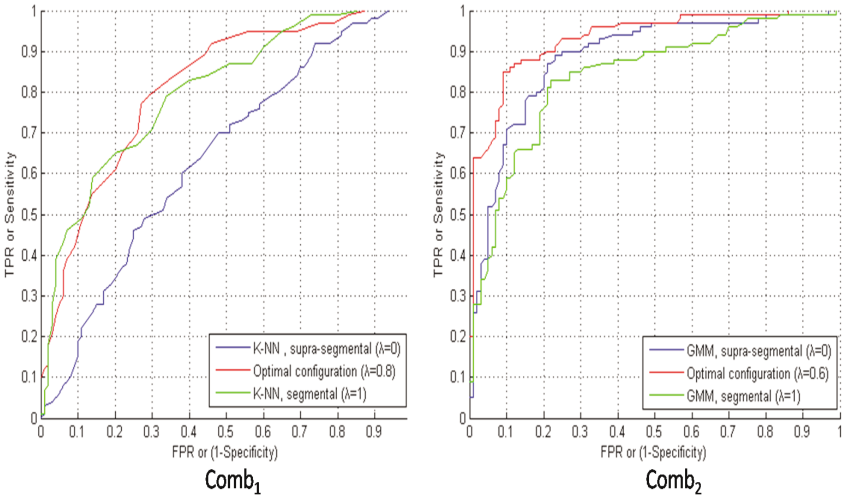

Fig. 2. ROC curves for $Comb_1$ $(k - NN_{(seg,supra)})$ and $Comb_2$ $(GMM_{(seg,supra)})$

Table 3. Optimal cross-combinations

Fusion	$Comb_1$		$Comb_2$		$Comb_3$		$Comb_4$		$Comb_5$		$Comb_6$	
	(K=1,K=11)		(M=12,M=15)		(M=12,K=11)		(K=1,M=15)		(K=11,M=12)		(K=5,M=16)	
λ	0.8	0.9	0.6	0.8	0.9	0.7	0.5	0.6	0.9	0.7	0.5	0.4
AUC	0,813	0,812	0,932	0.922	0,913	0.905	0,846	0.845	0,845	0.849	0,897	0.895
EER (%)	26	29	12	19	13	16	22	22	25	24	16	14

should be careful because the fusion of best configurations does not always give best results since the efficiency will depend on errors produced by the classifiers (independent vs dependent) [9]. Table 1 and section §3.3 show that 6 different fusion schemes can be investigated ($Comb_1$ to $Comb_6$) and for each of them we optimized classifiers (k,M) and weighting λ (eq. 6) parameters. In table 3 and figure 2 (Fig. 2) we can see that for the k-NN classifier, best scores (0,813/0,812) are obtained with an important contribution of the segmental features ($\lambda = 0.8$) which is in agreement with the results obtained without the fusion (table 2). The best GMM results (0,932/0,922) are obtained with a weighting factor equals to 0.6 revealing a balance between the two features. Despite that motherese is characterized by prosody, we showed that is interesting to combine the acoustic and prosodic features, also the fusion of two classifiers increased the performance of detection.

5 Conclusions

we have developed a first motherese detection system tested in a speaker-dependent mode. Using classification techniques that are also often used in speech and speaker recognition (GMM and k-NN) we have developed motherese detection system and we have tested them on mode dependent of speaker. Fusion of features and classifiers were also investigated. We obtained results from which we can draw interesting conclusions. Firstly, our results show that segmental features alone contain much useful information for discrimination between motherese and adult direct speech since they outperform supra segmental feature. Thus, we can conclude that segmental features can be used alone. However according to our detection results, prosodic features are also very promising features. Based on the previous two conclusions, we combined classifiers that use segmental features with classifiers that use supra-segmental features and found that this combination improves the performance of motherese detector considerably. For our motherese classification experiments, we used only segments that were already segmented (based on human transcription). In other words, detection of onset and offset of motherese was not investigated in this study but can be addressed in a follow-up study. Detection of onset and offset of motherese (motherese segmentation) can be seen as a separate problem that gives rise to other interesting questions such as how to define the beginning and end of motherese, and what kind of evaluation measures to use.

References

1. Muratori, F., Maestro, S.: Autism as a downstream effect of primary difficulties in intersubjectivity interacting with abnormal development of brain connectivity. International Journal for Dialogical Science Fall 2(1), 93–118 (2007)
2. Fernald, A., Kuhl, P.: Acoustic determinants of infant preference for Motherese speech. Infant Behavior and Development 10, 279–293 (1987)
3. Kuhl, P.K.: Early language acquisition: Cracking the speech code. Nature Reviews Neuroscience 5, 831–843 (2004)
4. Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V.: The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In: Proceedings of Interspeech, pp. 2253–2256 (2007)
5. Reynolds, D.: Speaker identification and verification using Gaussian mixture speaker models. Speech Communication 17, 91–108 (1995)
6. Duda, R., Hart, P., Stork, D.: Pattern Classification, 2nd edn (2000)
7. Shami, M., Verhelst, W.: An Evaluation of the Robustness of Existing Supervised Machine Learning Approaches to the Classification of Emotions. Speech. Speech Communication 49(3), 201–212 (2007)
8. 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 (October 2007)
9. Kuncheva, I.: Combining pattern classifiers: Methods & algorithms. Wiley, Chichester (2004)
10. `http://www.nist.gov/speech/tools/`