Motherese Detection Based On Segmental and Supra-Segmental Features

Ammar Mahdhaoui, Mohamed Chetouani, Cong Zong
Pierre et Marie Curie - Paris 6 University
ISIR/P&M Laboratory, CNRS FRE 2507
3 rue Galile, 94200 Ivry sur Seine, France
Ammar.Mahdhaoui@robot.jussieu.fr, Mohamed.Chetouani@upmc.fr

Abstract

In this paper, we present an automatic motherese detection system for the study of parent-infant interaction analysis. Motherese is a speech register directed towards infants and it is characterized by higher pitch, slower tempo, and exaggerated intonation. The goal of this paper is to propose and evaluate different approaches for the detection of motherese from home movies. We investigated the characterization by supra-segmental features (prosody) but also by segmental ones namely the MFCC (Mel Frequency Cepstral Coefficients). Concerning the classification stage, we investigated two different methods: the k-nn (k-nearest neighbors) and the GMM (Gaussian Mixture Models). Experimental results show that segmental features play a major role on the detection.

1. Introduction

Parent-infant interaction plays a major role on the development of cognitive, perceptual and motor skills and this role is emphasized for disorder developments. 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, and by a restricted, stereotyped, repetitive repertoire of behaviour, interest and activities (ICD 10; DSM VI)[1].

In this paper, we focus on 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]. 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. Current studies in early signs of autism show that the infants, who will become autistic, are clearly sensitive to the motherese since the interaction (verbal and/or nonverbal) seems to be usually started by motherese. The lack of attention for other kind of registers is mainly due to the social deficits in autism.

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. 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 "normal" directed speech which implies the design of robust motherese detector.

Section 2 presents the longitudinal corpus. The proposed method is described in section 3 which needs specific attentions to the different stages: feature extraction, classification and decision fusion. The experimental protocols, metrics and results are discussed in section 4. Finally, we give conclusions and future plans from the proposed work.

2. Home movies corpus

Home videos are usually collections of natural and spontaneous interactions which is a clear advantage. 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

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

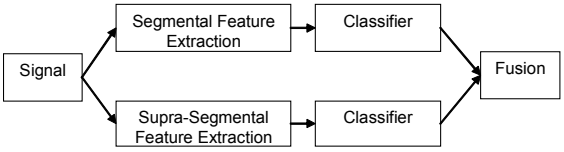

Figure 1. Classification scheme.

conditions (noise, camera, microphones...). We focus on one home video totaling 3 hours during 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.

3. Motherese detection

In this paper, the motherese detection is speaker dependent (mother) and attempts to detect this special register in home videos. Due to the longitudinal corpus, different mismatches occur (noise, camera, microphone) and make the detection difficult. In the literature, motherese is essentially characterized by prosody (supra-segmental) and especially the evolution of the fundamental frequency [2]. In the presence of noisy data, the efficiency of supra-segmental feature extractor is known to be reduced. Consequently, we propose to combine supra-segmental characterization process (feature extraction and classification) with a segmental one as described in Figure 1. The segmental features aims at characterizing the short-time spectral information.

3.1 Feature Extraction

Feature extraction is an important stage and it has been shown that emotional speech can be characterized by a large number of features (acoustic, voice quality, prosodic, phonetic, lexical) [3]. In this paper we focus on 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 on both the fundamental frequency (F0) and the short-time energy. 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) on 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 (cf. §3.3). The feature vectors are normalized (zero mean, unit standard deviation).

3.2 A Posteriori Probabilities Estimation

Two different classifiers are used in this study a statistical based: the Gaussian Mixture Models (GMM) and a distanced based: the k-nearest neighbors (k-nn). A Gaussian mixture density is a weighted sum of M component densities [4] given by:

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

Where $p(x|C_m)$ is the probability density function (PDF) of class C_m evaluated at x. Due to the binary classification task, we define C_1 as the "motherese" class and C_2 as "normal directed speech". x is a d-dimensional vector, $g_{(\mu,\Sigma)}(x)$ are the component densities and ω_i the mixture weights. Each component density is a d-variate Gaussian function:

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

With mean vector μ_i and covariance matrix Σ_i. 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:

$$P_{gmm}(C_m|x) = \frac{p(x|C_m)P(C_m)}{p(x)} \quad (3)$$

where $P(C_m)$ is the prior probability for class C_m, we assume equal prior probabilities. $p(x)$ is the overall PDF evaluated at x

The k-nn classifer [5] 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 [5]:

$$P_{knn}(x|C_m) = \frac{k_m}{k} \quad (4)$$

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

3.3 Segmental and supra-segmental based characterization

3.3.1 Utterance level

Segmental features (i.e. MFCC) are extracted from all the frames of an utterance U_x independently to

the voiced or unvoiced parts. One should note that the nature of the segments can also be exploited (vowels/consonants) [6]. 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_{x_i}) 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.

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

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

3.3.2 Fusion level

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:

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

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 ($P_{gmm,seg}(C_m|U_x)$, $P_{gmm,supra}(C_m|U_x)$)[4]. 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 el al. [8], which consists in using the posteriori probability instead of the log probability of the k-nn:

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

Consequently, for the k-nn classifier we used equation 7 while for the GMM the likelihood is conventionally computed. We investigated cross combinations: $Comb_1$: $P_{knn,seg}$/$P_{knn,supra}$, $Comb_2$: $P_{gmm,seg}$/$P_{gmm,supra}$, $Comb_3$: $P_{knn,seg}$/$P_{gmm,supra}$, $Comb_4$: $P_{gmm,seg}$/$P_{knn,supra}$.

4. Results and discussions

Motherese detection is a binary classification problem and from given confusion matrix we have different decisions: true/false positive (TP, FP) and true/false negative (TN,FN). We evaluated, from a 10 folds cross-validation, the accuracy: $(TP+TN)/(TP+TN+FP+FN)$. We also estimated the Positive Predictive Value $(PPV) : TP/(TP+FP)$ which evaluates the proportion of correctly detected motherese out of all utterances labelled as motherese. PPV can be viewed as the reliability of positive predictions induced by the classifier. Negative Predictive Value (NPV) gives information about the detection of normal speech.

Firstly, we optimized the parameters of the GMM and k-nn classifiers such as the M component densities (eq. 1) and the k neighbors (eq. 4) for both the segmental and supra-segmental features. As a result, the best accuracy rates are obtained for the configurations presented in table 1 using the 10 folds cross-validation. One should note that motherese detection performance depends on both feature extraction and classification schemes, confirming the difficulty for designing emotion recognizers[3]. However, this result motivates an investigation on fusion of both features and classifiers following the statistical approach described in section §3.3.

	segmental	supra-segmental
k-nn	72.5% (k=11)	61% (k=7)
GMM	78% (M=16)	82% (M=16)

Table 1. Optimal configurations

Improvements by the combination of features and classifiers is known to be efficient [9]. However, one should be careful because the fusion of best configurations do not always give better results since the efficiency will depend on errors produced by the classifiers (independent vs dependent) [9]. Table 1 and section §3.3 show that 4 different fusion schemes can be investigated ($Comb_1$ to $Comb_4$) and for each of them we optimized classifiers (k,M) and weighting λ (eq. 6) parameters. In table 2, best results in terms of accuracy but also PPV and NPV for the same classifier since a high accuracy is not enough. As one can see on this table different combinations can reach similar accuracies but with different PPVs. Concerning the k-nn, best scores (74%/77.91%) are obtained with an important contribution of the segmental features ($\lambda = 0.9$) which is in agreement with the results obtained without the fusion (table 1). The best GMM results (87.5%/88.47%) are obtained with a weighting factor equals to 0.5 revealing a balance between the two features.

		Accuracy	PPV	NPV	λ
$Comb_1$	Seg (k=11) Supra (k=1)	74%	77.91%	71.05%	0.9
	Seg (k=7) Supra (k=11)	72.5%	74.42%	71.84%	0.8
$Comb_2$	Seg (M=12) Supra (M=15)	87.5%	88.47%	86.41%	0.4
	Seg (M=16) Supra (M=16)	86.5%	84.85%	88.42%	0.5

Table 2. Optimal combinations

We also investigated cross-classifiers fusion (table 3). We can see that the accuracy can be lower than for the single classifiers (table 2). However significant improvements of the PPV can be reached (90.7%) for the $Comb_3 : P_{knn,seg}/P_{gmm,supra}$ combination ($\lambda = 0.6$). This result shows the importance of evaluation metrics for fusion which is here dependent on the task (motherese detection).

		Accuracy	PPV	NPV	λ
$Comb_3$	Seg (k=11) Supra (M=12)	85.5%	89.41%	83.81%	0.7
	Seg (k=5) Supra (M=12)	85%	90.7%	82.69%	0.6
$Comb_4$	Seg (M=15) Supra (k=11)	80.5%	79.61%	81.44%	0.2
	Seg (M=16) Supra (k=1)	79.5%	77.57%	81.72%	0.3

Table 3. Optimal cross-combinations

5. Conclusion

In this paper, we explored motherese/normal directed speech discrimination from parent-infant verbal interaction extracted from home movies. Within this corpus, a high variability is observed due to the spontaneous interactions, noisy environment (house) and mismatches (longitudinal). We investigated features and classifiers combinations by posteriori probabilities estimation. Concerning the features, we extracted segmental (MFCC) and also supra-segmental features due to the prosodic characteristics of motherese. The classifier combination is based on a statistical weighting taking into account the specificity of each classifier (non-parametric, statistical). The results show that the accuracy is not enough for the estimation of the efficiency of our detector since best results in terms of PPV have been obtained with a lower accuracy. In addition, even if motherese is essentially characterized by prosody, combinations of features and classifiers show that the relevance of segmental features. This result can be justified by the robustness of segmental features compared to supra-segmental ones. Our future works will be devoted to the acquisition of a more important data and also on the application of other evaluation metrics.

References

[1] American Psychiatric Association, "The Diagnostic and Statistical Manual of Mental Disorders, IV", Washington, D.C.: American Psychiatric Association, 1994.

[2] A. Fernald, P. Kuhl, "Acoustic determinants of infant preference for Motherese speech". Infant Behavior and Development, 10, 279-293, 1987.

[3] Schuller, B., et al. "The relevance of feature type for the automatic classification of emotional user states": low level descriptors and functionals. Proceedings of Interspeech, pp. 2253-2256, 2007

[4] Reynolds, D. Speaker identification and verification using Gaussian mixture speaker models, Speech Communication, vol. 17, pp. 91108, 1995.

[5] Duda, R., Hart, P., Stork, D.. "Pattern Classification", second edition, 2000.

[6] Ringeval, F. and Chetouani, M. "Exploiting a Vowel Based Approach for Acted Emotion Recognition". International Workshop on Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction (Springer), 2008.

[7] Shami M., Verhelst, W. "An Evaluation of the Robustness of Existing Supervised Machine Learning Approaches to the Classification of Emotions", Speech. Speech Communication, vol. 49, issue 3, pages 201-212, 2007.

[8] Kim,S., Georgiou, P., Lee, S., and Narayanan, S., "Real-time emotion detection system using speech: Multi-modal fusion of different timescale features". IEEE International Workshop on Multimedia Signal Processing, October 2007.

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