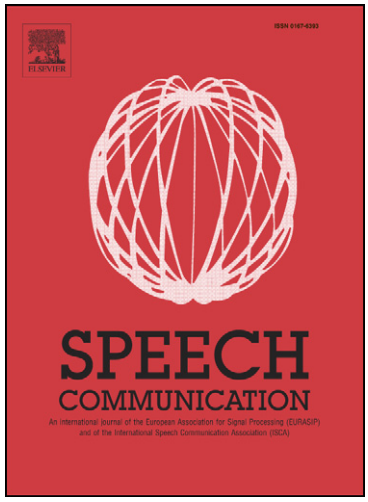
Supervised and semi-supervised infant-directed speech classification for parent-infant interaction analysis

Ammar Mahdhaoui, Mohamed Chetouani



Supervised and semi-supervised infant-directed speech classification for parent-infant interaction analysis

Ammar Mahdhaoui, Mohamed Chetouani

PII: S0167-6393(11)00070-7
DOI: 10.1016/j.specom.2011.05.005
Reference: SPECOM 1994

To appear in: *Speech Communication*

Received Date: 1 May 2010
Revised Date: 29 April 2011
Accepted Date: 4 May 2011

Please cite this article as: Mahdhaoui, A., Chetouani, M., Supervised and semi-supervised infant-directed speech classification for parent-infant interaction analysis, *Speech Communication* (2011), doi: 10.1016/j.specom.2011.05.005

Supervised and semi-supervised infant-directed speech classification for parent-infant interaction analysis

Ammar Mahdhaoui and Mohamed Chetouani
Univ Paris 06, F-75005, Paris, France CNRS, UMR 7222
ISIR, Institut des Systèmes Intelligents et de Robotique, F-75005, Paris, France

Abstract:

This paper describes the development of an infant-directed speech discrimination system for parent-infant interaction analysis. Different feature sets for emotion recognition were investigated using two classification techniques: supervised and semi-supervised. The classification experiments were carried out with short pre-segmented adult-directed speech and infant-directed speech segments extracted from real-life family home movies (with durations typically between 0.5 s and 4 s). The experimental results show that in the case of supervised learning, spectral features play a major role in the infant-directed speech discrimination. However, a major difficulty of using natural corpora is that the annotation process is time-consuming, and the expression of emotion is much more complex than in acted speech. Furthermore, interlabeler agreement and annotation label confidences are important issues to address. To overcome these problems, we propose a new semi-supervised approach based on the standard co-training algorithm exploiting labelled and unlabelled data. It offers a framework to take advantage of supervised classifiers trained by different features. The proposed dynamic weighted co-training approach combines various features and classifiers usually used in emotion recognition in order to learn from different views. Our experiments demonstrate the validity and effectiveness of this method for a real-life corpus such as home movies.

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 00 (2011) 1–15
Speech Communication

Supervised and semi-supervised infant-directed speech classification for parent-infant interaction analysis

Ammar Mahdhaoui*, Mohamed Chetouani

Univ Paris 06, F-75005, Paris, France CNRS, UMR 7222
ISIR, Institut des Systèmes Intelligents et de Robotique, F-75005, Paris, France

Abstract

This paper describes the development of an infant-directed speech discrimination system for parent-infant interaction analysis. Different feature sets for emotion recognition were investigated using two classification techniques: supervised and semi-supervised. The classification experiments were carried out with short pre-segmented adult-directed speech and infant-directed speech segments extracted from real-life family home movies (with durations typically between 0.5 s and 4 s). The experimental results show that in the case of supervised learning, spectral features play a major role in the infant-directed speech discrimination. However, a major difficulty of using natural corpora is that the annotation process is time-consuming, and the expression of emotion is much more complex than in acted speech. Furthermore, interlabeler agreement and annotation label confidences are important issues to address. To overcome these problems, we propose a new semi-supervised approach based on the standard co-training algorithm exploiting labelled and unlabelled data. It offers a framework to take advantage of supervised classifiers trained by different features. The proposed dynamic weighted co-training approach combines various features and classifiers usually used in emotion recognition in order to learn from different views. Our experiments demonstrate the validity and effectiveness of this method for a real-life corpus such as home movies.

Keywords: Infant-directed speech, emotion recognition, face-to-face interaction, data fusion, semi-supervised learning

1. Introduction

Parent-infant interactions play a major role in the development of the cognitive, perceptual and motor skills of infants, and this role is emphasised for developmental disorders. Typically developing infants gaze at people, turn toward voices and express interest in communication. In contrast, infants who will later become autistic are characterised by the presence of abnormalities in reciprocal social interactions and by a restricted, stereotyped and repetitive repertoire of behaviours, interests and activities (autism pathology is defined by ICD 10: International classification of diseases and related health problems[1] and DSM IV: Diagnostic and statistical manual of mental disorders[2])[1]. The quality of parent-infant interaction depends on a reciprocal process, an active dialogue between parent and child based on the infant's early competencies and the mother's (or father's) stimulations. In addition, the infant's development depends on social interaction with a caregiver who serves the infant's needs for emotional attachment.

Researchers in language acquisition and researchers

*Corresponding author
Email addresses: Ammar.Mahdhaoui@isir.upmc.fr (Ammar Mahdhaoui), Mohamed.Chetouani@upmc.fr (Mohamed Chetouani)
URL: people.isir.upmc.fr/mahdhaoui (Ammar Mahdhaoui)

[1] http://www.who.int/classifications/icd/en/
[2] http://www.psych.org/mainmenu/research/dsmiv.aspx

in early social interactions have identified an important peculiarity that affects both the language and social development of infants; i.e., the way adults speak to infants. The special kind of speech that is directed towards infants, called infant-directed speech or "motherese" is a simplified language/dialect/register [2] that has recently been shown to be crucial for engaging interactions between parents and infant and very important for language acquisition [3]. Moreover, this speech register has been shown to be preferred by infants over adult-directed speech [4] and might assist infants in learning speech sounds [5]. From an acoustic point of view, infant-directed speech has a clear signature (high pitch, exaggerated intonation contours) [5] [6]. The phonemes, and especially the vowels, are more clearly articulated [7].

The importance of infant-directed speech has also been highlighted by recent research on autism [8] [9] [10]. Manual investigations (i.e., manual annotations) [11], of parent-infant interactions in home movies have shown that most positive sequences (i.e., multimodal responses of the infant: vocalisation, gaze, facial expression) were induced by infant-directed speech. To study more specifically the influence on engagement in an ecological environment, we followed a method usually employed for the study of infant development: home movie analysis [12].

The study of home movies is very important for future research, but the use of this kind of database makes the work very difficult and time-consuming. The manual annotation of these films is very costly, and the automatic detection of relevant events would be of great benefit to longitudinal studies. For the analysis of the role of infant-directed speech during interaction, we developed an automatic infant-directed speech detection system [10] [13] [14], to enable emotion classification.

Motherese or infant-directed speech has been highly studied by psychological community. However, in our knowledge there are no studies of infant-directed speech, in real-life interaction, employing machine learning techniques. In the literature, researchers in affective computing and in emotion recognition have studied infant-directed speech from acted databases [15]; the speech samples were recorded in laboratory. Recently, Inoue et al. [16] have developed a novel approach to discriminate between infant-directed speech and adult-directed speech by using mel-frequency cepstrum coefficient and a hidden Markov model-based speech discrimination algorithm. The average discrimination accuracy of the proposed algorithm is 84.34%, but still in laboratory conditions (acted data). Paralinguistic characteristics of motherese motivate several researchers to employ recognition systems intially developped for emotion processing [15][17].

In this paper, we implemented a traditional supervised method. We tested different machine learning techniques, both statistical and parametric, with different feature extraction methods (time/frequency domains). The GMM classifier with cepstral MFCC (Mel-frequency cepstral coding) features was found to be most efficient.

However, the supervised methods still have some significant limitations. Large amounts of labelled data are usually required, which is difficult in real-life applications; manual annotation of data are very costly and time consuming. Therefore, we investigate a semi-supervised approach that does not require a large amount of annotated data for training. This method combines labelled and unlabelled utterances to learn to discriminate between infant-directed speech and adult-directed speech.

In the area of classification, many semi-supervised learning algorithms have been proposed, one of which is the co-training approach [18]. Most applications of co-training algorithm have been devoted to text classification [19] [20] and web page categorisation [18] [21]. However, there are a few studies related to semi-supervised learning for emotional speech recognition. The co-training algorithm proposed by Blum and Mitchell [18] is a prominent achievement in semi-supervised learning. It initially defines two classifiers on distinct attribute views of a small set of labelled data. Either of the views is required to be conditionally independent to the other and sufficient for learning a classification system. Then, iteratively the predictions of each classifier on unlabelled examples are selected to increase the training data set. This co-training algorithm and its variations [22] have been applied in many areas because of their theoretical justifications and experimental success.

In this study, we propose a semi-supervised algorithm based on multi-view characterisation, which combines the classification results of different views to obtain a single estimate for each observation. The proposed algorithm is a novel form of co-training, which is more suitable for problems involving both classification and data fusion. Algorithmically, the proposed co-training algorithm is quite similar to other co-training methods available in the literature. However, a number of novel improvements, using different feature sets and dynamic weighting classifier fusion, have been incorporated to make the proposed algorithm more suitable for multi-view classification problems.

The paper is organised as follows. Section 2 presents

the longitudinal speech corpus. Section 3 presents the different feature extraction methods. Sections 4 and 5 present the supervised and the semi-supervised methods. Section 6 presents the details of the proposed method of semi-supervised classification of emotional speech with multi-view features. Section 7 reports experimental comparisons of supervised and semi-supervised methods on a discrimination task. In the last section, some concluding remarks and the direction for future works are presented.

2. Home movie: speech corpus

The speech corpus used in our study contains real parent/child interactions and consists of recordings of Italian mothers as they addressed their infants. It is a collection of natural and spontaneous interactions. This corpus contains expressions of non-linguistic communication (affective intent) conveyed by a parent to a preverbal child.

We decided to focus on the analysis of home movies (real-life data) as it enables longitudinal study (months or years) and gives information about the early behaviours of autistic infants long before the diagnosis was made by clinicians. However, this large corpus makes it inconvenient for people to review. Additionally, the recordings were not made by professionals (they were made by parents), resulting in adverse conditions (noise and camera and microphones limitations, etc). In addition, the recordings were made randomly in diverse conditions and situations (interaction situation, dinner, birthday, bath, etc.), and only parents and other family members (e.g., grand-parent, uncle) are present during the recordings.

All sequences were extracted from the Pisa home movies database, which includes home movies from the first 18 months of life for three groups of children (typically developing, autistic, mentally retarded) [23].

The home movies were recorded by the parents themselves. Each family uses his personal camera with only one microphone. Due to the naturalness of home movies (uncontrolled conditions: TV, many speakers, etc.), we manually selected a set of videos with at least understandable audio data. The verbal interactions of the infant's mother were carefully annotated by two psycholinguists, independently, into two categories: infant-directed speech and adult-directed speech. To estimate the agreement between the two annotators, we computed the Cohen's kappa [24] as a measure of the intercoder agreement. Cohen's kappa agreement is given by the following equation:

$$kappa = \frac{p(a) - p(e)}{1 - p(e)} \quad (1)$$

where $p(a)$ is the observed probability of agreement between two annotators, and $p(e)$ is the theoretical probability of chance agreement, using the annotated sample of data to calculate the probabilities of each annotator. We found a Cohen's kappa equal to 0.82 (CI for Confidence Interval: [95%CI: 0.75-0.90]), measured on 500 samples, which corresponds to good agreement between the two annotators.

From this manual annotation, we randomly extracted 250 utterances for each category. The utterances are typically between 0.5 s and 4 s in length. Figure 1 shows a distribution of infant-directed speech and adult-directed speech utterances from 3 periods of the child's life (0-6 months, 6-12 months and 12-18 months). The total duration of utterances is about 15 minutes. Figure 2 shows the duration distribution of infant-directed speech and adult-directed speech utterances. It shows that there is no significant difference between the durations of infant-directed speech and adult-directed speech utterances.

We randomly divided the database into two parts: unlabelled data U (400 utterances balanced between motherese and adult-directed speech) and labelled data L (100 utterances balanced between motherese and adult-directed speech).

3. Emotional Speech Characterisation

Feature extraction is an important stage in emotion recognition, and it has been shown that emotional speech can be characterised by a large number of features (acoustics, voice quality, prosodic, phonetic, lexical) [25]. However, researchers on speech characterisation and feature extraction show that is difficult to have a consensus for emotional speech characterisation.

In this study, we computed temporal and frequential features, which are usually investigated in emotion recognition [26] [17]. Moreover, different statistics are applied, resulting in 16 cepstral ($f1$), 70 prosodic ($f2$, $f3$, $f4$ and $f5$) and 96 perceptive features ($f6$, $f7$, $f8$ and $f9$), all of which have been shown to be the most efficient [26] [27] [13]. We obtained 9 different feature vectors with different dimensions, which are presented in Table 1.

3.1. Cepstral features

Cepstral features such as MFCC are often successfully used in speech and emotion recognition. The

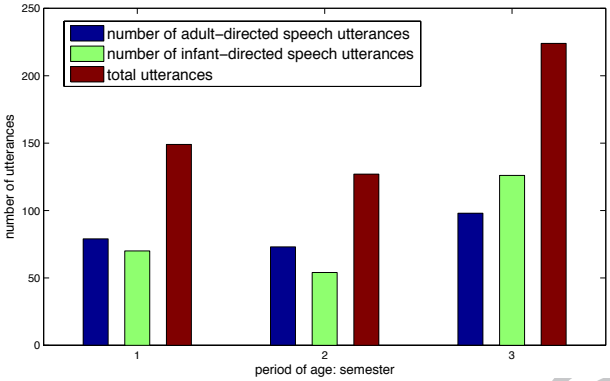


Figure 1: Distribution of infant-directed speech and adult-directed speech utterances during 3 periods of infant development

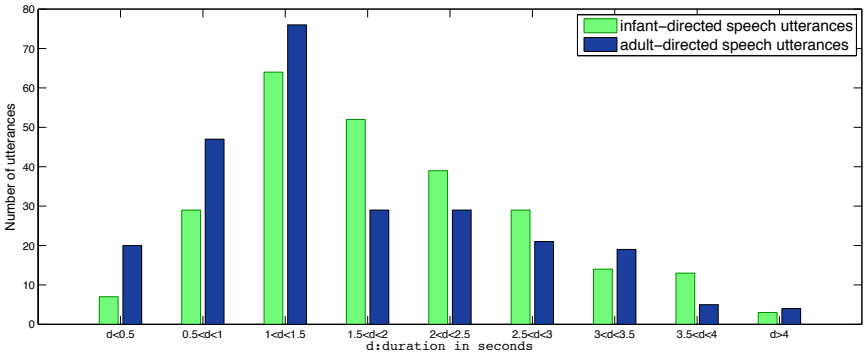


Figure 2: Duration distribution of infant-directed speech and adult-directed speech utterances

Table 1: Different features sets

$f1$	16 MFCCs
$f2$	Pitch (Min, Max, Range) + Energy (Min, Max, Range)
$f3$	35 statistics on the pitch
$f4$	35 statistics on the energy
$f5$	35 statistics on the pitch + 35 statistics on the energy
$f6$	Bark TL + SL + MV (96 statistics)
$f7$	Bark TL (32 statistics)
$f8$	Bark SL (32 statistics)
$f9$	Bark MV (32 statistics)

short-term cepstral signatures of both infant-directed speech and adult-directed speech are characterised by 16 MFCC features (often used for emotion recognition) and are extracted each 20 ms, so the number of the resulting feature vectors is variable and depends on the length of the utterance (Frame-level).

3.2. Prosodic features

Several studies have shown the relevance of both the fundamental frequency (F0) and energy features for emotion recognition applications [26]. F0 and energy were estimated every 20 ms [28], and we computed 3 statistics for each voiced segment (segment-based method) [29]: the mean, variance and range, for both F0 and short-time energy, resulting in a 6-dimensional vector.

In addition, 32 statistical features, presented in Table 2, are extracted from the pitch contour and the loudness contour. Three other features are also extracted from these contours with a histogram and by considering the maximum, the bin index of the maximum and the centre value of the corresponding bin. These 3 features are relevant for pitch and energy contour characterisation.

3.3. Perceptive features

Infant-directed speech and adult-directed speech sound perceptually different, [30], and in this work bark filters spectral representation are employed to investigate these perceptual differences.

The features based on the bark scale are considered to provide more information by characterising the hu-

Table 2: 32 statistics

Maximum, minimum and mean value
Standard deviation
Variance
Skewness
Kurtosis
Interquartile range
Mean absolute deviation (MAD)
MAD based on medians, i.e. MEDIAN(ABS(X-MEDIAN(X)))
First and second coeficients of linear regression,
First, second and third coefficients of quadratic regression
9 quantiles corresponding to the following cumulative probability values: 0.025, 0.125, 0.25, 0.375, 0.50, 0.625, 0.75, 0.875, 0.975
Quantile for cumulative probability values 1% and 9% and interquantile range between this two values
Absolute and sign of time interval between maximum and minimum appearances

man auditory system [31] [32]. We extracted the bark time/frequency representation using an analysis window duration of 15 ms and a time step of 5 ms with filters equally spaced by 1 bark (first filter centred on first bark critical band) [33]. This computation on the full spectrum results in 29 filter bands. This representation can be described as a discrete perceptive representation of the energy spectrum, which can be qualified as a perceptive spectrogram. We then extracted statistical features from this representation either along the time axis or along the frequency axis, as shown in Figure 3. We also considered the average of energy of the bands (a perceptive Long Term Average Spectrum) and extracted statistical features from it. Thirty-two statistical features were used and applied a) along the time axis (Approach TL), b) along the frequency axes (Approach SL) and c) on the average perceptive spectrum to obtain a first set of 32 features (Approach MV).

a) Approach TL (for 'Time Line') Figure 3. a.: (step 1) extracting 32 features on the spectral vector of each time frame, then (step 2) averaging the values for each of 32 features along the time axis to obtain a second set of 32 features.
b) Approach SL ('for Spectral Line') Figure 3. b.: (step 1) extracting 32 features along the time axis for each spectral band and (step 2) averaging the 32 features along the frequency axis to obtain a third set of 32 features.
c) Approach MV (for 'Mean Values'): (step 1) averaging the energy values of the bark spectral bands along the time axis to obtain a long term average spectrum using 29 bark bands and (step 2) extracting the 32 statistical features from this average spectrum.

The 32 statistical features, presented in Table 2, were computed to model the dynamic variations of the bark spectral perceptive representation.

4. Supervised Classification

The supervised classification assumes that there is already an existing categorisation of the data. In this classification form, the training data D are presented by an ensemble X of feature vectors and their corresponding labels Y:

$$D = \{(x_i, y_i)|x \in X, y \in Y\}_{i=1}^{n} \quad (2)$$

Supervised classification consists of two steps: feature extraction and pattern classification. The features extraction step consists of characterising the data. After the extraction of features, supervised classification is used to categorise the data into classes corresponding to user-defined training classes. This can be implemented using standard machine learning methods. In this study, four different classifiers, Gaussian mixture models (GMM) [34], k-nearest neighbour (k-NN) [35] classifiers), SVM [36] [37] and Neural networks (MLP) [38], were investigated.

In our work, all the classifiers were adapted to provide a posterior probability to maintain a statistical classification framework.

4.1. Gaussian mixture models

A Gaussian Mixture Model is a statistics based model for modelling a statistical distribution of Gaussian Probability Density Function (PDF). A Gaussian mixture density is a weighted sum of M component densities [34] given by:

$$p(x|C_m) = \sum_{i=1}^{M} \omega_i g_{(\mu_i,\Sigma_i)}(x) \quad (3)$$

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

$$g_{(\mu,\Sigma)}(x) = \frac{1}{(2\pi)^{d/2}\sqrt{det(\Sigma)}} e^{-1/2(x-\mu)^T\Sigma^{-1}((x-\mu))} \quad (4)$$

with mean vector μ_i and covariance matrix Σ_i. The mixture weights ω_i satisfy the following constraint:

$$\sum_{i=1}^{M} \omega_i = 1 \quad (5)$$

The feature vector x is then modelled by the following posterior probability:

$$P_{gmm}(C_m|x) = \frac{p(x|C_m)P(C_m)}{p(x)} \quad (6)$$

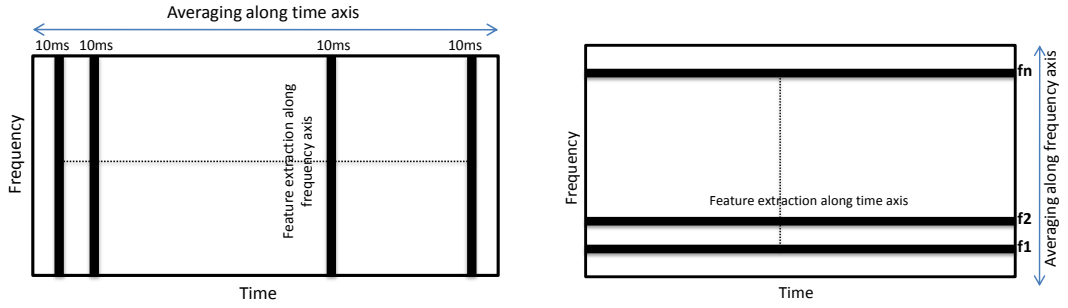


Figure 3: Method for Extraction of bark-based features: along time axis (a) and along frequency axis (b)

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

4.2. k-nearest neighbours

The k-NN classifier [35] is a non-parametric technique that classifies the input vector with the label of the majority of the k-nearest neighbours (prototypes). To maintain a common framework with the statistical classifiers, we estimate the posterior probability that a given feature vector x belongs to class C_m using k-NN estimation [35]:

$$P_{knn}(C_m|x) = \frac{k_m}{k} \tag{7}$$

where k_m denotes the number of prototypes that belong to the class C_m among the k nearest neighbours.

4.3. Support vector machines

The support vector machine (SVM) is the optimal margin linear discriminant trained from a sample of l independent and identically distributed instances: $(x_1, y_1), \cdots, (x_l, y_l)$, where x_i is the d-dimensional input and $y_i \in \{-1, +1\}$ its label in a two-class problem is $y_i = +1$ if is a positive (+) example, and $y_i = -1$ if x_i is a negative example.
The basic idea behind SVM is to solve the following model:

$$min \frac{1}{2}\|\omega\|^2\omega + C\sum_{i=1}^{l}\xi_i \tag{8}$$

$$\forall i, y_i\,(\omega x_i + b) \geq 1 - \xi_i \tag{9}$$

which is a C-soft margin algorithm where ω and b are the weight coefficients and bias term of the separating hyperplane, C is a predefined positive real number and ξ_i are slack variables [39]. The first term of the objective function given in (8) ensures the regularisation by minimising the norm of the weight coefficients. The second term tries to minimise the classification errors by introducing slack variables to allow some classification errors and then minimising them. The constraint given in (9) is the separation inequality, which tries to locate each instance on the correct side of the separating hyperplane. Once ω and b are optimised, during the test, the discrimination is used to estimate the labels:

$$\hat{y} = sign(\omega x + b) \tag{10}$$

and we choose the positive class if $\hat{y} = +1$ and the negative class if $\hat{y} = -1$. This model is generalised to learn nonlinear discriminants with kernel functions to map x to a new space and learning a linear discriminant there.

The standard SVM does not provide posterior probabilities. However, to maintain a common framework with other classifiers, the output of a classifier (SVM) should be a posterior probability to enable post-processing. Consequently, to map the SVM outputs into probabilities, as presented in [40], we must first train an SVM, and then train the parameters of an additional sigmoid function. In our work, we used LIBSVM [36] with posterior probabilities outputs $P_{svm}(C_m|x)$.

4.4. Neural network

The Neural Network structure used in this paper was the Multilayer Perceptron (MLP). An MLP is a network of simple neurons called perceptrons. The perceptron computes a single output from multiple real-valued inputs by forming a linear combination according to its input weights and then possibly transforming the output

by some nonlinear activation function. Mathematically this can be written as:

$$y = \varphi(\sum_{i=1}^{n} \omega_i x_i + b) = \varphi(w^T x + b) \quad (11)$$

where w denotes the vector of weights, x is the vector of inputs, b is the bias and φ is the activation function.

It is proved in [41] that for various parameter optimisation strategies (such as gradient descent) with minimisation of the Mean Square Error function or Cross-Entropy Error function and the back-propagation technique used to compute derivatives of the error function with respect to each of the free parameters, the trained network estimates the posterior probabilities of class membership $P_{mlp}(C_m|x)$ directly.

5. Semi-Supervised Classification

Supervised methods require a large number of labelled utterances to enable efficient learning in real emotional speech classification systems. However, the manual annotation of data is very costly and time consuming, so an extensive manual annotation of all the home movies is unrealistic. Therefore, a learning algorithm with only a few labelled data is required; i.e., a semi-supervised learning algorithm. In this section, we briefly describe two techniques for semi-supervised learning, namely, self-training and co-training. Self-training and co-training algorithms allow a classifier to start with a few labelled examples to produce an initial weak classifier and later to combine labelled and unlabelled data to improve the performance. In the following, let us assume that we have a set L (usually small) of labelled data, and a set U (usually large) of unlabelled data.

5.1. Self-training

The definition of self-training can be found in different forms in the literature; however, we adopted the definition of Nigam and Ghani [42]. In this method, we need only one classifier and then only one feature set. For several iterations, the classifier labels the unlabelled data and converts the most confidently predicted examples of each class into a labelled training example.

Table 3 shows the pseudo-code for a typical self-training algorithm. The self-training starts with a set of labelled data L, and builds a classifier h, which is then applied to the set of unlabelled data U. Only the n best classified utterances are added to the labelled set. The classifier is then retrained on the new set of labelled examples, and the process continues for several iterations.

Table 3: Self-training algorithm

Given:
- a set L of Labelled examples
- a set U of Unlabelled examples
- a number n of examples to be added to L in each iteraction

Loop:
- Use L to train the classifier h
- Allow h to label U
- Let T be the n examples in U on which h makes the most confident predictions
- Add T to L
- Remove T from U

End

Table 4: Co-Training algorithm

Given:
- a set L of Labelled examples
- a set U of Unlabelled examples

Loop:
- Use L to train each classifier h_1
- Use L to train each classifier h_2
- Allow h_1 to label p_1 positive and n_1 negative examples from U
- Allow h_2 to label p_2 positive and n_2 negative examples from U
- Add these self-labelled examples to L
- Remove these self-labelled examples from U

End

Notice that only one classifier is required, with no split of the features.

5.2. Co-training

The co-training algorithm proposed in [18] is a prominent achievement in semi-supervised learning. This algorithm and the related multi-view learning methods [43] assume that various classifiers are trained over multiple feature views of the same labelled examples. These classifiers are encouraged to make the same prediction on any unlabelled example.

As shown in Table 4, the method initially defines two classifiers (h_1 and h_2) on distinct attribute views of a small set of labelled data (L). Either of the views is required to be conditionally independent of the other and sufficient for learning a classification system. Then, iteratively, each classifier's predictions on the unlabelled examples are selected to increase the training data set. For each classifier, the unlabelled examples classified with the highest confidence are added to the labelled data set L, so that the two classifiers can contribute to increase the data set L. Both classifiers are re-trained on this augmented data set, and the process is repeated a given number of times. The rationale behind co-training is that one given classifier may assign correct labels to certain examples, while it may be difficult for others to do so. Therefore, each classifier can increase the training set by adding examples that are very informative for the other classifier.

This method can be generalised to be used with a large number of views. Figure 4 shows the general architecture of a generalised co-training method based on multi-view characterisation. It considers v different views. For each iteration, we select an ensemble of p_i positive examples and n_i negative examples that are classified with the highest confidence. Then, we add the ensemble $T = \sum_{i=1}^{v} p_i + n_i$ to the labelled data set L.

These semi-supervised algorithms and their variations [22] have been applied in many application areas because of their theoretical justifications and experimental success.

6. Co-Training Algorithm Based On Multi View Characterisation

Many researchers have shown that multiple-view algorithms are superior to single-view method in solving machine learning problems [18] [44] [45]. Different feature sets and classifiers (views) can be employed to characterize speech signals, and each of them may yield different prediction results. Therefore, the best solution is to use multiple-characterisation (views= feature + classifier) together to predict the common class variable. Thus, the generalised co-training algorithm shown in Figure 4 uses different views for classification. In the multi-view approach, the labelled data are represented by $\{(x_1^1, ..., x_1^v, y_1), ..., (x_m^1, ..., x_m^v, y_m)\}$, where v is the number of views and y_i are the corresponding labels, m is the number of labels.

However, the standard co-training algorithm does not allow the fusion of different views in the same framework to produce only one prediction per utterance. It takes the prediction of each classifier separately. To overcome this problem, we propose a co-training procedure that iteratively trains a base classifier within each view and then combines the classification results to obtain a single estimate for each observation. The proposed algorithm is a novel form of co-training, which is more suitable for problems involving both semi-supervised classification and data fusion.

The goal of the proposed co-training method is to incorporate all the information available from the different views to accurately predict the class variable. Each group of features provides its own perspective, and the performance improvements are obtained through the synergy between the different views. The co-training framework is based on the cooperation of different classifiers for the improvement of classification rates. Each of them gives an individual prediction weighted by its classification confidence value. This problem has a strong similarity to data fusion, which involves incorporating several disparate groups of views into a common framework for modelling data.

This algorithm is designed to improve the performance of a learning machine with a few labelled utterances and a large number of cheap unlabelled utterances.

Given a set L of labelled utterances, a set U of unlabelled utterances, and a set of different feature views V_i, the algorithm works as described in Table 5 and Figure 5. First, to initialise the algorithm, we found the best feature set for each classifier, as presented in Table 7. Second, we set all of the initial weights equally so that $\omega_k = 1/v$, where v is the number of views (9 in our case). Third, while the unlabelled database U is not empty, we repeat the following:

- Classification: to classify all the unlabelled utterances, the class of each utterance is obtained using a decision function. In our case we compute the maximum likelihood; otherwise we can use other decision functions.

- Update the labelled and unlabelled databases: first we take as U_1 the utterances from U classified on Class 1 and U_2 classified on Class 2, after that we calculate the classification confidence for each utterance that we called *margin*. This step consists of cooperating all the classifiers to have once prediction by combining the classifiers outputs using a simple weighted sum.

$$p(C_j|z_i) = \frac{\sum_{k=1}^{v} \omega_k \times h_k(C_j|z_i^k)}{\sum_{k=1}^{v} \omega_k} \tag{12}$$

$$margin_j = \frac{\sum_{1}^{n_j} p(C_j|z_i)}{n_j} \tag{13}$$

  where z_i^k is the feature view to be classified on the class C_j, ω_k is the weight of the classifier h_k, v is the number of views and n_j is the number of segments classified on class C_j. The *margin* value is in the interval [0,1]. This number can be interpreted as a measure of confidence, as is done for SVM [46]. Then we take T_j to be the utterances from U_j that were classified on $Class_j$ with a probability greater than the mean value of classification confidence (*margin*) of the $Class_j$.

- Update weights: finally, we update the weights of each view, as described in Table 5. The new weight

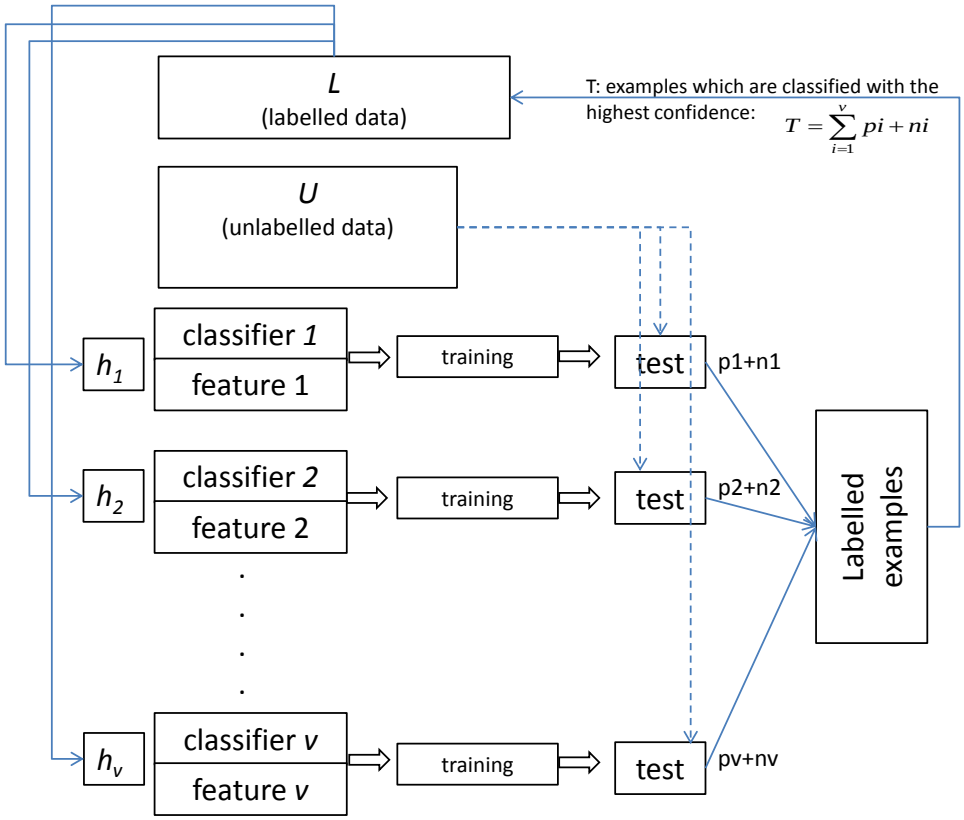


Figure 4: Standard existing co-training algorithm based on multi-view characterization

of each classifier is proportional to its contribution to the final classification. In other words, the weights of efficient classifiers will be increased.

7. Experimental Results

7.1. Experimental setup

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). For supervised classification, we evaluated, from a 10 folds cross validation, the accuracy rate to compare the performances of different separate classifiers: (TP+TN)/(TP+TN+FP+FN). We optimized the parameters of the different classifiers; such as M component densities for GMM, k optimal number of neighbours for k-NN, optimal kernel for SVM and the number of cells for MLP.

For the semi-supervised classification, the performance of the classification system is given for different data set. First we randomly selected an ensemble U containing 400 examples and an ensemble L containing 100 examples balanced between motherese and adult-directed speech. Then, in order to study the implication of the quantity of supervised learning data, we perform several experiments with different number of labelled data; from 10% (10 examples) to 100% (100 examples).

Notice that for the standard co-training algorithm, first we compute the standard algorithm with only two classifiers (the two best classifiers) such as proposed in [18] (Table 4), then we perform this algorithm using all the classifiers as shown in the Figure 4.

The supervised and semi-supervised classification systems were performed on multi-speaker data (speaker-independent). The speech segments were randomly extracted from 10 home movies (10 different mothers). In addition, as shown in Figure 1, the speech segments were extracted from three different periods of time (semester 1, semester 2, semester 3), which will augment the data diversity since the voice of the mothers changes from one semester to another.

Table 5: The proposed Co-Training algorithm

Given:
a set L of m Labelled examples $\{(l_1^1, ..., l_1^v, y_1), ..., (l_m^1, ..., l_m^v, y_m)\}$ with labels $y_i = \{1, 2\}$
a set U of n Unlabelled examples $\{(x_1^1, ..., x_1^v), ..., (x_n^1, ..., x_n^v)\}$
v = number of view (classifier)

Initialization:
ω_k (weights of classifier)= $1/v$ for all the view

While U not empty

A. Classify all the example of the test database:
Do for $k = 1, 2, ..., v$
1. Use L to train each classifier h_k
2. Classify all examples of U by each h_k
3. Calculate the probability of classification for each example x_i from U,
$p(C_j|x_i) = \sum_{k=1}^{v} \omega_k \times h_k(C_j|x_i^k)$
4. $Labels(x_i) = argmax(p(C_j|x_i))$

End for

B. Update the training (L) and test (U) databases:
$U_j = \{z_1, ..., z_{nj}\}$ the ensemble of example classified C_j
Do for $i = 1, 2, ..., nj$
$p(C_j|z_i) = \frac{\sum_{k=1}^{v} \omega_k \times h_k(C_j|z_i^k)}{\sum_{k=1}^{v} \omega_k}$
End for
$margin_j = \frac{\sum_1^{n_j} p(C_j|z_i)}{n_j}$
Take T_j from U_j the examples which has classified on C_j with a probability upper to $margin_j$.
$T = \sum T_j$
Add T to L and remove it from U

C. Update weights: $\omega_k = \frac{\sum_{i=1}^{size(T)} h_k(z_i^k)}{\sum_{k=1}^{v} \sum_{i=1}^{size(T)} h_k(z_i^k)}$

End While

Table 6: Accuracy of separate classifier using 10 folds cross validation

	Feature set	GMM	k-NN	SVM	MLP
Cepstral feature	$f1$	**72.8**	57.7	59.4	61.4
Prosodic features	$f2$	**59.5**	55.7	54.7	50.2
	$f3$	54.7	**55.0**	50.0	50.0
	$f4$	67.0	**68.5**	65.5	58.5
	$f5$	62.1	65.5	**65.5**	54.5
Perceptive features	$f6$	**61.0**	50.5	49.0	54.5
	$f7$	55.5	51.0	52.0	**58.5**
	$f8$	**65.0**	52.0	50.5	55.5
	$f9$	58.8	50.5	50.5	**64.0**

7.2. Results of supervised classifiers

The performance of the different classifiers, each trained with different feature sets ($f1$, $f2$, ... , $f9$), were evaluated on the home movies database.

Table 6 shows the best results of all the classifiers trained with different feature sets. The best result was obtained with GMM trained with cepstral MFCC (72.8% accuracy), and second best result was obtained with k-NN trained with $f4$ (35 statistics on energy). Therefore, Table 6 shows that cepstral MFCC outperforms the other features. Regarding the prosodic features, best results are not obtained with a GMM classifier but with k-NN and SVM classifiers. In addition to the GMM, perceptive features provide satisfactory results using the MLP classifier.

To summarise, comparing the results of different feature sets and taking into account the different classifiers, the best performing feature set for infant-directed speech discrimination appears to be the cepstral MFCC. Regarding the classifiers, we can observe that GMMs generalise better over different test cases than the other classifiers do.

7.3. Results of semi-supervised classifiers

The algorithm works as described in Figure 5. To initialise the co-training algorithm, we consider the best configuration of each features trained with all supervised classifiers, using 10 folds cross validation. We obtained 9 classifiers (views) *h1* to *h9* as described in Table 7.

The classification accuracy of the co-training algorithm using multi-view feature sets with different number of annotations is presented in Figure 6 and Table

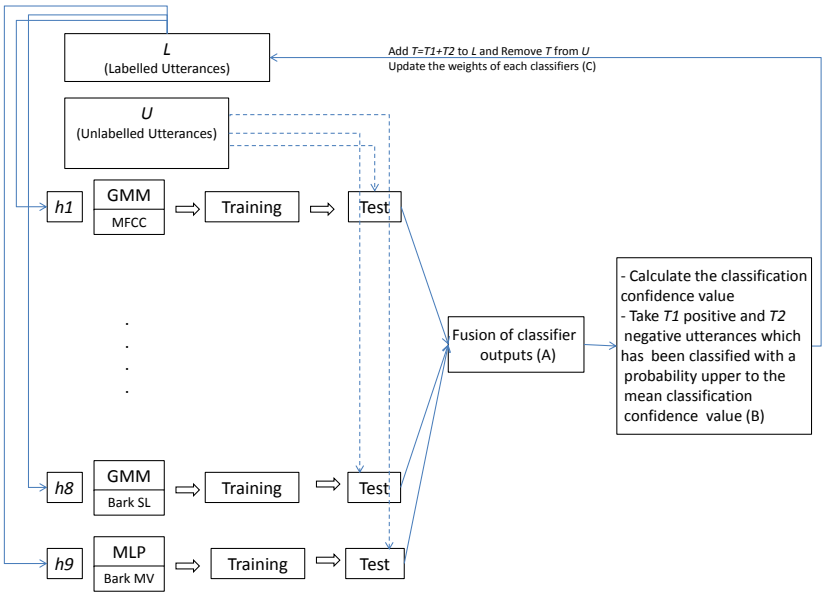


Figure 5: Structure of the proposed Co-training algorithm

8. It can be seen that our method can achieve efficient results in infant-directed speech discrimination.

To further illustrate the advantage of the proposed method, Table 8 and Figure 6 show a direct comparison between our co-training algorithm and the standard co-training algorithm. It shows that our method performs better results, 75.8% vs. 71.5%, using 100 labelled utterances. In addition, Figure 6 and Table 8 show that the performance of the standard co-training algorithm that uses all the classifiers is worse than the performance of the algorithm using only two classifiers, especially when we dispose of few labelled data for training. Although, the standard co-training algorithm was shown promising for different classification problems, it suffers from issues of divergence, where errors in the newly classified data could cause the system to run off track [47]. One approach to overcome this problem is combining the different predictions given by the different classifiers; such as all the classifiers cooperate to obtain only one prediction per utterance. The proposed co-training algorithm offers a framework to take advantage of co-training learning and data fusion. It combines the various features and classifiers in a co-training framework.

In addition, to illustrate the advantage of the proposed multi-view method, especially in cases with very few annotations, we compare our method with the self-training method with a single view. In our study, we investigated the basic self-training algorithm, which replaces multiple classifiers in the co-training procedure with the best classifier that employs the most efficient feature. We computed GMM with the cepstral MFCC (h_1) and prosodic features (h_2), and at each iteration we take only the utterance with the best posterior probability.

Figure 6 and Table 8 show a comparison between our co-training method and the self-training method. It can be seen that our method outperforms the self-training method, 75.8% vs. 70.3%, with 100 labelled utterances. In addition, the proposed co-training method gives a satisfactory result in the case of very few annotations, 66.8% with 10 labelled utterances vs. 52.0% for the self-training method. Comparing self-training and supervised methods, Figure 6 shows that supervised algorithm outperforms self-training algorithm since that self-training method suffers from issues of divergence (hight risk of divergence) [47]. The self-training algorithm makes error in the first iteration, therefore the error rate becomes important, and then the classifier will learn on falsely classified examples. The risk of divergence is the major problem of the self-training algorithm [47].

In addition, to illustrate the importance of the use of the semi-supervised method, we compared the performance of the proposed semi-supervised method and the best supervised method (GMM-MFCC) using different

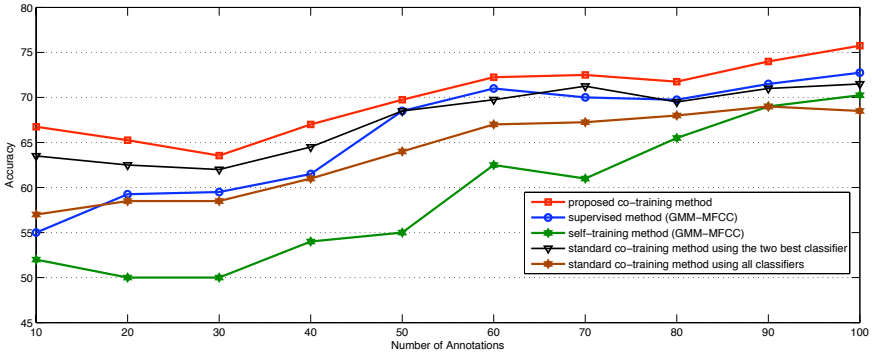


Figure 6: Classification accuracy with different numbers of annotations

Table 8: Classification accuracy with different numbers of annotations for training and 400 utterances for testing

Number of annotations for training	10	20	30	40	50	60	70	80	90	100
Proposed Co-training method	66.8	65.3	63.5	67.0	69.8	72.3	72.5	71.8	74.0	75.8
Co-training standard (using h1 and h4)	63.5	62.5	62.0	64.5	68.5	69.8	71.3	69.5	71.0	71.5
Co-training standard (using all the classifiers h1-h9)	57.0	58.5	58.5	61.0	64.0	67.0	67.3	68.0	69.0	68.5
Self-training (using h1: MFCC-GMM)	52.0	50.0	50.0	54.0	55.0	62.5	61.0	65.0	69.0	70.3
Self-training (using h2: prosody-GMM)	54.0	52.5	53.0	52.0	53.5	58.0	59.0	62.0	64.5	67.8
Supervised method: MFCC-GMM (best configuration)	55.0	59.3	59.5	61.5	68.5	71.0	70.0	69.8	71.5	72.8

Table 7: Initialization of Co-training algorithm

Classifiers (views)	Combination
h1	GMM trained with $f1$
h2	GMM trained with $f2$
h3	k-NN trained with $f3$
h4	k-NN trained with $f4$
h5	SVM trained with $f5$
h6	GMM trained with $f6$
h7	MLP trained with $f7$
h8	GMM trained with $f8$
h9	MLP trained with $f9$

numbers of annotations (from 10 labelled data to 100 labelled data). Figure 6 and Table 8 show that the proposed co-training method outperforms the supervised method especially with limited labelled data for training (always 400 utterances for testing), 66.8% vs. 55.0% with 10 labelled utterances.

Moreover, Figure 8 demonstrates that the proposed co-training algorithm performs better in the first several iterations (93.5% accuracy in the first iteration). This result is quite reasonable because, as shown in Figure 7, there are many more correctly classified than falsely classified utterances in the first iteration (101 correctly classified utterances vs. 7 falsely classified utterances). However, the performance of the classification decreases in the last iterations because we are retraining the system on misclassified utterances detected incorrectly in previous iterations.

8. Conclusion

In this article, a co-training algorithm was presented to combine different views to predict the common class variable for emotional speech classification. Our goal was to develop a motherese detector by computing multi-features and multi-classifiers to automatically discriminate pre-segmented infant-directed speech segments from manually pre-segmented adult-directed segments, so as to enable the study of parent-infant interactions and the investigation of the influence of this kind of speech on interaction engagement. By using the more conventional features often used in emotion recognition, such as cepstral MFCC, and other features, including prosodic features with some statistics on the pitch and energy and bark features, we were able to automatically discriminate infant-directed speech segments. Using classification techniques that are often used in speech/emotion recognition (GMM, k-NN, SVM and MLP) we developed different classifiers and we have tested them on real-life home movies database. Our experimental results show that spectral features alone contain much useful information for discrimination because they outperform all other features investigated in this study. Thus, we can conclude that cepstral MFCC alone

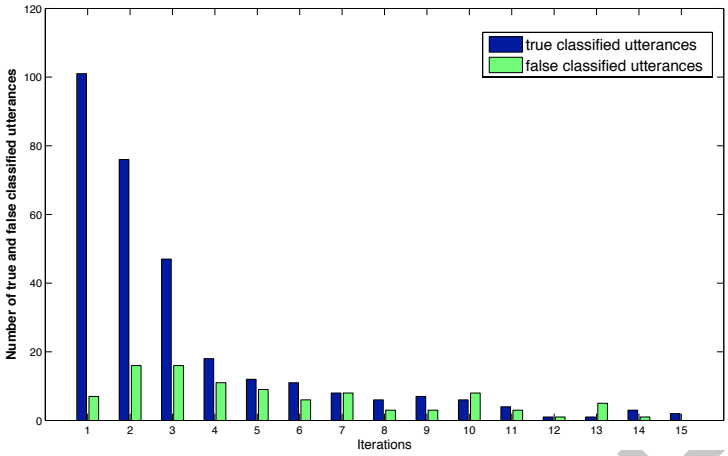


Figure 7: Number of accurately and falsely classified utterances by iteration

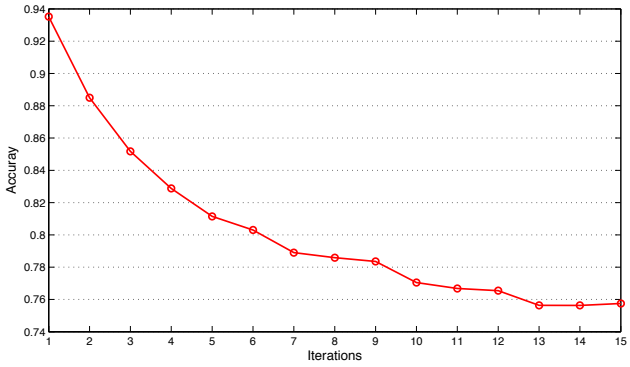


Figure 8: Accuracy by iteration

can be used effectively to discriminate infant-directed speech.

However, this method requires a large amount of labelled data. Therefore, we investigated a semi-supervised approach that combines labelled and unlabelled data for classification. The proposed semi-supervised classification framework allows the combination of multi-features and the dynamic penalisation of each classifier by iteratively calculating its classification confidence. The experimental results demonstrate the efficiency of this method.

For our infant-directed speech classification experiments, we used only utterances that were already segmented (based on a human transcription). In other words, the automatic segmentation of infant-directed speech was not investigated in this study, but it can be addressed in a follow-up study. Automatic infant-directed speech segmentation can be seen as a separate problem, which gives rise to other interesting questions, such as how to define the beginning and the end of infant-directed speech, and what kind of evaluation measures to use.

In addition, other issues remain to be investigated in the future. We plan to test our semi-supervised classification method on larger emotional speech databases. Then it will be interesting to investigate the complementarities of the different views by analysing the evolution of weights of each classifier and to compare our algorithm with other semi-supervised algorithms, especially algorithms using multi-view features.

9. Acknowledgments

The authors would like to thank Filippo Muratori and Fabio Apicella from Scientific Institute Stella Maris of University of Pisa, Italy, who have provided data; family home movies. We would also like to extend our thanks to David Cohen and his staff, Raquel Sofia Cassel and

Catherine Saint-Georges, from the Department of Child and Adolescent Psychiatry, AP-HP, Groupe Hospitalier Pitié-Salpétrière, Université Pierre et Marie Curie, Paris France, for their collaboration and the manual database annotation and data analysis. Finally, this work has been partially funded by La Fondation de France.

References

[1] A. P. Association, The Diagnostic and Statistical Manual of Mental Disorders, IV, Washington, D.C, 1994.
[2] A. Fernald, P. Kuhl, Acoustic determinants of infant preference for motherese speech, Infant Behavior and Development 10 (1987) 279–293.
[3] P. Kuhl, Early language acquisition: Cracking the speech code, Nature Reviews Neuroscience 5 (2004) 831–843.
[4] R. Cooper, R. Aslin, Preference for infant-directed speech in the first month after birth, Child Development 61 (1990) 1584–1595.
[5] A. Fernald, Four-month-old infants prefer to listen to motherese, Infant Behavior and Development 8 (1985) 181–195.
[6] D. Grieser, P. Kuhl, Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese, Developmental Psychology 24 (1988) 14–20.
[7] C. Burnham, C. Kitamura, U. Vollmer-Conna, What's new pussycat: On talking to animals and babies, Science 296 (2002) 1435.
[8] F. Muratori, S. Maestro, Autism as a downstream effect of primary difficulties in intersubjectivity interacting with abnormal development of brain connectivity, International Journal Dialogical Science Fall 2(1) (2007) 93–118.
[9] A. Mahdhaoui, M. Chetouani, C. Zong, R. Cassel, M.-C. Saint-Georges, C. Laznik, S. Maestro, F. Apicella, F. Muratori, D. Cohen, Computerized home video detection for motherese may help to study impaired interaction between infants who become autistic and their parents, International Journal of Methods in Psychiatry Research In press.
[10] A. Mahdhaoui, M. Chetouani, C. Zong, R. Cassel, M.-C. Saint-Georges, C. Laznik, S. Maestro, F. Apicella, F. Muratori, D. Cohen, Multimodal signals: cognitive and algorithmic issues, Springer, 2009, Ch. Automatic Motherese detection for face-to-face interaction analysis, pp. 248–55.
[11] M. Laznik, S. Maestro, F. Muratori, E. Parlato, Au commencement tait la voix, Ramonville Saint-Agne: Eres, 2005, Ch. Les interactions sonores entre les bebes devenus autistes et leur parents, pp. 81–171.
[12] C. Saint-Georges, R. Cassel, D. Cohen, M. Chetouani, M. Laznik, S. Maestro, F. Muratori, What studies of family home movies can teach us about autistic infants: A literature review, Research in Autism Spectrum Disorders 4(3) (2010) 355–366.
[13] A. Mahdhaoui, M. Chetouani, C. Zong, Motherese detection based on segmental and supra-segmental features, in: International Conference on Pattern Recognition-ICPR, 2008, pp. 8–11.
[14] M. Chetouani, A. Mahdhaoui, F. Ringeval, Time-scale feature extractions for emotional speech characterization, Cognitive Computation 1 (2009) 194–201.
[15] M. Slaney, G. McRoberts, Babyears: A recognition system for affective vocalizations, Speech Communication 39 (2003) 367–384.
[16] T. Inouea, R. Nakagawab, M. Kondoua, T. Kogac, K. Shinoharaa, Discrimination betweenmothersinfant-andadult-directedspeechusinghidden markov models, Neuroscience Research (2011) 1–9.
[17] M. Shami, W. Verhelst, An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech, Speech Communication 49 (2007) 201–212.
[18] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: Conference on computational learning theory, 1998.
[19] K. Nigam, A. McCallum, S. Thrun, T. Mitchell, Text classification from labeled and unlabeled document using em, in: International Conference on Machine Learning, 2000.
[20] X. Zhu, J. Lafferty, Z. Ghahramani, Semi-supervised learning using gaussian fields and harmonic functions, in: International Conference on Machine Learning, 2003, pp. 912–919.
[21] D. Zhou, B. Schlkopf, T. Hofmann, Advances in Neural Information Processing Systems (NIPS) 17, MIT Press, Cambridge, MA, 2005, Ch. Semi-Supervised Learning on Directed Graphs, pp. 1633–1640.
[22] S. Goldman, Y. Zhou, Enhancing supervised learning with unlabeled data, in: International Conference on Machine Learning, 2000, pp. 327–334.
[23] S. Maestro, F. Muratori, M. Cavallaro, C. Pecini, A. Cesari, A. Paziente, D. Stern, B. Golse, F. Palasio-Espasa, How young children treat objects and people: an empirical study of the first year of life in autism, Child psychiatry and Human Development 35(4) (2005) 83–396.
[24] J. Cohen, Educational and Psychological Measurement, 1960, Ch. A coefficient of agreement for nominal scales, p. 3746.
[25] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, L. Amir, N.and Kessous, V. Aharonson, The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals, in: Interspeech, 2007, pp. 2253–2256.
[26] K. Truong, D. van Leeuwen, Automatic discrimination between laughter and speech., Speech Communication 49 (2007) 144–158.
[27] L. Kessous, N. Amir, R. Cohen, Evaluation of perceptual time/frequency representations for automatic classification of expressive speech, in: paraling, 2007.
[28] P. Boersma, D. Weenink, Praat, doing phonetics by computer, Tech. rep., Institute of Phonetic Sciences, University of Amsterdam, Pays-Bas, (2005).
URL `www.praat.org`
[29] M. Shami, M. Kamel, Segment-based approach to the recognition of emotions in speech, in: IEEE Multimedia and Expo, 2005.
[30] R. Cooper, J. Abraham, S. Berman, M. Staska, The development of infants preference for motherese, Infant Behavior and Development 20(4) (1997) 477–488.
[31] E. Zwicker, Subdivision of the audible frequency range into critical bands, Acoustical Society of America 33(2) (1961) 248.
[32] A. Esposito, M. Marinaro, Nonlinear speech modeling and applications, Springer, Berlin, 2005, Ch. Some notes on nonlinearities of speech, pp. 1–14.
[33] E. Zwicker, H. Fastl, Psychoacoustics: Facts and Models, Berlin, Springer Verlag, 1999.
[34] D. Reynolds, Speaker identifcation and verification using gaussian mixture speaker models, Speech Communication 17(1-2) (1995) 91–108.
[35] R. Duda, P. Hart, D. Stork, Pattern Classification, second edition, Wiley-interscience, 2000.
[36] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, Tech. rep., Department of Computer Science, National Taiwan University, Taipei (2001).

URL http://www.csie.ntu.edu.tw/ cjlin/libsvm/
[37] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[38] F. Eibe, I. Witten, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, The Morgan Kaufmann Series in Data Management Systems, 1999.
[39] V. Vapnik, Statistical Learning Theory, New York: Wiley, 1998.
[40] J. Platt, Advances in Large Margin Classifiers, MIT Press, Cambridge, MA, 1999, Ch. Probabilistic Outputs for SVM and comparison to regularized likelihood methods.
[41] C. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995.
[42] K. Nigam, R. Ghani, Analyzing the effectiveness and applicability of co-training, in: Ninth International Conference on Information and Knowledge Management, 2000, pp. 86–93.
[43] U. Brefeld, T. Gaertner, T. Scheffer, S. Wrobel, Efficient co-regularized least squares regression, in: International conference on machine learning, 2006.
[44] I. Muslea, S. Minton, C. Knoblock, Selective sampling with redundant views, in: Proceedings of Association for the Advancement of Artificial Intelligence, 2000, pp. 621–626.
[45] Q. Zhang, S. Sun, Multiple-view multiple-learner active learning, Pattern Recognition 43(9) (2010) 3113–3119.
[46] R. Schapire, Y. Singer, Improved boosting algorithms using confidence-rated predictions, Machine Learning, Springer Netherlands 37(3) (1999) 297–336.
[47] A. Carlson, Coupled semi-supervised learning, Ph.D. thesis, Carnegie Mellon University, Machine Learning Department (2009).