A NEW APPROACH FOR MOTHERESE DETECTION USING A SEMI-SUPERVISED ALGORITHM

Ammar Mahdhaoui and Mohamed Chetouani

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

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

ABSTRACT

Authentic and natural infant-parent interactions analysis requires the development of efficient detectors such as the discrimination between infant and adult-directed speech. Supervised methods have been found to be efficient for labeled data. The annotation process is time-consuming and the eventual divergence between annotators increases the difficulty. Semi-supervised approaches such as co-training offers a framework allowing to take advantage of supervised classifiers trained by different features. The proposed motherese detector system combined various features and classifiers used in emotion recognition in a co-training framework. The results show the relevance of this approach for real-life corpora such as home movies.

1. INTRODUCTION

Researchers in autism pathology and parent-infant interaction highlighted the importance of infant-directed speech for infants who will become autistic [14] [16]. Given that home movies offer a unique opportunity to follow infant development and Parent/Infant interactions. The study of home movies is very important for future research, but the use of this kind of database makes the study very difficult and long, the manual annotation of these films is very costly and time consuming. In addition, from the manual analysis, we have found that 80% of positive sequences (i.e. multimodal response of the infant: vocalization, gaze, facial expression) have been induced by motherese. So, for the analysis of the role of infant-directed speech during interaction, we developed an automatic motherese detection system.

Motherese is a highly communicative and social event in parents child communication that can elicit emotional reactions. If we look to the definition of infant-directed speech and the acoustic characteristics of this kind of speech, hight pitch/dialect/register [3], the problem of classification seems to be not very complicated, a simple method based only on prosodic feature should immediately discriminate motherese from normal speech. However, in [13] [14], we have shown that the processing of natural and authentic interactions requires the development of methods beyond the strict definition: prosodic features alone were not sufficient to resolve this problem. We tested many machine learning techniques, statistical and parametric, with different feature extraction methods (time/frequency domains). GMM (Gaussian mixture model) with MFCC (Mel-frequency cepstrum) features were found to be efficient (73.5% accuracy). However, the training of GMM classifiers requires a large amount of data, while until now we have collected around 300 labeled segments of motherese and normal speech from family home movies annotated by two psycholinguists on two categories (motherese and normal speech).

In this work, we are interested on non-acted databases such as home movies. The study of home movies is very important for future research, but the use of this kind of database makes the study very difficult and long, the manual annotation of these films is very costly in time. In addition to annotated utterances (300 on total), we have a lot of utterances not yet annotated, so we investigated a semi-supervised approach which does not require a large number of annotated data, this method combines labeled and unlabeled utterances to enable the infant-directed speech discrimination.

In the area of classification, many semi-supervised learning algorithms have been proposed, one of these algorithms is the co-training approach [2]. Most applications of co-training approach have been devoted to text classification [12] [21] and web page categorization [2] [7]. However, there are few studies related to semi-supervised learning for speech emotion recognition. The co-training algorithm proposed by Blum and Mitchell [2] is a prominent achievement in semi-supervised learning. It initially defines two classifiers on distinct attribute views of few labeled data. Either of the views is required to be conditionally independent to the other and sufficient for learning a classification system. Then iteratively, each classifier's predictions on unlabeled examples are selected to increase the training data set. This co-training algorithm and its variations [18] have been applied in many application areas because of their theoretical

978-1-4244-4948-4/09/$25.00 © 2009 IEEE

justifications and experimental success. The originality of our work is to use multi features and multi classifiers to minimize the error of classification. To obtain a more accurate system, we employ all the investigated features and classifiers to predict the correct class label.
This paper is organized as follows. In section 2, different supervised learning and feature extraction methods are described. Section 3 presents the details of our co-training algorithm. In section 4, a description of the database used in the experiments is given and different experimental results are shown. Section 5 concludes this paper.

2. FEATURES AND CLASSIFIERS

2.1. Features extraction

Temporal and frequential features are usually investigated in emotion recognition [11] [17]. In this study, 70 prosodic, 16 cepstral and 96 spectral features were extracted, which were shown to be the most efficients [9] [10] [13]. In addition, we computed frame-level and utterance-level features that can be used in different modeling techniques to develop motherese detection system.

2.1.1. Frame-level features extraction

Frame-level features refer to features extracted each 20 ms of the utterance, so the number of the resulting feature vectors is variable and depends on the length of the utterance.
Cepstral features such as MFCC are often successfully used in speech and emotion recognition. Short-term cepstral signatures of both motherese and normal speech are characterized by MFCC features.
Several studies have shown that fundamental frequency (F0) and energy features are very important to emotion recognition applications [9]. F0 and energy were estimated each 20 ms [4], and we computed, for each voiced segment, 3 statistics: mean, variance and range for both F0 and short-time energy resulting in a 6 dimensional vector.

2.1.2. Utterance-level features extraction

Utterance-level features refer to features extracted per whole utterance, so we have one feature vector by utterance.
In addition to pitch estimations per frame, we also measured some more global higher-level pitch features to capture the fluctuations and variability of fundamental frequency. We computed 32 statistics: maximum, minimum and mean values, 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 coefficients 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 0.1 and 0.9, and interquantile range between these two values, absolute and sign of time interval between maximum and minimum appearances. These 32 statistical features are also extracted in order to model the dynamic variations of the bark spectral perceptive representation. Three other features are also extracted from the pitch contour and the loudness contour, by using histograms and considering the maximum, the bin index of the maximum and the center value of the corresponding bin. These 3 features are relevant for pitch and energy contours.
To extract spectral bark features, for a given spectral model we performed the analysis on successive time frames along each utterance. We then extracted a set of F (in our case 32) statistical features from these representations. These features can be applied either along the time axis or along the frequency axis as shown in figure 1.
Indeed as described in [1], three approaches, for extraction of a feature vector of a defined dimension F, were used:

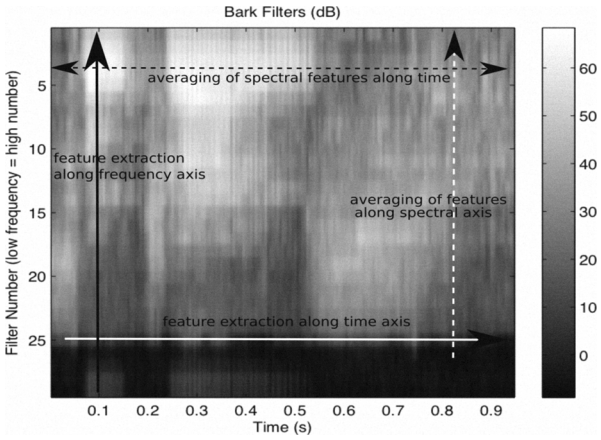

Fig. 1. bark-based features extraction

- TL Approach (for 'Time Line'): (1) extracting the F features on the spectral vector of each time frame, (2) averaging the values of the F features along time.

- SL Approach ('for Spectral Line'): (1) extracting the F features along the time axis for each spectral band, (2) averaging the F features along the frequency axis.

- MV Approach (for 'Mean Values'): (1) averaging the values along the time axis to obtain a spectral vector of dimension S, (2) extracting the F features on this spectral vector.

Finally, we have 9 kind of feature vectors with different dimensions and are presented in Table 1.

Table 1. Feature extraction

	Features	Dimension
1	MFCC	16 MFCCs
2	Pitch+Energy-6	6 statistics on pitch and energy
3	Pitch-35	35 statistics on pitch
4	Energy-35	35 statistics on energy
5	Pitch+energy-70	70 statistics on pitch and energy
6	Bark TL+SL+MV	96 statistics
7	Bark TL	32 statistics
8	Bark SL	32 statistics
9	Bark MV	32 statistics

2.2. Modeling techniques

After feature extraction, the classification task can be achieved using standard machine learning methods. In this study, four different classifiers: Gaussian mixture model (GMM) [19], Neural network (MLP) [8], SVM [5] [20] and k-nearest neighbor (k-NN) [6] classifiers, were investigated. Each classifier considers the selected features as the most efficient for the two-classes discrimination problem. With a relatively small number of training samples compared to the dimensionality of the data, a high risk of bias due to variances in training material is present. In order to improve instable classifiers we investigated a novel approach which combines labeled and unlabeled utterances for motherese detection using a co-training algorithm.

3. CO-TRAINING ALGORITHM

In the previous section, for supervised method, we tried to optimize each classifier for each feature set. However, all the classifiers presented previously require a large number of utterances to enable efficient learning, therefore we proposed a novel method based on a co-training approach described in Table 2. Given a set L of labeled training utterances and a set U of unlabeled utterances, we extract frame-level and utterance-level features from all these utterances. First, to initialize the algorithm we found the best feature set for each classifier. We obtained a table of correspondence which link each classifier with the best feature set as described in Table 4. Secondly, a loop for n iterations (n is the number of classifier, 9 in our case) on :

- select the ensemble of segments T for which label predictions from all the classifiers are identical and add them to the training base and remove them from test database U.
- remove the classifier which is the least correlated with the other classifiers.
- compute correlation: for each couple of classifiers we calculate the number of utterance whose label predictions are identical.

Table 2. The Co-Training algorithm

Given:
  a set L of labeled training examples
  a set U of unlabeled examples
  n = number of classifier
Initialization:
  Find the best feature set for each classifier
For n iterations:
  Use L to train each classifier *hi*
  Classify all examples of U by each *hi*
  Take p positive examples and n negative examples from U
  which has identical annotation by all the classifier
  $T = p+v$
  Add T to L and remove T from U
  Remove the classifier which is least correlated with the
  other classifiers
End for
If (U is empty) then
  exit the program
else
  Classify the remaining examples from U using best configuration *hi*
END

The correlation is estimated from the classifier outputs on the unlabeled data. For example, to estimate the correlation P_i of classifier C_i, as shown in Fig.2, we calculate the number $N1$ of identical positive segments, $N2$ number of segment annotated as positive for classifier C_i and negative for the other classifiers, $M1$ number of negative identical segments and $M2$ number of segment labeled as negative for C_i and positive for the others classifiers . Then, we calculate the correlation coefficient P_i:

$$P_i = (N1 + M1)/(N1 + M1 + N2 + M2)$$

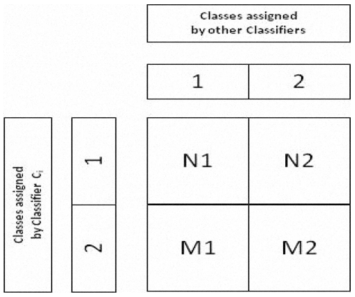

Fig. 2. Correlation analysis matrix

If we remove the most correlated classifier, we risk to have a lot of utterances not classified after the last iteration. And these utterances will be classified using the best single classifier which can cause a decrease in performance of classification. We will test the two approaches in section 4.2.2. At the end, if the data test U is not empty, we classify the remaining examples using the best classifier *hi*.

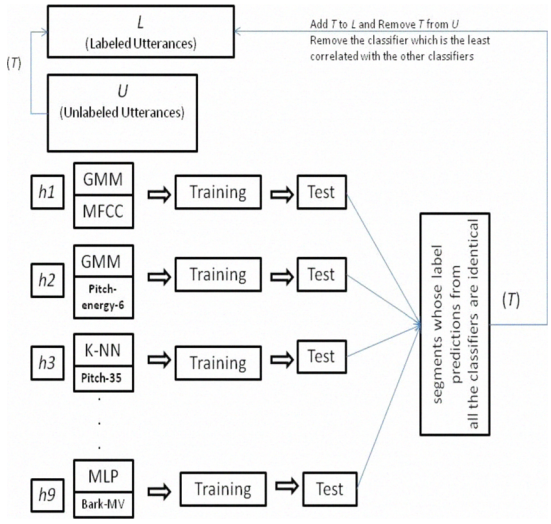

Fig. 3. Structure of Co-training Algorithm

4. CLASSIFICATION EXPERIMENTS AND RESULTS

4.1. Database description

The speech corpora used in our study are a collection of natural and spontaneous interactions. These corpora contain expressions of non-linguistic communication (affective intent) conveyed by a parent to a preverbal child. The corpora are a real parents/child interactions and consist of recordings of Italian mothers and fathers as they addressed their infants. We decided to focus on the analysis of home movies as it makes 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. All sequences were extracted from the Pisa home movies video database [15]. 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 different home videos of the first year of an infant. Verbal interactions of the mothers have been carefully annotated by two psycholinguists on two categories: infant directed speech (motherese) and adult directed speech (normal). From this manual annotation, we extracted 152 utterances for each class. The utterances are typically between 0.5s and 4s length.
We divided the database into two parts, test database L which contains 104 utterances of motherese and normal speech and base learning U that contains 200 utterances of motherese and normal speech.

Table 3. Accuracy of separate classifier (in %)

	Features	GMM	k-NN	SVM	MLP
Frame level	MFCC	**73.5**	57.71	59.41	61.35
	Pitch+Energy-6	**59.5**	55.73	54.65	50.18
Utterance level	Pitch-35	54.69	**55**	50	50
	Energy-35	67	**68.5**	65.5	58.5
	Pitch+energy-70	62.12	65.5	**65.5**	54.5
	Bark TL+SL+MV	**61**	50.5	49	54.5
	Bark TL	55.5	51	52	**58.5**
	Bark SL	**65**	52	50.5	55.5
	Bark MV	58.76	50.5	50.5	**64**

4.2. Experiments results

4.2.1. Supervised Classifiers

The performances of the different classifiers, each trained with different feature sets (MFCC, Pitch+Energy-6, Pitch-35, Energy-35, Pitch+energy-70, Bark TL, Bark SL, Bark MV and Bark TL+SL+MV) were evaluated on the home movies database. We use the accuracy rate to compare the performances of the different separate classifiers.
Table 3 shows the best result of all classifiers trained with different feature sets. Best result obtained with GMM trained with cepstral MFCC features, second best result is obtained with k-NN trained with Energy-35 features. Therefore, table 3 shows that frame level features MFCC outperform the other features, but in comparing the results of all the feature sets and taking into account the different classifiers, we can observe that utterance level features in most of the cases, perform better than frame levels.
To summarize, the best performing feature set for motherese detection appears to be the frame-level cepstral (MFCC). Regarding the classifiers, we can observe that GMMs generalize better over different test cases than other classifiers do.

4.2.2. Semi-Supervised Classifiers

The algorithm works as described in figure 2. To initialize the co-training algorithm, we take into account the best configuration of each features trained with all supervised classifiers. We obtained 9 classifiers, *h1* to *h9* as described in table 4. After that, we take the ensemble of utterances T which has same label predictions from all the classifiers *hi*, we add T to train data L and remove it from test data U. Figure 3 presents the number of utterances for which the label predictions from all the classifiers are identical. Then, the least correlated classifier is removed. In the first iteration we removed *h5*.
The accuracy of co-training algorithm is about 75.5 %. Table 5 shows that, with only 104 labeled utterances, we obtained statistically significant improvements in classification accuracy compared to the best results of the supervised classifier GMM trained with cepstral MFCC features). If we modify our algorithm by removing the most correlated clas-

Table 4. Initialization of Co-training algorithm

Classifiers	Combination
h1	GMM trained with MFCC
h2	GMM trained with Pitch+Energy-6
h3	k-NN trained with Pitch-35
h4	k-NN trained with Energy-35
h5	SVM trained with Pitch+energy-70
h6	GMM trained with Bark TL+SL+MV
h7	MLP trained with Bark TL
h8	GMM trained with Bark SL
h9	MLP trained with Bark MV

Table 5. Best Results

Method	Accuracy
Supervised training	73.5%
Semi Supervised training	75.5%

sifier within each iteration, the accuracy decreases to 64.5 % because it remains 49 utterances to be classified with *h1*.

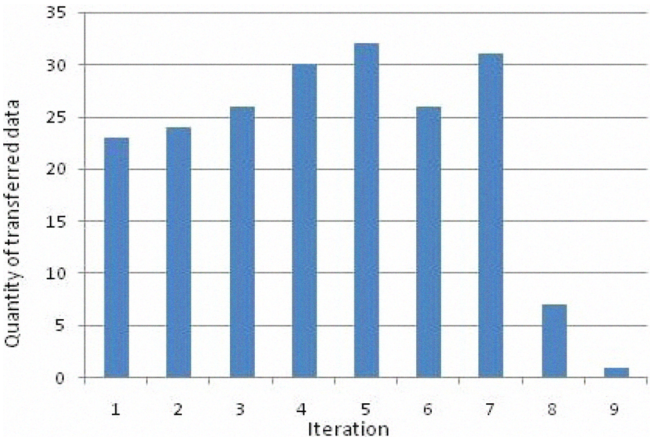

Fig. 4. Quantity of transferred utterances from training data to test data by iteration

5. CONCLUSIONS

Using unlabeled data to enhance the performance of classifiers trained with few labeled data has many applications in pattern recognition. In this paper, we proposed a new approach for infant-directed speech discrimination based on semi-supervised learning. We computed frame level and utterance level features used to train different classification models.

The performance of a pattern recognition system highly depends on the discriminant ability of the features. We started this work by comparing different supervised classification techniques for the discrimination of infant-directed speech from normal speech. We found that GMM classifier trained with cepstral MFCC feature gives the best result. However, the best system obtained with GMM has many limitations such as the need of a large quantity of annotated data for training. To enhance our detector, we investigated the semi-supervised learning approach. A co-training algorithm was presented to utilize the different features of manually pre-segmented unlabeled examples to reduce the need of expensive labeled data. The semi-supervised method achieved better performance than the method based on supervised learning. In the future, we intend to test other semi-supervised algorithms and to integrate a non-supervised method like k-means to improve the performance of the co-training algorithm.

6. ACKNOWLEDGEMENTS

Thanks to 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 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.

7. REFERENCES

[1] Amir, N., Cohen, R., "Characterizing Emotion In the Soundtrack of an Animated Film: Credible or Incredible?" ACII 2007,(pp. 148-158).

[2] A. Blum and T. Mitchell, Combining from labeled and unlabeled data with co-training, Proc. of Annual Conference on Computational Learning Theory, pp. 92 100, 1998.

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

[4] Boersma, P., Weenink, D., 2005. Praat: doing phonetics by computer (Version 4.3.01) [Computer program]. Retrieved from www.praat.org.

[5] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001. http://www.csie.ntu.edu.tw/ cjlin/libsvm/

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

[7] D. Zhou, B. Scholkopf, and T. Hofmann, Semisupervised learning on directed graphs, Advances in Neural Information Processing System, pp. 16331640, 2005.

[8] Eibe Frank Ian H.Witten, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor, October 1999.

[9] Khiet P. Truong, David A. van Leeuwen., "Automatic discrimination between laughter and speech", Speech Communication 49 (2007) 144158.

[10] Kessous, L., Amir, Noam., Cohen, Rachel ., "Evaluation of perceptual time/frequency representations for automatic classification of expressive speech", paraling 2007.

[11] K.P. Truong, D.A. van Leeuwen, Automatic discrimination between laughter and speech. Speech Communication 49 (2007)144158

[12] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell, Text classification from labeled and unlabeled document using em, Proc. of International Conference on Machine Learning, vol. 39, no. 2, pp. 103134, 2000.

[13] Mahdhaoui, A., Chetouani, M., Zong, C., 2008, Motherese Detection Based On Segmental and Supra-Segmental Features, International Conference on Pattern Recognition-ICPR, December 8-11, 2008, Tampa, Florida, USA.

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

[15] Maestro S, Muratori F, Cavallaro MC, Pecini C, Cesari A, Paziente A, Stern D, Golse B, Palasio-Espasa F. 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).

[16] Muratori F, Maestro S. Autism as a downstream effect of primary difficulties in intersubjectivity interacting with abnormal development of brain connectivity. Int J Dialog Sci Fall. 2007;2(1):93-118.

[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) 201212

[18] S. Goldman and Y. Zhou, Enhancing supervised learning with unlabeled data, Proc. of International Conference on Machine Learning, pp. 327334, 2000.

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

[20] Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer, New York.

[21] X. Zhu, J. Lafferty, and Z. Ghahramani, Semisupervised learning using gaussian fields and harmonic functions, Proc. of International Conference on Machine Learning, pp. 912919, 2003.