Detecting and Labeling Speakers on Overlapping Speech using Vector Taylor Series

Pranay Dighe, Marc Ferràs, Hervé Bourlard

Idiap Research Institute, CH-1920 Martigny, Switzerland

pranay.dighe@idiap.ch, marc.ferras@idiap.ch, bourlard@idiap.ch

Abstract

Successfully modeling overlapping speech is a crucial step towards improving the performance of current speaker diarization systems. In this direction, we present ongoing work on a novel Multi-Class Vector Taylor Series (MC-VTS) approach that models overlapping speech from knowledge of the individual speaker models and the feature extraction process. We explore several variants of the MC-VTS technique that aim at modeling overlapping speech more precisely. Bootstrapping the algorithm with both oracle and diarization output segmentations, we show the potential of this approach in terms of overlapping speech detection and speaker labeling performances through a set of experiments on far-field microphone meeting data.

Index Terms: Multi-Class Vector Taylor Series, Overlap Detection, Speaker Diarization

1. Introduction

Speech overlap is a common phenomenon in a multi-party conversation. In situations involving an open exchange of ideas, discussion and debate, such as in a meeting scenario, it has been found [1] that speech overlaps can occur nearly 20% of the speech time. For speaker diarization systems, determining "who spoke when" in an audio recording, performance suffers when it comes to detecting and labeling speakers in overlapping speech segments. In addition, the presence of overlapping speech also results in inaccurate modelling of individual speaker models which in turn again degrades the diarization performance.

Though overlapping speech is simply a linear combination of the individual speech sources in the signal and spectral domains, statistical modeling in such high-dimensional vector spaces is challenging. Speech technologies typically seek low-dimensional vectors that gather enough information to solve a given task. Overlapping speech can be modeled as a set of linear and non-linear operations on the individual speech sources in the cepstral domain. As we did in our previous work [17], we model these non-linearities through a Vector Taylor Series approximation of the overlapping speech model. In this paper we extend the work on overlap detection towards increasing detection performance but also diarization performance by labeling overlapping speech speakers from the speaker diarization output.

This work was supported by the European Union under the FP7 Integrated Project inEvent (Accessing Dynamic Networked Multimedia Events), grant agreement 287872. The authors gratefully thank the EU for their financial support and all project partners for a fruitful collaboration.

A range of studies has been done on overlap detection and speaker diarization of overlapping speech. Boakye et al. used a HMM-based segmenter [2] to detect speech, non-speech and overlapping speech from meeting audio, where the models are trained using cepstral features together with instantaneous and LPC residual energies and diarization posterior entropy from ground truth alignments. The detected overlapped segments are labeled with multiple speakers in the output given by a standard diarization system. Convolutive non-negative sparse coding (CNSC) used in [3], [4], [5] and [6] has also been found effective for overlap detection. Features extracted using CNSC are used together with a wide variety of features such as cepstral features, energy, jitter, shimmer and even linguistic features in a probabilistic framework. The knowledge of the silence distribution in meeting recordings in [7] and analysis of long-term conversational features (distribution of overlap occurence) in [8] has also been shown to improve detection and diarization performance.

This paper is organized as follows: Section 2 describes the multi-class VTS approach to modeling overlapping speech. We discuss the framework and describe how the overlapping speech detection system works. Sections 3 provides experimental details and results, evaluating this approach on meeting data for the overlapping speech detection and speaker diarization tasks. Conclusions are given in Section 4.

2. Modeling Overlapping Speech using VTS

In our previous work [17], we proposed to use Vector Taylor Series (VTS) to model overlapping speech in the cepstral domain. Just as signal and additive noise add together to obtain the noisy signal, the speech signals from two speakers add together to obtain an overlapping speech signal. However, noise is often stationary while speech is not. We addressed this point in [17] by proposing a multi-class variant of the VTS framework. On the other side, noise is usually assumed to be uncorrelated with speech, at least in the long term. When two speakers speak simultaneously this is hard to assume as the speech production system is similar for human speakers. We assume that the individual speaker speech signals may be correlated to each other in this work.

The process of superposing two speech signals, $x_1(t)$ and $x_2(t)$ from different speakers,

$$y(t) = x_1(t) + x_2(t) \quad , \tag{1}$$

translates into a complex non-linear transformation in the cepstral case. For Mel-Frequency Cepstral Coefficients (MFCC) we can arrange such expression in terms of the feature vector of one speaker $\mathbf{x}_1$ and a non-linear function of the feature vectors from both speakers $\mathbf{x}_1$ and $\mathbf{x}_2$, as

$$\mathbf{y}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1 + g(\mathbf{x}_2 - \mathbf{x}_1) \tag{2}$$

with

$$\begin{aligned} g(\mathbf{x}_2 - \mathbf{x_1}) = \mathbf{C}\ln(1 + \exp(\mathbf{C}^{-1}(\mathbf{x}_2 - \mathbf{x}_1) \\ +2\alpha \cdot \exp(\mathbf{C}^{-1}(\mathbf{x}_2 - \mathbf{x}_1)/2)) \end{aligned} \tag{3}$$

where α is the correlation between the spectra of the two speakers, typically unknown. The term $\mathbf{x_2}$ is modeled using multiple acoustic classes to deal with the phone variability expressed by the second speaker within a speech segment. Standard VTS uses a single Gaussian assuming noise to be stationary.

We use prior knowledge of two individual-speaker GMM, trained using the feature vectors $\mathbf{X}_1 = (\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{1,T})$ and $\mathbf{X}_2 = (\mathbf{x}_{2,1}, \ldots, \mathbf{x}_{2,T})$, with T frames duration. We let the VTS technique estimate the corrupted GMM parameters assuming these two sources are overlapping.

2.1. Overlapping Speech Detection

The overlapping speech detection (OSD) system requires individual speaker models and speech data as inputs. We assume the speaker models are available, either trained from oracle speaker segmentations or via automatic speaker diarization.

To detect overlapping speech, we perform a series of hypothesis tests on each sliding window of the input features, assessing how more likely it is for overlapping speech to occur compared to single-speaker speech. Since we consider overlap from two speakers only, the number of possible overlapping speech models is $N^2 - N$, with N^2 hypotheses to test and N being the number of speakers. A faster approach is to assign a main speaker to each segment, the speaker obtaining the largest average likelihood score, to decrease the number of hypotheses from N^2 to N. In this case, if speaker i is the main speaker, we obtain the set of likelihood ratios

$$\frac{p(\mathbf{X}|\mathcal{S}_{1,i})}{p(\mathbf{X}|\mathcal{S}_i)}, \ldots, \frac{p(\mathbf{X}|\mathcal{S}_{i-1,i})}{p(\mathbf{X}|\mathcal{S}_i)}, 1, \ldots, \frac{p(\mathbf{X}|\mathcal{S}_{N,i})}{p(\mathbf{X}|\mathcal{S}_i)} \tag{4}$$

with $\mathcal{S}_{i,j}$ representing the hypothesis of speaker overlap between speakers i and j, and $\mathcal{S}_i$ representing the hypothesis of only speaker i speaking. Overlap and non-overlap hypotheses are modeled as follows:

- **Overlap:** For speaker pairs j, i in (4), we estimate the models $p(\mathbf{X}|\mathcal{S}_{j,i})$ using VTS mean adaptation as described in Section 2.
- **Single-speaker:** For the main speaker i, we adapt the mean vectors of the corresponding GMM using MAP adaptation [13] as

$$\hat{\boldsymbol{\mu}}_{x_m} = \alpha E_m[\mathbf{x}] + (1 - \alpha)\boldsymbol{\mu}_{x_m} \tag{5}$$

  where $E_m[\mathbf{x}] = \frac{1}{n_m}\sum_{t=1}^{T}\gamma_{mt}\mathbf{x}_t$ and $n_m = \sum_t^T \gamma_{mt}$. We determine the value of the interpolating factor α experimentally.

To determine whether overlap occured in the current window, we just pick the largest likelihood ratio value and decide on the corresponding hypothesis, i.e. single-speaker for hypothesis i, overlap otherwise.

2.2. Approximating the corrupted model using VTS

The goal of VTS is to approximate an overlapping speech model from the individual speaker models given some data assuming it contains overlapping speech. Since overlapping speech is non-stationary and several sounds might be being uttered by the main and corrupting speakers, we use the multi-class approach proposed in [17]. The acoustic space of the corrupting speaker is clustered into multiple classes and VTS adaptation is performed for each class independently.

Keeping these assumptions in mind, we let $\boldsymbol{\mu}_{y_m}$, $\boldsymbol{\mu}_{x_{1m}}$ and $\boldsymbol{\mu}_{x_{2m}}$ be the mean vectors of the m^{th} Gaussian component of the corrupted speech model, main speaker models and corrupting speaker model GMM respectively.

To cluster the acoustic space of the corrupting speaker, we start assuming that all the Gaussian components are observed for the corrupting speaker. If the average number of frames, ($\bar{\gamma}_{mt} = \frac{1}{T}\sum_{t=1}^{T}\gamma_{mt}$), assigned to a given Gaussian component is below a threshold, η, that component joins the Gaussian with the closest mean vector in terms of Euclidean distance. The average gamma for the new cluster becomes the sum of the corresponding average gammas. We use the mean of the Gaussian with largest gamma as the new cluster centroid. So the mean of the cluster c of corrupting speaker is denoted by $\boldsymbol{\mu}_{x_{2c}}$. In practice, matrix inversion issues are avoided if η is large enough.

The first-order VTS expansion of (2) w.r.t. vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ for Gaussian m around the point ($\boldsymbol{\mu}_{x_{1m}0}$,$\boldsymbol{\mu}_{x_{2c}0}$) is

$$\begin{aligned} \boldsymbol{\mu}_{y_m} \approx \boldsymbol{\mu}_{x_{1m}0} + g(\boldsymbol{\mu}_{x_{2c}0} - \boldsymbol{\mu}_{x_{1m}0}) + \mathbf{G}_m(\boldsymbol{\mu}_{x_{1m}} - \boldsymbol{\mu}_{x_{1m}0}) \\ +\mathbf{F}_m(\boldsymbol{\mu}_{x_{2c}} - \boldsymbol{\mu}_{x_{2}0}) \end{aligned} \tag{6}$$

where $\mathbf{G}_m$ and $\mathbf{F}_m$ are the derivatives of $\mathbf{y}$ w.r.t. $\mathbf{x}_1$ and $\mathbf{x}_2$ evaluated at the point ($\boldsymbol{\mu}_{x_{1m}0}$,$\boldsymbol{\mu}_{x_{2c}0}$), that is,

$$\mathbf{G}_m = \mathbf{C}\text{diag}(\frac{\exp(\beta) + \alpha \cdot \exp(\beta/2)}{1 + \exp(\beta) + 2\alpha \cdot \exp(\beta/2)})\mathbf{C}^{-1} \tag{7}$$

where

$$\beta = \mathbf{C}^{-1}(\boldsymbol{\mu}_{x_{2c}0} - \boldsymbol{\mu}_{x_{1m}0}) \tag{8}$$

$$\mathbf{F}_m = \mathbf{I} - \mathbf{G}_m \tag{9}$$

The mean of $\mathbf{y}$ for Gaussian m, i.e. $\boldsymbol{\mu}_{y_m}$, can then be obtained by taking the expectation operator on both sides of (6) which can be reduced to

$$\boldsymbol{\mu}_{y_m} \approx \boldsymbol{\mu}_{x_{1m}0} + g(\boldsymbol{\mu}_{x_{2c}0} - \boldsymbol{\mu}_{x_{1m}0}) \tag{10}$$

Similarly, using $\boldsymbol{\Sigma}_{x_{1m}}$ and $\boldsymbol{\Sigma}_{x_{2c}}$ the covariance matrices for Gaussian m of the main speaker and the corrupting speaker respectively, the corrupted covariance matrix $\boldsymbol{\Sigma}_{y_m}$ for Gaussian m can be approximated by

$$\boldsymbol{\Sigma}_{y_m} \approx \mathbf{G}_m\boldsymbol{\Sigma}_{x_{1m}}\mathbf{G}_m^T + \mathbf{F}_m\boldsymbol{\Sigma}_{x_{2c}}\mathbf{F}_m^T \tag{11}$$

Given that both main and corrupting speaker GMM are MAP-adapted from the same reference GMM, we assume that the m^{th} gaussian of the main speaker will be corrupted by the cluster, c, that contains the m^{th} component of the corrupting speaker GMM.

2.3. Estimation of VTS parameters

Given T frames of overlapping speech data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ and an initially corrupted GMM with M mixtures and parameters given by (10) and (11), the expectation-maximization (EM) algorithm iteratively finds estimates of $\boldsymbol{\mu}_{y_m}$ that further maximize the likelihood function

$$Q = \sum_{t \in T} \sum_{m \in M} \gamma_{t,m} \log(p(\mathbf{x}_t | \boldsymbol{\mu}_{y_m}, \boldsymbol{\Sigma}_{y_m})) \quad , \qquad (12)$$

eventually converging to a local maximum.

$\boldsymbol{\mu}_{y_m}$ is a function of $\boldsymbol{\mu}_{x_2}$ and $\boldsymbol{\Sigma}_{x_2}$, and their parameters are optimized to maximize the Q function above. Replacing the expectation of (6) into (12) and then differentiate w.r.t. $\boldsymbol{\mu}_{x_{2c}}$, the follwing update equation for $\boldsymbol{\mu}_{x_{2_c}}$ can be derived

$$\begin{aligned} \boldsymbol{\mu}_{x_{2c}} = \boldsymbol{\mu}_{x_{2c}0} + \{ \sum_{t \in T, m \in C} \gamma_{m,t} \mathbf{F}_m^T \boldsymbol{\Sigma}_{y_m}^{-1} \mathbf{F}_m \}^{-1} \\ \times \{ \sum_{t \in T, m \in C} \gamma_{m,t} \mathbf{F}_m^T \boldsymbol{\Sigma}_{\mathbf{y}_m}^{-1} (\mathbf{y}_t - \boldsymbol{\mu}_{y_m}) \} \end{aligned} \qquad (13)$$

where $\boldsymbol{\mu}_{x_{2c}0}$ is the previous estimate of the corrupting mean vector.

In this work, we also experiment with covariance adaptation of $\boldsymbol{\Sigma}_{x_2}$. We use the non-linear Conjugate Gradient (NLCG) algorithm [16] to update $\boldsymbol{\Sigma}_{x_2}$ assuming diagonal covariance matrices. As in [10], we update the Cholesky decomposition of $\boldsymbol{\Sigma}_{x_2}$, the diagonal matrix $\mathbf{U}_{x_2}$, to ensure that $\boldsymbol{\Sigma}_{x_2}$ is positive-definite after the update. Using the Fletcher-Reeves variant of the NLCG algorithm, we use the following gradient matrix:

$$\begin{aligned} \frac{\partial Q}{\partial \mathbf{U}_{x_{2c}}} = \sum_{t \in T, m \in C} \gamma_{m,t} [-\mathbf{U}_{x_{2c}} \mathbf{F}_m^T \boldsymbol{\Sigma}_{y_m}^{-1} \mathbf{F}_m \\ + \mathbf{U}_{x_{2c}} \mathbf{F}_m^T \boldsymbol{\Sigma}_{\mathbf{y}_m}^{-1} (\mathbf{y}_t - \boldsymbol{\mu}_{y_m}) (\mathbf{y}_t - \boldsymbol{\mu}_{y_m})^T \boldsymbol{\Sigma}_{\mathbf{y}_m}^{-1} \mathbf{F}_m^T] \end{aligned} \qquad (14)$$

We keep the diagonal of the matrix products in (14) to obtain an updated covariance matrix after applying the gradient in the NLCG optimization. The number of computations can also be reduced by using this assumption and, it has a big effect on the real time factor, especially when considering that adaptation is performed on every window.

Finally, the algorithm used to update the VTS model for each Gaussian m is summarized as follows:

1. Initialize the overlapping speech model parameters $\boldsymbol{\mu}_{y_m}$ and $\boldsymbol{\Sigma}_{y_m}$ using the main speaker model parameters $(\boldsymbol{\mu}_{x_{1m}0}, \boldsymbol{\Sigma}_{x_{1m}0})$ and the corrupting speaker parameters $(\boldsymbol{\mu}_{x_{2c}0}, \boldsymbol{\Sigma}_{x_{2c}0})$ using (10) and (11). $\boldsymbol{\mu}_{x_{2c}0}$ is mean vector of cluster c of corrupting speaker such that the m^{th} component of corrupting speaker GMM lies in it.
2. Update $\boldsymbol{\mu}_{x_{2_c}}$ using (13) using the current estimates of $\boldsymbol{\mu}_{y_m}, \boldsymbol{\Sigma}_{y_m}, \mathbf{G}_m$ and $\mathbf{F}_m$.
3. Update $\boldsymbol{\Sigma}_{x_2}$ using the gradient of Eq. 14 using the current estimates of $\boldsymbol{\mu}_{y_m}, \boldsymbol{\Sigma}_{y_m}, \mathbf{G}_m$ and $\mathbf{F}_m$.
4. Replace $\boldsymbol{\mu}_{x_{2c}0}$ with $\boldsymbol{\mu}_{x_{2c}}$ obtained in step 2 and recompute the overlapping speech model parameters $(\boldsymbol{\mu}_{y_m}, \boldsymbol{\Sigma}_{y_m})$.
5. Go to 2 until a number of iterations has been reached.

After running this algorithm, the $\boldsymbol{\mu}_{y_m}$ and $\boldsymbol{\Sigma}_{y_m}$ obtained in the last iteration are retained as the optimal parameters modeling the overlapping speech data $\mathbf{X}$.

3. Experiments

We ran experiments for the overlap detection and speaker diarization tasks, the latter using the VTS system output.

3.1. Overlap Detection

We evaluated the proposed approach on 10 meeting recordings from the AMI Meeting Corpus given in Table 1. The calibration threshold was optimized on a development data set consisting of 10 other meetings from AMI, also shown on Table 1. The recordings, involving 4 participants each, vary from 17 to 57 minutes in length, with a total of 11 hours of audio, of which 20% are overlapping speech. We use one of the single distant microphones channels to extract 19 MFCC every 10ms over 30ms long windows. The individual speaker models are MAP-adapted from a reference 64-component GMM using ML training and the speech from each recording.

Development Set				
EN2004a	EN2013c	IS1001c	IS1001d	IS1005a
IS1007b	IS1001c	TS3006a	TS3007c	TS3012b

Evaluation Set				
EN2003a	EN2009b	ES2008a	ES2015d	IN1008
IN1012	IS1002c	IS1003b	IS1008b	TS3009c

Table 1: AMI corpus meetings used for development and evaluation of the VTS overlapping speech detection system.

We measure the precision and recall performances as well as the overlap detection error, defined as the sum of false alarms and miss errors in the whole recording over the number of labeled speaker overlap time. Note that this measure can take values over 100%, since the labeled overlap time is much shorter than the recording length.

The system has some tuning parameters like analysis windowsize W, clustering threshold η and phase-factor α. Based on the studies done in [17], we use $W = 3.2s$ and $\eta = 1$ frame. Note that, although α is the cosine of the angle between the two speaker short-term spectra, it is here considered to be constant with the same value for all meetings. We experiment with varying values of α to evaluate overlap detection.

3.1.1. Results

The overlap detection framework can either use the oracle speaker segmentations or the segmentation output of an automatic speaker diarization system to train the initial single speaker GMM. As discussed and shown in [17], the system using oracle segmentations begins with the purest speaker GMMs we can obtain and it largely outperforms the system using inaccurate single speaker GMM. For the current work, we focus on the case where speaker GMM are trained using the oracle segmentations.

First, we study the effect of the phase-factor α on the overlap detection performance, shown on Table 2. We observe the same trend reported in [11] for the automatic speech recognition task, with performance asymptotically improving with increasing values of α w.r.t setting the phase factor to 0. These results suggest that VTS modeling can be improved by considering cross-speaker correlation, although results also improve when modeling noisy speech in [11], where noise and speech can be assumed to be more uncorrelated. In any case, since the

gains obtained with α over 1 are minor we take the value of $\alpha = 1$ for further experiments.

Phase-Factor α	**Prec./Rec.** (%)	**F meas.**	**Error (%)**
-1.0	60.2/16.5	.259	94.39
-0.5	59.8/11.1	.187	96.37
0.0	65.7/41.8	.511	80.03
0.5	66.1/43.1	.522	79.02
1.0	65.7/44.6	.531	78.70
1.5	65.6/44.8	.532	78.68
2.0	65.4/45.3	.535	78.62

Table 2: Precision, Recall, F-measure and Overlap Detection Error for varying values of the phase factor α.

Table 3 compares the results of experiments using the phase factor, covariance adaptation and all possible overlap hypotheses with the best results obtained in our previous work [17].

Although updating the covariance matrices gives significant relative gains in likelihood terms, the F-ratio and the overlap detection error slightly increased. On the other side, covariance adaptation is computationally heavy, scaling the real time factor by 10 in our implementation.

We also ran an experiment where we consider all N^2 speech hypotheses (N being the number of total speakers) instead of picking a main speaker as described in section 2.1. The results show that, although we compute N times more hypotheses, performance stays the same. Picking a main speaker and finding a corrupting speaker is as good a strategy as considering all possible hypotheses.

System	**Prec./Rec.** (%)	**F meas.**	**Error (%)**
Diarization seg.	51.0/17.5	.260	99.32
Oracle seg.	65.7/41.8	.510	80.05
Covariance adaptation	66.4/38.6	.488	80.89
α=1.0, N hyp.	65.7/44.6	.531	78.70
α=1.0, N^2 hyp.	65.6/44.6	.531	78.67

Table 3: Precision, Recall, F-measure and Overlap Detection Error for several systems. In the first block are experiments using the diarization output and the oracle segmentations to train the speaker models. In the second block are experiments using covariance adaptation, N and N^2 hypotheses.

3.2. Speaker Labeling

Besides detecting overlapping speech segments, the VTS framework also identifies two speakers involved in the overlap. As discussed in the section above, we used the overlap detection framework trained using oracle segmentations to detect overlaps and label the active speakers in the overlapping segments. We ran these experiments for a set of 32 meeting recordings taken from the AMI corpus. 16 meetings were used as development set, to find the operating point of the overlap detection system that maximizes the performance of diarization. The remaining 16 meetings were used for evaluation.

For speaker labeling, we ran three experiments. First, given the diarization output providing a single speaker label for a given time instant, we assign multiple speaker labels to those regions detected as overlapping speech. Second, we generate the complete diarization output for all time instants using the winning hypothesis of the overlapping speech detection system, with models trained using the oracle segmentation. Third, we generate the diarization outputs using the overlapping speech detection system trained using the diarization output. This is the most realistic system, and it is prone to errors because of impurity in the initial speaker models.

3.2.1. Results

Table 4 compares the results of the three diarization experiments described above. As the baseline, we use the diarization output of the system described in [15], obtaining 29.2% DER. When the two speakers involved in overlapping speech regions are labeled the DER is reduced by 5% relative, down to 27.6% absolute. A very low DER of 10.44% is observed when all speakers are assigned using our approach from speaker models trained from the oracle segmentation. This confirms that the system is able to properly label speakers, from N^2 classes when the models are pure. Note that labeling is local to the window and no global strategy such as Viterbi decoding is used, as it is the case in most speaker diarization systems. However, the DER increases up to 38.6% when the speaker models are trained directly from the diarization output, highlighting the high sensitivity of the algorithm to speaker model impurity.

System	**Missed** (%)	**False Alarm %**	**DER (%)**
Baseline	10.2	2.2	29.21
Label Overlap Oracle	10.7	0.3	27.62
Label All Oracle	7.9	0.8	10.44
Label All Diar.	13.2	0.1	38.62

Table 4: Precision, Recall, F-measure and Overlap Detection Error for systems using different ways of labeling speakers in the diarization output. We show experiments labeling overlap regions only using speaker models trained with the oracle segmentation (Label Overlap Oracle). We also consider labeling all the data using oracle (Label All Oracle) and diarization output (Label All Diar.) segmentations.

4. Conclusions

In this work we have shown the potential of the multi-class Vector Taylor Series (VTS) framework for overlap detection and speaker labelling in meeting recordings. This framework accounts for multiple sounds being uttered by two overlapping speakers. In this paper, we modeled the correlation between the sounds uttered by two speakers, decreasing the overlap detection error and F-measure. We also found that assuming as many hypotheses as the number of speakers performed as good as considering all possible pairs of hypotheses. For the speaker diarization task, we observed a large performance gap between the diarization output with single-speaker labels and labeling the data using multiple speakers using the oracle models. When using these models, the VTS technique is able to reduce this gap. A future line of work will focus on purifying the speaker models generated from the diarization output to obtain a more realistic system.

5. References

[1] Shriberg, Elizabeth, Andreas Stolcke, and Don Baron. "Observations on overlap: findings and implications for automatic processing of multi-party conversation." INTERSPEECH. 2001.

[2] Boakye, K.; Trueba-Hornero, B.; Vinyals, O.; Friedland, G., "Overlapped speech detection for improved speaker diarization in multiparty meetings," Acoustics, Speech and Signal Processing,

2008. ICASSP 2008. IEEE International Conference on , vol., no., pp.4353,4356, March 31 2008-April 4 2008

[3] Vipperla, Ravichander, et al. "Speech overlap detection and attribution using convolutive non-negative sparse coding." Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE, 2012.

[4] Geiger, Jurgen T., et al. "Speech overlap detection using convolutive non-negative sparse coding: New improvements and insights." Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European. IEEE, 2012.

[5] Geiger, Jurgen T., Florian Eyben, Nicholas Evans, Bjrn Schuller, and Gerhard Rigoll. "Using Linguistic Information to Detect Overlapping Speech." INTERSPEECH 2013.

[6] Geiger, Jurgen T., Florian Eyben, Bjrn Schuller, and Gerhard Rigoll. "Detecting Overlapping Speech with Long Short-Term Memory Recurrent Neural Networks." INTERSPEECH 2013.

[7] Yella, Sree Harsha, and Fabio Valente. "Speaker diarization of overlapping speech based on silence distribution in meeting recordings." In INTERSPEECH. 2012.

[8] Yella, Sree Harsha and Bourlard, Herv. "Improved Overlap Speech Diarization of Meeting Recordings using Long-term Conversational Features" in ICASSP, 2013.

[9] P.J.Moreno. "Speech Recognition in Noisy Environments". Ph.D. Thesis, Carnegie Mellon University, 1996.

[10] Lei, Yun, Luk Burget, and Nicolas Scheffer. "A Noise Robust i-Vector Extractor Using Vector Taylor Series for Speaker Recognition" in ICASSP 2013.

[11] Li, Jinyu, Li Deng, Dong Yu, Yifan Gong, and Alex Acero. "A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions."Computer Speech & Language 23,no.3(2009):389-405.

[12] Zhao, Yong, and Biing-Hwang Juang. "On noise estimation for robust speech recognition using vector Taylor series." Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. IEEE, 2010.

[13] Reynolds, Douglas A., Thomas F. Quatieri, and Robert B. Dunn. "Speaker verification using adapted Gaussian mixture models." Digital signal processing 10.1 (2000): 19-41.

[14] Anguera, Xavier, Wooters, Chuck, Hernando, Javier, "Acoustic beamforming for speaker diarization of meetings", IEEE Transactions on Audio, Speech and Language Processing September 2007, volume 15, number 7, pp.2011-2023.

[15] Vijayasenan, Deepu, Valente, Fabio, Bourlard Hervé, "An Information Theoretic Approach to Speaker Diarization of Meeting Data", IEEE Transactions on Audio Speech and Language Processing, September 2009, volume 17, number 7, pp.1382-1393.

[16] Jonathan Richard Shewchuk, "An Introduction to the Conjugate Gradient Method Without the Agonizing Pain.

[17] Pranay Dighe, Marc Ferras, Hervé Bourlard, "Modeling Overlapping Speech using Vector Taylor Series", Submitted to the IEEE Workshop Speaker Odyssey 2014, notification of acceptance on April 10th. Available for this review on `http://www.idiap.ch/%7Emferras/Odyssey2014.pdf`

.