Modeling Overlapping Speech using Vector Taylor Series

Pranay Dighe, Marc Ferràs and Hervé Bourlard

Idiap Research Institute, CH-1920 Martigny, Switzerland

Abstract

Current speaker diarization systems typically fail to succesfully assign multiple speakers speaking simultaneously. According to previous studies, overlapping errors account for a large proportion of the total errors in multi-party speech diarization. In this work, we propose a new approach using Vector Taylor Series (VTS) to obtain overlapping speech models assuming individual speaker models are available, e.g. from the diarization output. We extend the VTS framework to use multiple acoustic classes to account for the non-stationarity of corrupting speaker speech. We propose a system using multi-class VTS to detect single-speaker and two-speaker overlapping speech as well as the speakers involved. We show the effectivity of the approach on distant microphone meeting data, especially with the multi-class approach performing at the state-of-the-art.

1. Introduction

Speaker diarization is the task of determining "who spoke when" as well as the number of speakers in a recording. State-of-the-art systems are known to struggle to assign speech segments to the right speakers on multi-party spontaneous speech such as meetings, especially in the presence of overlapping speech. For the cepstral features, the overlapping speech can be modeled as a set of linear and non-linear operations, whereas it simply boils down to a linear combination of the individual sources in the signal and even spectral domains. Unfortunately, statistical modeling is more challenging in these domains as well.

Speaker diarization systems are affected by the presence of overlapping speech in two different ways [1]. First, the speaker segments output by the diarization system are used to train pure speaker models and, since they are corrupted with overlapping speech, the resulting models are less precise. Second, the system is asked to attribute a single speaker label to a segment which actually contains overlapping speech from two or more speakers. Given that around 20% [1] time of meetings have speech overlaps, both types of errors contribute to a significant increase in Diarization Error Rate (DER). This provides a genuine motivation to devise approaches to detect and model overlapping speech.

Overlapping speech detection has been addressed by previous studies. A HMM-based segmenter is used in [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. Assigning the highest scoring speakers to the overlapping speech segments output by the diarization system improves the DER performance. Another family of approaches [3], [4], [5] and [6] use convolutive non-negative sparse coding (CNSC) to detect overlap, where the bases learnt for each speaker are concatenated into a single basis matrix and the decomposition provides the activity of each speaker for each frame. A wide variety of features have been used including cepstral features, energy, jitter, shimmer, CNSC features and even linguistic features in a probabilistic framework. The work in [6] uses the same set of features in a Long Short-Term Memory (LSTM). Two other approaches use knowledge of the silence distribution in meeting recordings [7] and long-term conversational features, namely the distribution of overlap occurrence around speaker changes [8]. Similar goals have been addressed for multi-speaker speech recognition using graphical models in [9]. Otherwise, techniques derived from speaker identification using Gaussian mixture models have also been proposed [10]. In this paper, we propose to model the feature space of two simultaneous speakers using the Vector Taylor Series (VTS) technique.

The paper is organized as follows: Section 2 describes the VTS approach to noisy speech modeling and how it can be applied to model overlapping speech. We give details about the algorithm and parameter estimation. Section 3 focuses on how VTS can be used for the task of overlapping speech detection. We extend de VTS framework to deal with multiple acoustic classes in Section 4. Sections 5 provides experimental results of the approach on meeting data. Conclusions are given in Section 6.

2. VTS Modeling of Overlapping Speech

The standard VTS approach estimates a noisy speech model from a clean speech model and some statistics about additive and convolutional noises assumed to be present in the noisy speech signal. The process of corrupting a clean speech signal $x(t)$ with additive, $n(t)$, and convolutional, $h(t)$, noises can be written by

$$y(t) = x(t) * h(t) + n(t) \tag{1}$$

The corresponding relation in the feature domain, where we do the modeling, is much more complex [11]. For MFCC features this becomes

$$\mathbf{y} = \mathbf{x} + \mathbf{h} + \mathbf{C}\ln(1 + \exp(\mathbf{C}^{-1}(\mathbf{n} - \mathbf{x} - \mathbf{h}))) \tag{2}$$

where $\mathbf{y}$,$\mathbf{x}$,$\mathbf{n}$ and $\mathbf{h}$ are cepstral vectors for noisy speech, clean speech, additive, and convolutional noise, respectively. $\mathbf{C}$ and is the discrete cosine transform (DCT) matrix and $\mathbf{C}^{-1}$ is its pseudo-inverse.

In essence, VTS uses a multivariate linear approximation of (2) to estimate a Gaussian Mixture Model (GMM) for the corrupted features $\mathbf{y}$.

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.

Overlapping speech is actually the superposition of two or more individual-speaker speech signals. For the two speaker case, an equation analogous to (1) can be derived as

$$y(t) = x_1(t) + x_2(t) \quad , \tag{3}$$

with $x_1(t)$ and $x_2(t)$ being the speech signals spoken by speaker 1 and 2 respectively. The expression

$$\mathbf{y}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1 + g(\mathbf{x}_2 - \mathbf{x}_1) \tag{4}$$

with

$$g(\mathbf{x}_2 - \mathbf{x_1}) = \mathbf{C}\ln(1 + \exp(\mathbf{C}^{-1}(\mathbf{x}_2 - \mathbf{x}_1)) \tag{5}$$

is the corresponding relation in the MFCC feature domain. Note that the convolutional noise term $\mathbf{h}$ has been neglected and the corrupting speaker term $\mathbf{x_2}$, which is additive just like noise, has been introduced. Although the acoustic structure of speech is expected to be much richer than that of noise, the latter is typically modeled using a single Gaussian, thus assuming stationarity. We focus on the overlapping speech case from here on, first by using a single Gaussian to model the corrupting speaker and using multiple classes later in Section 4.

In this work, 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, and we let the VTS technique estimate the corrupted GMM parameters assuming these two sources are overlapping.

2.1. Approximating the corrupted model using VTS

In the context of overlapping speech, we assign noisy and clean speeches to main and corrupting speaker speeches in our approach. Keeping these assumptions in mind, we let $\boldsymbol{\mu}_{y_m}$ and $\boldsymbol{\mu}_{x_{1m}}$ be the mean vectors of the m_{th} Gaussian component of the corrupted and main speaker GMM respectively. The mean of the single Gaussian representing the corrupting speaker is denoted by $\boldsymbol{\mu}_{x_2}$. The first-order VTS expansion of (4) for Gaussian m w.r.t. vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ around the point ($\boldsymbol{\mu}_{x_{1m}0}$,$\boldsymbol{\mu}_{x_20}$) is

$$\begin{aligned}\mathbf{y} \approx \boldsymbol{\mu}_{x_{1m}0} + g(\boldsymbol{\mu}_{x_20} - \boldsymbol{\mu}_{x_{1m}0}) + \mathbf{G}_m(\mathbf{x}_1 - \boldsymbol{\mu}_{x_{1m}0}) \\ + \mathbf{F}_m(\mathbf{x}_2 - \boldsymbol{\mu}_{x_20})\end{aligned} \tag{6}$$

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 ($\boldsymbol{\mu}_{x_{1m}0}$,$\boldsymbol{\mu}_{x_20}$), that is,

$$\mathbf{G}_m = \mathbf{C}\text{diag}(\frac{1}{1 + \exp(\mathbf{C}^{-1}(\boldsymbol{\mu}_{x_20} - \boldsymbol{\mu}_{x_{1m}0}))})\mathbf{C}^{-1} \tag{7}$$

$$\mathbf{F}_m = \mathbf{I} - \mathbf{G}_m \tag{8}$$

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

$$\boldsymbol{\mu}_{y_m} \approx \boldsymbol{\mu}_{x_{1m}0} + g(\boldsymbol{\mu}_{x_20} - \boldsymbol{\mu}_{x_{1m}0}) \tag{9}$$

Similarly, using $\boldsymbol{\Sigma}_{x_{1m}}$ and $\boldsymbol{\Sigma}_{x_2}$ 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

$$\boldsymbol{\Sigma}_{y_m} \approx \mathbf{G}_m\boldsymbol{\Sigma}_{x_{1m}}\mathbf{G}_m^T + \mathbf{F}_m\boldsymbol{\Sigma}_{x_2}\mathbf{F}_m^T \tag{10}$$

2.2. 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 (9) and (10), the expectation-maximization (EM) algorithm iteratively finds estimates of $\boldsymbol{\mu}_{y_m}$ that further maximize the likelihood function

$$Q = \sum_{t\in T}\sum_{m\in M}\gamma_{t,m}\log(p(\mathbf{x}_t|\boldsymbol{\mu}_{y_m}, \boldsymbol{\Sigma}_{y_m})) \quad , \tag{11}$$

eventually converging to a local maximum.

Since VTS can be a resource consuming technique, only $\boldsymbol{\mu}_{x_2}$ is optimized in thsi work. In turn, $\boldsymbol{\mu}_{y_m}$, $\mathbf{F}_m$, $\mathbf{G}_m$ and $\boldsymbol{\Sigma}_{y_m}$ are updated accordingly. If we replace the expectation of (6) into (11) and then differentiate w.r.t. $\boldsymbol{\mu}_{x_2}$, the update equation for $\boldsymbol{\mu}_{x_2}$ becomes

$$\begin{aligned}\boldsymbol{\mu}_{x_2} = \boldsymbol{\mu}_{x_20} + \{\sum_{t\in T, m\in M}\gamma_{m,t}\mathbf{F}_m^T\boldsymbol{\Sigma}_{y_m}^{-1}\mathbf{F}_m\}^{-1} \\ \times\{\sum_{t\in T, m\in M}\gamma_{m,t}\mathbf{F}_m^T\boldsymbol{\Sigma}_{\mathbf{y}_m}^{-1}(\mathbf{y}_t - \boldsymbol{\mu}_{y_m})\}\end{aligned} \tag{12}$$

where $\boldsymbol{\mu}_{x_20}$ is the previous estimate of the corrupting mean vector.

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 $(\boldsymbol{\mu}_{x_20}, \boldsymbol{\Sigma}_{x_20})$ using (9) and (10). $\boldsymbol{\mu}_{x_20}$ is taken as the mean vector of the Gaussian component with hightest average posterior probability over $\mathbf{X}$, and $\boldsymbol{\Sigma}_{x_20}$ as the corresponding covariance matrix.
2. Update $\boldsymbol{\mu}_{x_2}$ using (12) using the current estimates of $\boldsymbol{\mu}_{y_m}$, $\boldsymbol{\Sigma}_{y_m}$, $\mathbf{G}_m$ and $\mathbf{F}_m$.
3. Replace $\boldsymbol{\mu}_{x_20}$ with $\boldsymbol{\mu}_{x_2}$ obtained in step 2 and recompute the overlapping speech model parameters $(\boldsymbol{\mu}_{y_m}, \boldsymbol{\Sigma}_{y_m})$.
4. 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. Overlapping Speech Detection System

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 segmentation or after automatic speaker diarization. We use the data collected for the Augmented Multiparty Interaction (AMI) project to evaluate the OSD system. These data consist of meeting audio of around 30 minutes long involving four participants recorded using far-field microphones. We preprocess the audio for beamforming using the BeamformIt toolkit [16] and also for Speech Activity Detection (SAD) so that the OSD system can focus on rather homogeneous speech segments, usually spoken by one speaker with overlaps and interruptions from other speaker. Together with the VTS technique described above, this setup allows the system to detect non-speech, single-speaker speech and overlapping speech regions.

The overlap detection, involving the modeling and decision steps, is done on a window sliding over each of the speech segments mentioned above. For each window, a set of hypothesis tests are performed comparing 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 with N being the number of speakers. We choose 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 $N-1$. In short, if speaker i is the main speaker, we obtain the set of likelihood ratios

$$\frac{p(\mathbf{X}|\mathcal{S}_{1,i})}{p(\mathbf{X}|\mathcal{S}_i)}, \ldots, \frac{p(\mathbf{X}|\mathcal{S}_{i-1,i})}{p(\mathbf{X}|\mathcal{S}_i)}, 1, \ldots, \frac{p(\mathbf{X}|\mathcal{S}_{N,i})}{p(\mathbf{X}|\mathcal{S}_i)} \tag{13}$$

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. In this work, we model the former using VTS adaptation and the latter using Maximum A Posteriori (MAP) adaptation [15], as

- **Overlap:** For the speaker pairs j, i in (13), 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 as

$$\hat{\boldsymbol{\mu}}_{x_m} = \alpha E_m[\mathbf{x}] + (1-\alpha)\boldsymbol{\mu}_{x_m} \tag{14}$$

  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.

4. Multi-Class VTS

The stark difference between overlapping and noisy speech is that the former is essentially non-stationary. Within the span of a window there might be several sounds being uttered by both main and corrupting speakers. This point not addressed by the standard VTS approach motivated us to look for an alternative way to represent the corrupting speaker.

In this work, we propose a multi-class version of the VTS framework to model multiple phonemes uttered by the corrupting speaker. The acoustic space of the corrupting speaker is clustered into multiple classes and VTS is used to adapt the mean vectors of the representatives of each cluster separately.

We start assuming that all the Gaussian components are observed in the data. 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 squared Euclidean distance[1]. 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. Using a large enough threshold η avoids singularity issues during VTS estimation.

[1]The squared euclidean distance was used here for the sake of speed.

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.

A few modifications are required to deal with multiple clusters in VTS. $\boldsymbol{\mu}_{x_2}$ for the single-gaussian case becomes cluster-dependent, i.e. $\boldsymbol{\mu}_{x_{2c}}$. Furthermore, we now iterate over the set of Gaussians, C, assigned to cluster c. The VTS estimation equation (12) becomes

$$\begin{aligned}\boldsymbol{\mu}_{x_{2c}} = \boldsymbol{\mu}_{x_{2c}0} + \{\sum_{t\in T, m\in C}\gamma_{m,t}\mathbf{F}_m^T\boldsymbol{\Sigma}_{y_m}^{-1}\mathbf{F}_m\}^{-1} \\ \times\{\sum_{t\in T, m\in C}\gamma_{m,t}\mathbf{F}_m^T\boldsymbol{\Sigma}_{\mathbf{y}_m}^{-1}(\mathbf{y}_t - \boldsymbol{\mu}_{y_m})\}\end{aligned} \tag{15}$$

The VTS approximation equations given in Section 2 are shown in Table 1.

Equations for Multi-Class VTS Approximation
$\boldsymbol{\mu}_{y_m} \approx \boldsymbol{\mu}_{x_{1m}0} + g(\boldsymbol{\mu}_{x_{2c}0} - \boldsymbol{\mu}_{x_{1m}0}) + \mathbf{G}_m(\boldsymbol{\mu}_{x_{1m}} - \boldsymbol{\mu}_{x_{1m}0}) + \mathbf{F}_m(\boldsymbol{\mu}_{x_{2c}} - \boldsymbol{\mu}_{x_20})$
$\mathbf{G}_m = \mathbf{C}.\text{diag}(\frac{1}{1+\exp(\mathbf{C}^{-1}(\boldsymbol{\mu}_{x_{2c}0} - \boldsymbol{\mu}_{x_{1m}0}))}).\mathbf{C}^{-1}$
$\mathbf{F}_m = \mathbf{I} - \mathbf{G}_m$
$\boldsymbol{\mu}_{y_m} \approx \mu_{x_{1m}0} + g(\boldsymbol{\mu}_{x_{2c}0} - \boldsymbol{\mu}_{x_{1m}0})$
$\boldsymbol{\Sigma}_{y_m} \approx \boldsymbol{G}_m\boldsymbol{\Sigma}_{x_{1m}}\mathbf{G}_m^T + \mathbf{F}_m\boldsymbol{\Sigma}_{x_{2c}}\mathbf{F}_m^T$

Table 1: Equations for Section 2 modified for the Multi-class VTS Approximation

5. Experiments

5.1. Experimental Setup

We evaluated the proposed approach on 10 meeting recordings from the AMI Meeting Corpus given in Table 2. We optimize the calibration threshold on a development data set consisting of 10 meetings from AMI also shown on Table 2. 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 GMM trained using the speech from each recording using maximum likelihood estimation and 64 Gaussian components.

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 2: AMI corpus meetings used for development and evaluation of VTS framework

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.

5.2. Results

We ran two sets of experiments. The first set uses the standard VTS algorithm to perform overlapping detection. We used the oracle speaker segmentation, manually annotated, to train the individual speaker GMM. These are the purest we can obtain and we expect VTS to provide the best overlap detection resuts as well. Since this is not a realistic choice for a practical system, we also ran experiments that use the segmentation output of an automatic speaker diarization system [17]. The resulting GMM are then prone to errors that make the overlapping speech modeling and detection less precise. For both oracle and diarization segmentations we test both standard and multi-class VTS modeling.

For the standard VTS system we focus on the evaluation of the effect of the window length on detection performance. In general terms, the longer the window the more data are available for modeling. Nonetheless, long windows are likely to be coarsely modeled if few Gaussians are used. Too long a window can also make decisions not local. In this set of experiments we explored the window lengths 0.4s, 0.8s, 1.6s and 3.2s with the results being shown in Table 3. The detection error is minimized for a window length of 3.2s. Note that multiple sounds can be uttered in 3.2s and it is still optimal to model the data using a single Gaussian component. On one side, this may reflect that long overlap durations are present in the corpus, thus minimizing the detection error when the window length is matched to the average overlap length. On the other side, a long window ensures that enough data is available to train the statistical models and that this is preferred over being accurate in the time domain for this corpus. These results seem reasonable in terms of precision but at the expense of low recall, which results in a very low F-measure. Figure 1 shows precision-recall curves for the explored system setups supporting the fact that a long analysis window of 3.2s results in the best performance across the majority of operating points.

WindowSize (s)	**Prec./Rec. (%)**	**F meas.**	**Error (%)**
VTS 0.4s	57.0/7.20	.127	98.3
VTS 0.8s	57.1/12.0	.198	97.0
VTS 1.6s	54.0/23.8	.330	96.4
VTS 3.2s	55.0/20.5	.299	**96.2**

Table 3: Precision, Recall, F-measure and Overlap Detection Error for 0.4s, 0.8s, 1.6s and 3.2s of analysis window in the standard VTS approach.

The second set of experiments explores the multi-class VTS approach and compares it to standard VTS for a window size of 3.2s, as found in the previous experiments. As shown in Table 4, the multi-class VTS approach (MC-VTS), largely outperforms standard VTS when at least one frame is enforced for each cluster, i.e. $\eta = 1$. This setup results in optimal overlap detection error with an average number of 24.7 classes out of the 64 initial Gaussian components used during modeling. Using larger thresholds resulted in more reasonble number of clusters found, but actual detection performance decreased as well.

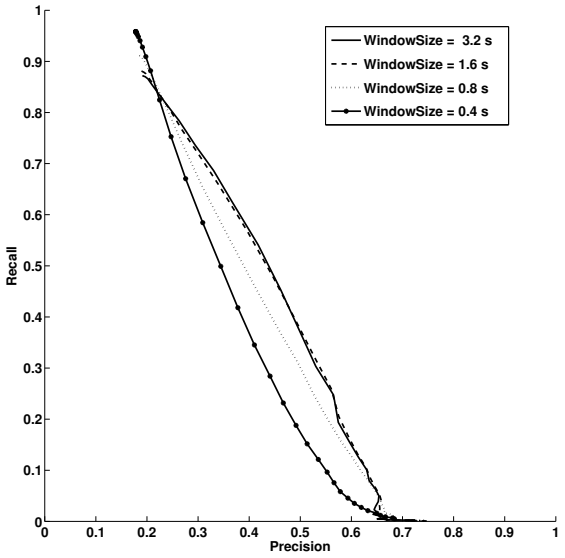

Figure 1: Precision and recall performance of standard VTS using different lengths for the analysis window. The output using the segmentation output by a diarization system is also shown.

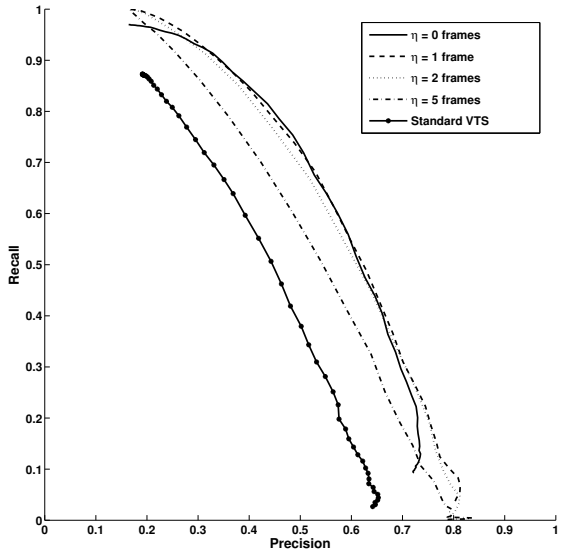

Figure 2: Precision and recall performance of multi-class VTS using different thresholds η as the minimum number of frames allowed for each cluster. The output using the segmentation output by a diarization system and $\eta = 1$ is also shown.

Gains up to 16% and 70% relative in precision and recall respectively are observed for an overlap detection error reduction of 15%. These results are in the line, or even better in terms of F-measure, when compared to other state-of-the-art approaches [5, 6, 8] found in the literature. However, these experiments were run using oracle alignments to train the individual speaker

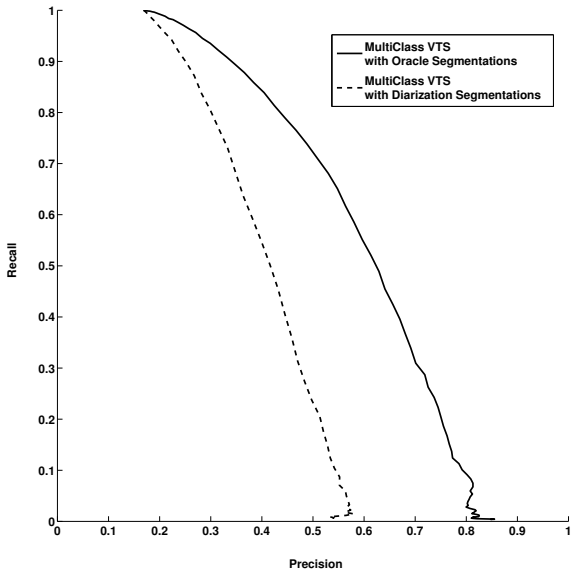

Figure 3: Precision and recall performance of the top performing multi-class VTS system using oracle and diarized speaker segmentations to train individual speaker models.

models, which makes the comparison not fair. It still allows us to evaluate the success of modeling overlapping speech using VTS. Figure 2 shows precision-recall curves for theses systems revealing a considerable performance gap between multi-class VTS setups and standard VTS.

System	η	#Class	Prec./Rec. (%)	F meas.	Error (%)
VTS	1	-	55.0/20.5	.299	96.2
MC-VTS	0	64	63.9/38.7	.482	84.0
MC-VTS	1	24.7	65.7/41.8	.510	**80.0**
MC-VTS	2	18.8	66.0/38.5	.486	81.3
MC-VTS	5	10.9	66.7/26.5	.379	86.7
MC-VTS Dia	1	24.7	51.0/17.5	.260	99.3

Table 4: Precision, Recall, F-measure and Overlap Detection Error for 0, 1 2 and 5 frames as the minimum number of frames allowed for each cluster in the Multi-class VTS approach. The first row shows the performance of the standard VTS approach, for reference. The last row gives results for individual speaker models trained using the segmentation output by a diarization system.

We also made a quick assessment of how using the diarization output to train the individual speaker models affects overlapping speech detection performance. Figure 3 shows precision-recall curves for these systems, highlighting the sensitivity of the multi-class VTS technique to the accuracy of the individual speaker models. Errors in the estimation of the number of speakers in the recording and the impurity of the speaker models might account for these results.

6. Conclusions

We proposed a new approach to overlapping speech modeling based on the Vector Taylor Series (VTS) framework that has been used to model noisy speech for the automatic speech recognition task. We have extended the VTS framework to account for the corrupting speaker uttering non-stationary sounds, i.e. multiple acoustic classes, as opposed to stationary noise within the analysis window. We used both standard and multi-class VTS to model different overlapping speech hypotheses to build an overlapping speech detection system. Rather long analysis windows of 3.2s were found to be optimal according to error detection. The multi-class VTS approach significantly outperformed the standard VTS providing a relative error reduction of 15% with relative gains of up to 40% recall and 70% precision. Alhoutgh the multi-class VTS approach has shown effective at modeling overlapping speech it still relies on the purity of individual speaker models to be state-of-the-art. Further work will focus on improving the diarization output so that purer models can be obtain for use with VTS.

7. 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 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 2013.

[9] J. Hershey, S. Rennie, P. Olsen, and T. Kristjansson, Superhuman multi-talker speech recognition: A graphical modeling approach, Elsevier Comp. Speech and Lang., vol. 24, no. 1, pp. 4566, Jan 2010.

[10] R. Saeidi, P. Mowlaee, T. Kinnunen, Z. - H. Tan, M. G. Christensen, S. H. Jensen, and Fra, "Signal-to-Signal Ratio Independent Speaker Identification for Co-channel

Speech Signals", Pattern Recognition (ICPR), 2010 20th International Conference on, pp. 4565 -4568, aug., 2010.

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

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

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.