UNSUPERVISED LEARNING OF OVERLAPPED SPEECH MODEL PARAMETERS FOR MULTICHANNEL SPEECH ACTIVITY DETECTION IN MEETINGS

Kornel Laskowski and Tanja Schultz

kornel|tanja@cs.cmu.edu
interACT, Carnegie Mellon University, Pittsburgh, PA, USA

ABSTRACT

The study of meetings, and multi-party conversation in general, is currently the focus of much attention, calling for more robust and more accurate speech activity detection systems. We present a novel multichannel speech activity detection algorithm, which explicitly models the overlap incurred by participants taking turns at speaking. Parameters for overlapped speech states are estimated during decoding by using and combining knowledge from other observed states in the same meeting, in an unsupervised manner. We demonstrate on the NIST Rich Transcription Spring 2004 data set that the new system almost halves the number of frames missed by a competitive algorithm within regions of overlapped speech. The overall speech detection error on unseen data is reduced by 36% relative.

1. INTRODUCTION

The study of meetings, and multi-party conversation in general, is currently the focus of much attention. As an almost canonical form of naturally occurring speech, meetings pose many new challenges, but also new horizons for applications. The majority of these implicitly assume error-free automatic speech-to-text transcription. Speech recognition, in turn, calls for speech activity detection algorithms capable of segmenting multichannel audio into contiguous intervals of foreground speech.

Speech activity detection in meetings was initially approached as a monochannel detection task, where audio frames in each channel were classified as either speech or non-speech, independently of other channels. Results were mostly unacceptable due to a high degree of acoustic inter-channel coupling, even with close-talking microphones. A multichannel extension to a single-channel, cepstral feature, hidden Markov model detection system was first introduced by Pfau et al [1], in which cross-correlation was used in a post-processing pass to discard hypothesized speech which could be accounted for by activity on other channels; Stolcke et al recently improved on the employed cross-correlation measure [2]. Feature selection specific to the task of multichannel speech and crosstalk classification was pioneered by Wrigley et al [3], and subsequently explored in the context of an ergodic hidden Markov model (eHMM) system [4]. In our own work on meetings, we introduced a multichannel speech activity detector [5] based only on cross-correlation. Most recently, the NIST Rich Transcription (RT) 2005 Spring Evaluation [6] produced much new activity and interest in the field.

In this paper we propose a novel algorithm, which combines our cross-correlation-based multispeaker speech activity detection [5] with the idea of overlapping speech states. In constrast to [3] in which the eHMM consisted of 4 states (foreground speech, non-speech, background speech, and crosstalk), we present here an eHMM with 2^K states, specifying every possibly combination of speech and non-speech for each of K participants. This eliminates the need for rule-based channel hypothesis recombination. The system presents much promise for tracking the turn-taking behavior and overlap patterns of individuals.

2. DATA

In all experiments presented in this paper, we use the NIST Rich Transcription 2004 development and evaluation datasets [6]. Each consists of eight 10-minute meeting excerpts, two collected at each of 4 sites: NIST, LDC, ICSI, and CMU. The number of participants varies between 3 and 7 in the development set; the evaluation set contains a meeting excerpt with 10 participants.

3. ANALYSIS OF BASELINE

As part of the NIST RT04s evaluation, we presented a parameter-free multichannel speech activity detection algorithm [5]; speech ($\mathcal{S}$) or non-speech ($\mathcal{N}$) were hypothesized independently for each participant, but based on a cross-correlation measure derived by considering all K microphone channels simultaneously:

$$\hat{\mathbf{\Psi}}_k \quad = \quad \begin{cases} \mathcal{S} & \text{if } \sum_{j \neq k} \log_{10} \left(\frac{\max \phi_{jk}}{\phi_{jj}[0]} \right) > 0 \\ \mathcal{N} & \text{otherwise} \end{cases} \tag{1}$$

where ϕ_{jk} is the cross-correlation between channels j and k, and $\phi_{jj}[0]$ is the channel energy. When exactly one source $y[n]$ is active in the meeting room, the response of each microphone under very simplistic assumptions is

$$x_k[n] \quad = \quad \frac{G_k}{d_{k,y}} y[n] + n_k[n] \tag{2}$$

where G_k, n_k and $d_{k,y}$ are the k-microphone-specific gain, uncorrelated noise and distance to source y, respectively. Under these assumptions, and by further assuming that the gains of all the microphones are equal, the criterion in Equation 1 reduces to

$$d_{k,y} < \left(\prod_{j \neq k} d_{j,y} \right)^{\frac{1}{K-1}} \tag{3}$$

In other words, the algorithm identifies as speech all those channel k frames for which the source is closer to microphone k than to the geometric mean of the distances to all the remaining microphones. While there exist other sound sources in the room, the overwhelming majority of sources that meet this criterion are the mouths of the microphone wearers, and the overwhelming majority of the audible output of the latter appears to be speech.

To reduce the effect of contact and breath noise, channel audio was pre-emphasized using a first-order high-pass IIR filter, $1 - z^{-1}$ and framed at 300ms intervals; the detector's output was smoothed by closing short non-speech gaps, discarding short speech blips, and padding contiguous speech activity intervals with a half-second collar. Parameters governing the smoothing were chosen empirically to equalize the miss and false alarm rates on the development set. We have found in subsequent experiments, however, that system performance can be significantly improved if these parameters are optimized jointly with the frame step and the frame size. These very simple modifications produce an improved baseline (B1) system whose performance we show alongside that of the original RT04s baseline (B0) in Table 1.

Baseline	pre-smoothing			post-smoothing		
Algorithm	MS	FA	1 − F	MS	FA	1 − F
B0 (orig)	32.90	4.18	21.07	14.41	14.17	14.29
B1 (impr)	37.63	3.77	24.32	11.70	12.18	11.94

Table 1. *Error rates (%) on the RT04s devset using the baseline detectors, both prior to and following their respective smoothing passes (MS = miss rate, FA = false alarm rate, 1–F = complement of F-score).*

Closer analysis of the algorithm reveals that, while only 12% of the development set is transcribed as containing 2 simultaneous speakers, more than 35% of the total miss rate occurs in these regions. In particular, only in 6% of those frames are both speakers hypothesized as speaking; in 69% of them only one of the correct speakers is detected, and in 24% neither speaker is detected. The latter is due to the absence of clear maxima in cross-channel correlations in the presence of multiple sources. However, the algorithm exhibits a surprisingly low false alarm rate. The remainder of this paper is concerned with addressing the large miss rate by training model-based detectors on the high-precision labels provided by the B1 baseline.

4. MODEL-BASED DETECTION FRAMEWORK

We investigate model-based detection of multispeaker speech activity in the hope that models generalize to missed sections of a meeting which are temporally adjacent to sections classified correctly by the baseline. As always, we aim to select the multispeaker assignment of speech and nonspeech, whose posterior probability given the multichannel signal is the supremum for the whole meeting:

$$\begin{aligned}
\hat{\mathbf{\Psi}} &\equiv \left\{\hat{\mathbf{\Psi}}_1, \ldots, \hat{\mathbf{\Psi}}_K\right\} \\
&= \operatorname*{argmax}_{\mathbf{\Psi}} P\left(\mathbf{\Psi} \mid \mathbf{X}\right) && (4) \\
&\doteq \operatorname*{argmax}_{\mathbf{\Psi}} \prod_n P\left(\mathbf{\Psi}\left[n\right] \mid \mathbf{X}\left[n\right]\right) && (5) \\
&\doteq \operatorname*{argmax}_{\mathbf{\Psi}} \prod_n \prod_k P\left(\psi_k\left[n\right] \mid \mathbf{X}\left[n\right]\right) && (6) \\
&\doteq \operatorname*{argmax}_{\mathbf{\Psi}} \prod_n \prod_k P\left(\psi_k\left[n\right] \mid \mathbf{x}_k\left[n\right]\right) && (7)
\end{aligned}$$

where in Equation 5 we make an assumption of conditional independence between states across frames, in Equation 6 we further assume independence between speakers within frames, and in Equation 7 we assume, given each participant's channel, the conditional independence of participant state on all other channels. We explore each of these assumptions in reverse order.

For simplicity, we use a single Gaussian to model states. The mean and covariance are estimated directly,

$$\hat{\mu}_m = \frac{M_{1,m}}{M_{0,m}} \quad (8)$$

$$\hat{\mathbf{\Sigma}}_m = \frac{M_{2,m}}{M_{0,m}} - \left(\frac{M_{1,m}}{M_{0,m}}\right)^T \left(\frac{M_{1,m}}{M_{0,m}}\right) \quad (9)$$

where the terms $M_{i,m}$ are the ith uncentered moments computed using frame data labeled by the improved B1 baseline algorithm as falling into each class m,

$$M_{i,m} = \sum_{\mathbf{\Psi}[n]=m} \left[\mathbf{f}\left(\mathbf{X}\left[n\right]\right)\right]^i \quad (10)$$

where $\mathbf{f}$ is a feature vector extracted from the audio $\mathbf{X}\left[n\right]$.

We note that the effect of the selected frame size and frame step, both of 110 ms in duration, is to discretize the references. The discretized references, relative to the "ground truth", utterance-level references, exhibit a miss rate and false alarm rate of 0.92% and 0.96%, respectively; this is small relative to the errors made by either baseline algorithm. The discretized references make it possible to compare the performance of detection systems trained on the output of the improved baseline to that of systems trained on the ground truth.

5. FACTORIZED STATE MODELING

In this section, we treat the state of each participant k independently, constructing a single speech $m = \mathcal{S}$ model and a single non-speech $m = \mathcal{N}$ model for each. The total number of states for a meeting of K participants is $2K$. We also restrict the feature space to one feature, namely the channel energy in dB.

We construct a detector based on Equation 7, in which, given the audio in channel k, the state of participant k is assumed independent of the audio in other channels. If we additionally assume an equiprobable prior for both states, $\hat{\psi}_k\left[n\right]$ can be selected using a maximum likelihood (ML) classifier,

$$\begin{aligned}
\hat{\psi}_k\left[n\right] &= \operatorname*{argmax}_{\psi_k[n]\in\{\mathcal{S},\mathcal{N}\}} P\left(\psi_k\left[n\right] \mid \mathbf{x}_k\left[n\right]\right) \\
&\doteq \operatorname*{argmax}_{\psi_k[n]\in\{\mathcal{S},\mathcal{N}\}} P\left(\mathbf{x}_k\left[n\right] \mid \psi_k\left[n\right]\right) && (11)
\end{aligned}$$

Alternately, lifting the assumption of conditional independence and instead basing a detector on Equation 6, which allows each participant state to be estimated from the audio on all channels, leads to

$$\begin{aligned}
\hat{\psi}_k\left[n\right] &= \operatorname*{argmax}_{\psi_k[n]\in\{\mathcal{S},\mathcal{N}\}} P\left(\psi_k\left[n\right] \mid \mathbf{X}\left[n\right]\right) \\
&\doteq \operatorname*{argmax}_{\psi_k[n]\in\{\mathcal{S},\mathcal{N}\}} P\left(\mathbf{X}\left[n\right] \mid \psi_k\left[n\right]\right) && (12)
\end{aligned}$$

The results of both systems are shown in Table 2.

Comparing the first three rows to the last six in Table 2 reveals that using all channels is advantageous, even when the participant states are assumed independent. Furthermore, models trained using labels produced by the B1 baseline (without smoothing) are much better than those trained using the "ground truth" discretized reference labels, as well as the smoothed B1 baseline labels. We believe this is because the reference segmentation naturally contains

Initial Labels	MS	FA	$1-\mathrm{F}$
$P(\mathbf{x}_k\|\psi_k)$			
discretized references	23.29	36.96	30.79
B1 baseline w/ smoothing	23.94	35.73	30.33
B1 baseline (w/o smoothing)	30.46	24.98	27.83
$P(\mathbf{X}\|\psi_k)$, diag Σ			
discretized references	18.82	33.10	26.65
B1 baseline w/ smoothing	17.74	33.07	26.19
B1 baseline (w/o smoothing)	31.77	11.50	22.95
$P(\mathbf{X}\|\psi_k)$, full Σ			
discretized references	13.42	25.65	20.00
B1 baseline w/ smoothing	13.63	31.63	23.68
B1 baseline (w/o smoothing)	**28.81**	**7.59**	**19.58**

Table 2. *Error rates (%) on the RT04s devset using factorized state models, with initial labels supplied by references and by the B1 algorithm, both before and after smoothing; abbreviations as in Table 1.*

many intra- and inter-word pauses, leading to much broader models. Finally, detection using full covariance matrices shows a vast improvement over diagonal matrices. In the remainder of this paper we therefore employ only the unsmoothed B1 baseline labels, full covariance matrices, and all channels for the estimation of every state.

6. JOINT STATE MODELING

We now discard the assumption of independence among participants (made in Equation 6), and estimate the *joint* speech activity state directly as in Equation 5. This requires consideration of the full 2^N state space; it is useful to think of each state m as identified by a binary codeword whose 1 bits represent speaking participants and whose 0 bits represent silent participants. A corresponding ML classifier has the form

$$\hat{\mathbf{\Psi}}[n] = \underset{\mathbf{\Psi}[n]}{\operatorname{argmax}} P(\mathbf{\Psi}[n] \mid \mathbf{X}[n]) \quad (13)$$

$$\doteq \underset{\mathbf{\Psi}[n]}{\operatorname{argmax}} P(\mathbf{X}[n] \mid \mathbf{\Psi}[n]) \quad (14)$$

and allows for the explicit modeling of overlap between arbitrary participants. However, because the baseline algorithm is poor at detecting overlap, there is very little training data for the overwhelming number of states. In this section, we propose three approaches to "increasing" the amount of data for training these state models during decoding.

The experiments presented here use a restricted feature set of only energy (ENE) and zero crossing rate (ZCR) per channel. While we have experimented with other features, for example with probability of voicing and kurtosis (shown to be useful in [3]), system performance using them alone and in combination with energy has been consistenly low. We believe this is due to the fact that such features are often bimodally distributed in speech and are therefore poorly modeled by our single Gaussian states. The gain from the zero crossing rate feature is minimal, but we retain it here.

A standard approach to dealing with miserly training data conditions is to share parameters. We explore one particular type of sharing, namely the interpolation of class covariances with the global covariance. We add, for the purposes of covariance estimation only, to each moment $M_{i,m}$ a quantity

$$\Delta M_{i,m} = \lambda_G \sum_{\Psi[n] \neq m} [\mathbf{f}(\mathbf{X}[n])]^i \quad (15)$$

While shrinking state covariances towards the global covariance mitigates the problem of near-singular or simply poor estimates, it does not improve the estimates of the class means. It is reasonable to expect some improvement in detection accuracy by considering how speaker j sounds on their channel when predicting what speaker k will sound like on theirs. We extend this idea to all states, by defining a rotation operator, $R(\cdot)$, over channels, which exchanges channel j with channel k; we use the same notation to define an identical operator over the bit positions of class codewords. We implement this *multichannel rotation* by adding to each moment the quantity

$$\Delta M_{i,m} = \lambda_R \sum_{R(\Psi[n])=m} [\mathbf{f}(R(\mathbf{X}[n]))]^i \quad (16)$$

This has the effect of sharing data mass among states with the same numbers of active speakers; single speaker states for infrequently speaking participants are just as likely to benefit as overlap states.

Finally, we propose to synthesize overlapped speech directly by "imagining" what overlap from two speakers sounds like, assuming it is known with high precision what both speakers sound like in isolation. If one participant is known to be speaking during frame p and a second participant is known to be speaking during frame q, an overlap frame between those two speakers can be synthesized by summing, at sample level, the audio for all channels from both frames; features can then be computed as if the frame were undoctored. This *multichannel synthesis* consists of adding to each moment the quantity

$$\Delta M_{i,m} = \lambda_S \sum_{\substack{\Psi[p] \cup \Psi[q] = m \\ \Psi[p] \cap \Psi[q] = 0}} [\mathbf{f}(\mathbf{X}[p] + \mathbf{X}[q])]^i \quad (17)$$

where we indicate explicitly that candidate addend frames must not share any speakers prior to addition. The above sum over p and q is partial; we identify for each frame p up to K candidates for addition.

Results for all three algorithms are shown in Table 3, separately and in combination. The parameters λ_G, λ_R and λ_S are empirically determined by minimizing the F-complement score on the development set. While rotation alone gives very modest gains, in combination with covariance sharing the improvement is considerable. It is also interesting to note that when applied alone, multichannel synthesis has the highest impact when the parameter λ_S is 1.0 — in other words, when each synthesized frame is treated like a real, undoctored frame of audio. Using all three algorithms together has a cumulative effect; combination also modifies the location of the minimum in λ-space.

Algorithm	MS	FA	$1-F$	rel impr
ENE+ZCR	31.51	8.10	21.51	—
+ sharing	28.67	7.69	19.52	9.3
+ rotation	28.79	7.04	19.36	10.0
+ synthesis	25.20	9.16	17.96	16.5

Table 3. *Error rates (%) on the RT04s devset using maximum likelihood joint state classifiers and $1-F$ score improvement, relative to unbiased joint state model estimation in the first row; abbreviations as in Table 1.*

7. STATE TRANSITION MODELING

In previous sections, we approximated $P(\mathbf{\Psi} \mid \mathbf{X})$ by ignoring the prior probability of each state. Here we explore lifting the assumption made in Equation 5, that of frame-wise temporal independence,

by applying transition probabilities within a fully ergodic hidden Markov model (eHMM) paradigm as in [4]:

$$\hat{\mathbf{\Psi}} = \underset{\mathbf{\Psi}}{\operatorname{argmax}} P(\mathbf{\Psi} \mid \mathbf{X}) \quad (18)$$

$$\doteq \underset{\mathbf{\Psi}[n]}{\operatorname{argmax}} \prod_n P(\mathbf{X}[n] \mid \mathbf{\Psi}[n]) \cdot P(\mathbf{\Psi}[n] \mid \mathbf{\Psi}[n-1]) \quad (19)$$

The transition probabilities were trained separately using 25 ISL meetings, with no overlap with the RT04s data; each probability depends only on the number of active speakers in the state transited from, the number of active speakers in the state transited to, and the number of speakers in common between the two. We use Viterbi decoding to find the single best path through each meeting and apply a smoothing pass to fill in short gaps as for the baseline. The performance improvements from applying Viterbi decoding and from smoothing are relatively large; we present them in Table 4.

Algorithm	MS	FA	$1-F$	rel impr
Viterbi	22.81	4.77	14.73	31.5
+ smoothing	**9.81**	**9.24**	**9.53**	**55.7**

Table 4. *Error rates (%) on the RT04s devset using Viterbi decoding over all 2^K states and $1-F$ score improvement, relative to unbiased joint state model estimation in the first row of Table 3; abbreviations as in Table 1.*

8. DISCUSSION

Of interest is whether the performance seen on the development set generalizes to unseen data. We note that in essense all data is unseen, because the final system is a flat-start system whose acoustic models are trained only during decoding, based on labels provided by the B1 baseline, which also requires no training. The state transition probabilities were trained on completely different data. However, the values of λ_G, λ_R and λ_S, as well as the final smoothing parameters, were determined empirically by maximizing the F-score on the development set. In Table 5 we show how each major change, described in previous sections, affects performance on the NIST RT04s evaluation set, which was unseen during system development.

Algorithm	MS	FA	$1-$F	rel impr
ENE+ZCR	41.61	8.09	28.58	—
+ sharing	37.93	7.19	25.61	10.4
+ rotation	38.30	5.95	25.48	10.9
+ synthesis	35.41	8.43	24.25	15.2
+ viterbi	32.46	2.78	20.29	29.0
+ smoothing	**21.40**	**6.94**	**14.78**	**48.3**

Table 5. *Error rates (%) on the RT04s evalset using joint state classifiers and $1-F$ score improvement, relative to unbiased joint state model estimation in the first row; abbreviations as in Table 1.*

The relative improvements on the evaluation set in Table 5 are almost identical to those in Tables 3&4 on the development set. The final system achieves a $1-$F of 9.53% and 14.78%, for the development and evaluation set, respectively, as compared to 14.29% and 23.23%, respectively, achieved using our published B0 baseline with smoothing. This represents a relative decrease in $1-$F of **33.3%** for the development set and **36.4%** for the evaluation set.

Finally, we examine to what extent this work, whose aim was to reduce the miss rate and in particular to reduce the miss rate during overlapped speech, has achieved its purpose. Table 6 shows that the absolute contribution to the overal miss rate, for durations transcribed as one participant speaking and as two participants speaking simulataneously, has been reduced by 37% and 42% respectively, as computed on the development set prior to smoothing. For all two-speaker frames the new system correctly detects both speakers 59% of the time, one of the speakers 33% of the time, and zero speakers only 1.6% of the time.

# spk	prop (%)	B1 baseline		final system	
		MS	FA	MS	FA
0	8.80	0.00	1.48	0.00	1.56
1	77.04	19.31	2.12	**12.15**	2.72
2	12.01	13.53	0.16	**7.82**	0.40
> 2	2.15	4.79	0.01	2.84	0.09
total	100.00	37.63	3.77	22.81	4.77

Table 6. *Comparison of the miss rates (MS) and false alarm rates (FA) for regions of zero, one, two, or more than two participants speaking simultaneously, between the baseline B1 system and the final system prior to smoothing, for the NIST RT04s development set. Also shown (in the second column) are the relative proportions of durations of regions.*

9. CONCLUSION

We have presented a new speech activity detection system for meetings recorded with multiple close-talking microphones, which employs Gaussian models for detection, trained at decoding time, using labels provided by a baseline algorithm which also requires no prior training. The performance exhibits a relative error reduction of 36% on unseen data over a previous system [5] [5]. In future work, we anticipate achieving significant improvements by employing mixtures of Gaussians instead of single Gaussian models. This should additionally allow for the inclusion of other features.

10. REFERENCES

[1] T. Pfau, D. Ellis, and A. Stolcke, "Multispeaker speech activity detecction for the icsi meeting recorder," in *Proc. IEEE Automatic Speecch Recognition and Understanding Workshop*, Madonna di Campiglio, Italy, 2001.

[2] A. Stolcke et al, "Further progress in meeting recognition: The icsi-sri spring 2005 speech-to-text evaluation system," in *Proc. NIST MLMI Meeting Recognition Workshop*, Edinburgh, Scotland, July 2005, pp. 39–50.

[3] S. Wrigley, G. Brown, V. Wan, and S. Renals, "Feature selection for the classification of crosstalk in multi-channel audio," in *Proc. EuroSpeech*, 2003.

[4] S. Wrigley, G. Brown, V. Wan, and S. Renals, "Speech and crosstalk detection in multi-channel audio," in *IEEE Transactions on Speech and Audio Processing*, 2002.

[5] K. Laskowski and T. Schultz, "Crosscorrelation-based multispeaker speech activity detection," in *Proc. ICSLP*, Jeju Island, Korea, October 2004.

[6] NIST, "Rich transcription ," `http://www.nist.gov/speech/tests/rt/`.