

# A SUPERVISED FACTORIAL ACOUSTIC MODEL FOR SIMULTANEOUS MULTIPARTICIPANT VOCAL ACTIVITY DETECTION IN CLOSE-TALK MICROPHONE RECORDINGS OF MEETINGS

*Kornel Laskowski and Tanja Schultz*

interACT, Carnegie Mellon University, Pittsburgh PA, USA  
Cognitive Systems Lab, Universität Karlsruhe, Germany  
kornel@cs.cmu.edu

## 1. INTRODUCTION

Vocal activity detection in close-talk microphone recordings of multiparty conversation continues to pose problems for meeting recognition systems, as evidenced by a 2-3% absolute gap in word error rates achieved with automatic and manual segmentations. State-of-the-art segmentation systems in this domain have adopted one of three different acoustic model approaches:

- independent decoding of each of  $K$  close-talk channels, in a space of two microphone states (*speech / silence*) per channel (ie. [1]);
- independent decoding of each of  $K$  close-talk channels, in a space of more than two microphone states (*speech / silence / crosstalk* [2] or *speech / silence / crosstalk / overlap* [3]), followed by rule-based channel hypothesis recombination; and
- joint simultaneous decoding of all  $K$  close-talk channels, in a space of  $2^K$  multi-microphone states (where each microphone can be in one of two states, *speech / silence*), without (the need for) recombination.

Using automatic speech recognition (ASR) word error rates (WERs) as a metric, the systems in (1) and (3) appear to have yield similar performance, in spite of significant additional architectural differences. Systems of type (2) have not been fielded for segmentation for ASR, and therefore cannot be directly compared.

Although approaches of type (3) offer a significant advantage, namely the opportunity to directly constrain the number of simultaneously vocalizing participants, they come with the caveat of a variable acoustic vector size, since conversations/meetings can have variable numbers of participants. To overcome this difficulty, unsupervised acoustic models have been deployed [4], which do not require acoustic model training data (or training time). Our previous work has shown that this severely limits the number of features, as well as the minimum frame size. The aim of the current work is to develop a supervised acoustic model, capable of producing accurate density estimates for large feature vectors extracted from short frames, for scenario (3).

## 2. BASELINE VAD SYSTEM

Our baseline VAD system was most recently described in [5]. Rather than detecting the 2-state speech ( $\mathcal{V}$ ) vs. non-speech ( $\mathcal{N}$ ) activity of each participant independently, the baseline implements a Viterbi search for the best path through a  $2^K$ -state vocal interaction space, where  $K$  is the number of participants. Our state vector,  $\mathbf{q}_t$ , formed by concatenating the concurrent binary vocal activity states  $\mathbf{q}_t[k]$ ,  $1 \leq k \leq K$ , of all participants, is allowed to evolve freely over the vocal interaction space hypercube, under stochastic transition constraints imposed by a fully-connected, ergodic hidden Markov model (eHMM). Once the best vocal interaction state path  $\mathbf{q}^*$  is found, we index out the corresponding best vocal activity state path  $\mathbf{q}^*[k]$  for each participant  $k$ . The underlying motivation for this approach is that it allows us to model the constraints that participants exert on one another; it is generally accepted that participants are more likely to begin vocalizing in silence than when someone else is already vocalizing [6].

The architecture of the baseline segmentation system is depicted in Figure 1. The system relies on a pre-trained vocal interaction model for transition probabilities. The acoustic model, on the other hand, is trained during decoding, using the unlabeled test audio and a set of labels obtained in an unsupervised way. Decoding is followed by a smoothing pass in which short talkspurts are eliminated and short inter-talkspurt gaps are bridged.

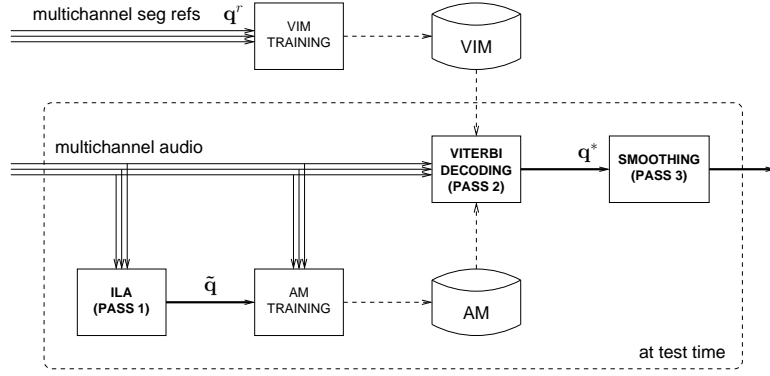


Fig. 1. Architecture of the baseline VAD system.

## 2.1. Transition Model

In our system, transition model probabilities are supplied by a model of vocal interaction [7] whose form is

$$\begin{aligned}
 P(\mathbf{q}_{t+1} = \mathbf{S}_j | \mathbf{q}_t = \mathbf{S}_i) = & \\
 & P(\|\mathbf{q}_{t+1}\| = n_j, \|\mathbf{q}_{t+1} \cdot \mathbf{q}_t\| = o_{ij} | \|\mathbf{q}_t\| = n_i) \times \\
 & P(\mathbf{q}_{t+1} = \mathbf{S}_j | \|\mathbf{q}_{t+1}\| = n_j, \|\mathbf{q}_{t+1} \cdot \mathbf{q}_t\| = o_{ij}, \|\mathbf{q}_t\| = n_i) ,
 \end{aligned} \tag{1}$$

The first factor is the Extended Degree of Overlap model [7], in which  $\|\mathbf{q}_t\|$  represents the number of participants vocalizing at time  $t$ , and  $\|\mathbf{q}_t \cdot \mathbf{q}_{t+1}\|$  represents the number of participants who were vocalizing at time  $t$  and who continue to vocalize at time  $t + 1$ .  $n_i \equiv \|\mathbf{S}_i\|$  and  $n_j \equiv \|\mathbf{S}_j\|$  are the number of vocally active participants in states  $\mathbf{S}_i$  and  $\mathbf{S}_j$ , respectively, and  $o_{ij} \equiv \|\mathbf{S}_i \cdot \mathbf{S}_j\| \leq \min(n_i, n_j)$  is the number of same participants which are vocally active in both  $\mathbf{S}_i$  and  $\mathbf{S}_j$ . The second factor is a uniform probability over possible specific-participant activity. Details regarding the motivation, structure, and training of the model are given in [7].

## 2.2. Unsupervised Baseline Acoustic Model

The baseline VAD system employs an unsupervised acoustic model, providing emission probabilities for the multi-channel observables  $\mathbf{X}_t$  according to

$$P_{uns}(\mathbf{X}_t | \mathbf{S}_i) = P(\mathbf{X}_t | \mathbf{N}(\mu_i, \Sigma_i)) \tag{2}$$

where  $\mathbf{N}(\mu, \Sigma)$  is a multivariate Gaussian distribution with a full covariance matrix. The number of dimensions  $D$  of each multivariate Gaussian is equal to  $K$ , the number of participants in the conversation in question.

The models  $\mathbf{N}(\mu, \Sigma)$  are estimated using the complete *unlabeled* (test) data  $\mathbf{X}_t$ , for  $1 \leq t \leq T$ , following an unsupervised initial label assignment (ILA) via

$$\tilde{\mathbf{q}}[k] = \begin{cases} \mathcal{V}, & \text{if } \sum_{j \neq k} \log \left( \frac{\max_{\tau} \phi_{jk}(\tau)}{\phi_{jj}(0)} \right) > 0 \\ \mathcal{N}, & \text{otherwise} \end{cases} , \tag{3}$$

where  $\phi_{jk}(\tau)$  is the crosscorrelation between IHM channels  $j$  and  $k$  at lag  $\tau$ , and  $\tilde{\mathbf{q}}[k]$  is the initial label assigned to the frame in question. We have shown, in [8], that under certain assumptions the criterion in Equation 3 is equivalent to declaring a participant as vocalizing when the distance between the location of the dominant sound source and that participant's microphone is smaller than the geometric mean of the distances from the source to each of the remaining microphones.

The initial label assignment described in Equation 3 produces a partitioning of the multichannel test audio. The labeled frames are used to train a single, full-covariance Gaussian for each of the  $2^K$  states in our search space, over a feature space of  $K$  features: a log-energy for each channel. Features are computed using 100 ms non-overlapping windows, following signal preemphasis  $(1 - z^{-1})$ .

For certain participants, and especially for frames in which more than one participant vocalizes, the ILA may identify too few frames in the test meeting for standard acoustic model training. To address this problem, we have proposed and evaluated two methods: feature space rotation, and sample-level overlap synthesis [4]. These methods are controlled by three parameters,  $\{\lambda_G, \lambda_R, \lambda_S\}$ , whose magnitudes empirically appear to depend on the number of features per channel and on the overall test meeting duration.

### 3. SUPERVISED FACTORIAL ACOUSTIC MODEL

In the current work, we propose to replace the unsupervised acoustic model with a supervised acoustic model, overcoming the problem of variable size feature vectors due to a variable number of conversation participants by factoring:

$$P_{sup}(\mathbf{X}_t | \mathbf{S}_i) = \prod_{k=1}^K P(\mathbf{X}_t[k] | \zeta(\mathbf{S}_i, k)) \quad (4)$$

with

$$P(\mathbf{X}[k] | \zeta(\mathbf{S}_i, k)) = \sum_{m=1}^M p_{\zeta(i,k),m} P(\mathbf{X}[k] | \mathbf{N}(\mu_{\zeta(i,k),m}, \sigma_{\zeta(i,k),m}^2)) \quad (5)$$

where  $\mathbf{N}(\mu, \sigma^2)$  is a multivariate Gaussian distribution with a diagonal variance matrix. The number of dimensions  $D$  of each multivariate Gaussian is equal to  $F$ , the number of single-participant features computed.

We have chosen to experiment with two feature sets. The first is a standard ASR vector, consisting of log-energy, 13 Mel-frequency cepstral coefficients (MFCCs; excluding  $\mathbf{C}_0$ ), their first-order differences (delta), and the first-order differences of their first-order differences (delta-delta). Cepstral mean subtraction (CMS) is also used in the present work; means are accumulated in a preliminary pass over the complete recording. The resulting feature vector has  $F \equiv 39$  features; its computation is shown schematically in Figure 2. The second feature set is this same vector, augmented with minimum and maximum NLED features, as proposed by [1]. These features were designed for differentiating between nearfield and farfield speech, when all participants to a conversation are instrumented with close-talk microphones. This second feature vector has  $F \equiv 41$  features; the computation of the NLED features is explained in [1].

The models  $\mathbf{N}(\mu, \sigma^2)$  are estimated using the multichannel audio  $\mathbf{X}_{r,t}$ ,  $1 \leq t \leq T_r$  and  $1 \leq r \leq R$ , of a training set of  $R$  multiparty conversations, each of duration  $T_r$ .

#### 3.1. 2 Substates Per Microphone

We allow each microphone to be in one of two states,  $\mathcal{V}^{NF}$  or  $\mathcal{N}^{NF}$ , corresponding to the presence of nearfield vocalization or its absence, respectively. In this case,

$$\zeta(i, k) \equiv \begin{cases} \mathcal{V}^{NF}, & \text{if } \mathbf{S}_i[k] \neq \mathcal{N} \\ \mathcal{N}^{NF}, & \text{if } \mathbf{S}_i[k] = \mathcal{N} \end{cases} \quad (6)$$

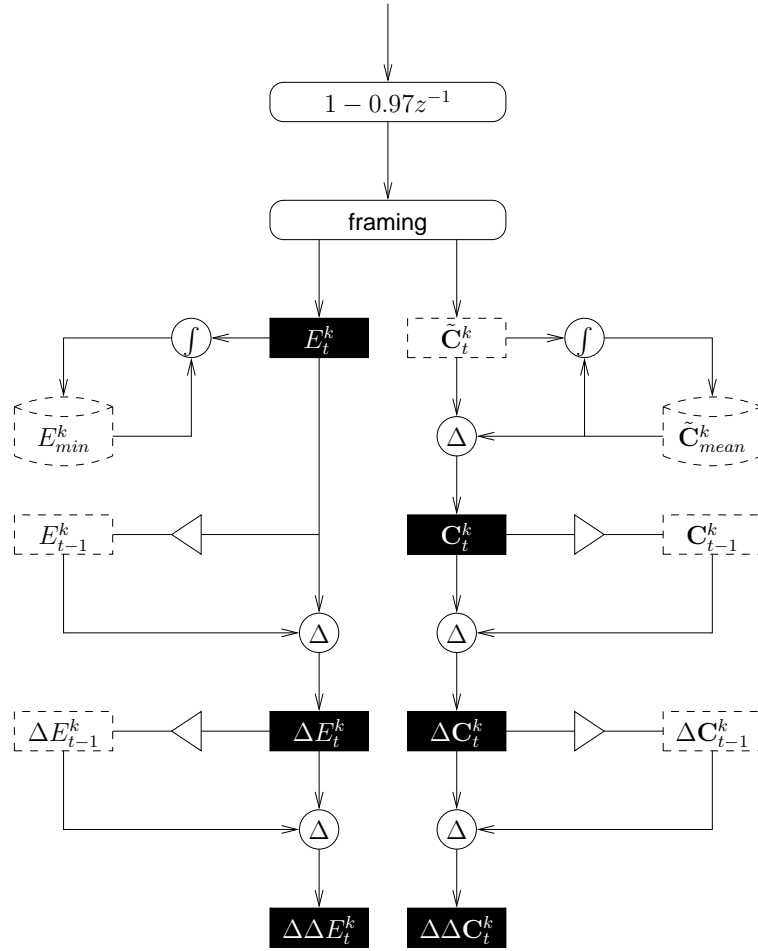
Since there are  $K$  microphones, there are  $2^K$  multi-microphone states, corresponding exactly to the number of multi-participant vocal activity states  $\mathbf{q}$ .

#### 3.2. 3 Substates Per Microphone

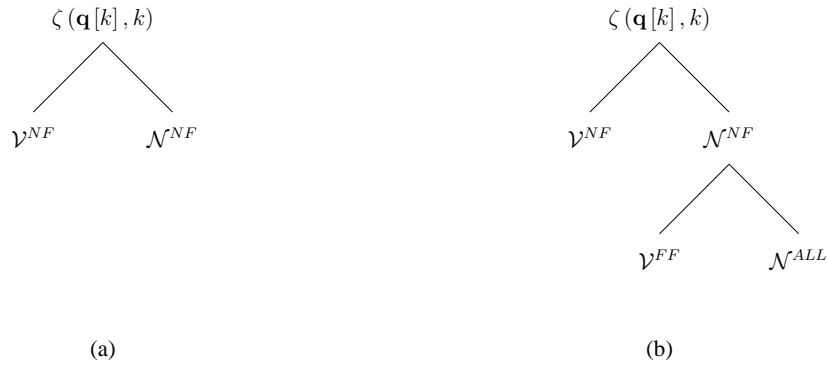
Although modeling each microphone as being in one of two states is the most natural approach to segmentation, several efforts in single-participant segmentation have been made to extend this model to farfield activity states (ie. [3]). In [2], three states were considered:  $\mathcal{V}^{NF}$ ,  $\mathcal{V}^{FF}$ , or  $\mathcal{N}^{ALL}$ , corresponding to the presence of nearfield vocalization, the absence of nearfield vocalization but the presence of farfield vocalization, and the absence of both nearfield and farfield vocalization, respectively. We explore this substate space in the current work, whereby

$$\zeta(i, k) \equiv \begin{cases} \mathcal{V}^{NF}, & \text{if } \mathbf{S}_i[k] \neq \mathcal{N} \\ \mathcal{V}^{FF}, & \text{if } \mathbf{S}_i[k] = \mathcal{N}, \exists j \text{ s.t. } \mathbf{S}_i[j] \neq \mathcal{N} \\ \mathcal{N}^{ALL}, & \text{if } \mathbf{S}_i[j] = \mathcal{N} \forall j \end{cases} \quad (7)$$

Here, there are  $3^K$  multi-microphone substates; however, only  $2^K$  of them correspond to valid conversation states. For example, the conversation state in which some microphones are in the  $\mathcal{V}^{FF}$  substate and some in the  $\mathcal{N}^{ALL}$  substate is not possible; either all microphones are in  $\mathcal{N}^{ALL}$  or none are.



**Fig. 2.** Standard features as used in ASR, including log-energy ( $E_t^k$ ), MFCC coefficients ( $C_t^k$ ), delta ( $\Delta E_t^k$  and  $\Delta C_t^k$ ), and delta-delta ( $\Delta\Delta E_t^k$  and  $\Delta\Delta C_t^k$ ) features. Cepstral mean ( $\tilde{C}_{mean}^k$ ) subtraction (CMS), and the accumulation of a log-energy floor ( $E_{min}^k$ ) are also shown. Framing involves a step size of 100ms, with a 100ms Hamming window.



**Fig. 3.** Alternative models of the substate  $\zeta(i, k)$  of the close-talk microphone belonging to participant  $k$  in conversation state  $S_i$ : (a) a two-substate model; and (b) a three-substate model.

### 3.3. Overall System Modifications

The replacement of the unsupervised acoustic model by the supervised acoustic model entails several modifications to the overall VAD system, as shown in Figure 4. In particular, since the acoustic model is trained prior to testing, audio processing during

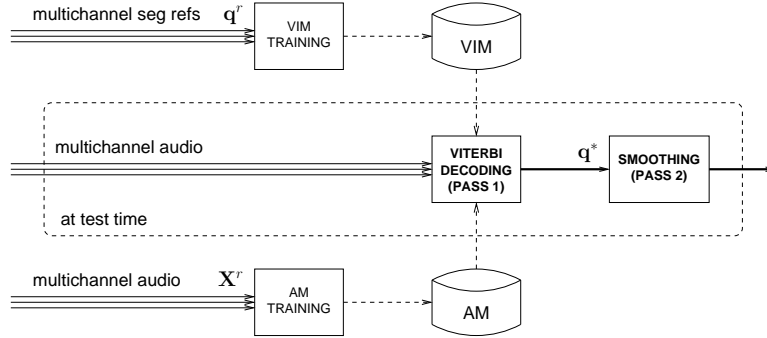


Fig. 4. Architecture of the proposed VAD system with supervised acoustic models.

testing is much simplified.

#### 4. DATA

The supervised acoustic models in this work were trained using the ICSI Meeting Corpus [9], making use of the forced alignment segmentation available in the associated MRDA Corpus [10]. The corpus consists of 75 meetings, amounting to over 66 hours of meeting time. The number of participants with close-talk microphones varies between 3 and 9.

The EDO transition probability model was trained using the ISL Meeting Corpus [11]. This corpus consists of 18 meetings, and the number of participants with close-talk microphones varies between 3 and 9. The total recorded meeting time is 9.65 hours.

The development and evaluation data used in our experiments consist of two datasets from the NIST RT-06s Meeting Recognition evaluation. `rt05s_eval` was used for development, and `rt06s_eval` was used for final evaluation; we have retained this separation in the current work. The two sets consist of ten 11-minute excerpts from several meetings recorded at different sites; the number of participants per meeting varies between 3 and 11. Segmentation system development was carried out while excluding a single meeting from `rt05s_eval` which contained a participant on speakerphone; the resulting development dataset is referred to as `rt05s_eval*` (this dataset was also referred to as *confDEV* in [12]).

#### 5. EXPERIMENTS

In the context of this work, an experiment consists of decoding, using a particular VAD system, all of the audio in the development set, `rt05s_eval*`, to produce a segmentation for each participant in each meeting excerpt. The hypothesized segmentation is then scored against a reference segmentation, which can be a manually produced utterance-level segmentation  $\Upsilon_U$ . However, in previous unpublished work using a development ASR system, we have observed that segmentations exist which lead to word error rates which are up to 1.5% absolute lower than those obtained with manual segmentation. After considerable experimentation, we have found that beginning with the forced alignment of partial and complete words (but not non-lexical human noises), bridging gaps shorter than 0.375s, and padding each resulting talkspurt with 0.015s and 0.025s at the beginning and the end, respectively, leads to a near-optimal segmentation from an ASR point of view. We refer to this reference segmentation as  $\Upsilon_T$ .

Scoring a hypothesized segmentation consists of identifying false alarms and misses, ie. intervals during which a participant is hypothesized as speaking when in fact they are not, and missing speech which they do produce, respectively. For diagnostic purposes during development, we further break down false alarms into false alarms occurring during intervals when no participant is vocalizing (FA0), intervals during which one participant is vocalizing (FA1), intervals during which two participants are simultaneously vocalizing (FA2), and intervals during which three or more participants are simultaneously vocalizing (FA3). Similarly, the miss rate is broken down into MS1, MS2, and MS3, representing missed speech from participant  $k$  when only participant  $k$  was vocalizing, when participant  $k$  was vocalizing simultaneously with some second participant, and when participant  $k$  was vocalizing simultaneously with two other participants, respectively.

We show such a scoring of the baseline in Table 1. Smoothing of VAD output, in this table, consists of bridging all gaps of 0.4s or shorter, and then pruning out all talkspurts of 0.2s or shorter. This particular smoothing policy was developed for this particular VAD system, and normally would need to be retuned when the VAD system changes.

Segm vs Reference Segm	Segmentation Errors						
	FA0	FA1	FA2	FA3	MS1	MS2	MS3
Baseline vs $\Upsilon_U$	2.55	2.67	0.25	0.03	6.66	8.07	4.13
Smoothed Baseline vs $\Upsilon_U$	2.80	2.62	0.24	0.04	5.96	7.98	4.13
Baseline vs $\Upsilon_T$	9.25	4.43	0.27	0.02	5.14	3.52	0.99
Smoothed Baseline vs $\Upsilon_T$	9.51	4.39	0.27	0.03	4.32	3.50	0.98
$\Upsilon_T$ vs $\Upsilon_U$	0.62	0.15	0.03	0.00	7.86	7.12	3.93

**Table 1.** Segmentation errors committed by the baseline, before and after smoothing, when scored against the manual utterance segmentation  $\Upsilon_U$  and the forced-alignment-mediated talkspurt segmentation  $\Upsilon_T$ . Also shown is the performance of  $\Upsilon_T$  as scored against  $\Upsilon_U$ .

As Table 1 shows, the better segmentation  $\Upsilon_T$  misses approximately 20% of what is considered to be speech by  $\Upsilon_U$ . This is reflected in the types of errors committed by the baseline. When scored against  $\Upsilon_U$ , the total false alarm rate is approximately 5.5%; this number increases to approximately 14% when scoring is performed against  $\Upsilon_T$ . This indicates that something which is considered by human annotators to be part of utterances, but which is not lexical (or is incorrectly forced-aligned), is being detected by the baseline VAD system as vocalization. These intervals are therefore penalized by  $\Upsilon_T$ . We note also that smoothing reduces the miss rate but increases the false alarm rate only negligibly, and, in addition, it is only FA0 which is increased. We believe these false alarms are less deleterious than other false alarms, because due to the absence of farfield speech they are likely to not lead to word insertions during ASR decoding.

As mentioned above, the smoothing operation, together with potential padding operations following it, needs to be recalibrated when the architecture of the VAD system changes. Since in the current work we intend to change the architecture quite drastically, the current smoothing pass consisting of 0.4s bridging and 0.2s pruning may not be appropriate for performance comparison. We will therefore compare systems using received operating characteristic curves, which show the performance of a system over a wide range of possible smoothing/padding parameters. The ROC curves we show in subsequent sections are produced in the following way:

1. segment audio using a particular VAD system
2. exhaustively sample a 4-dimensional smoothing/padding parameter space, over the range:
  - bridging: {0.05, 0.15, 0.25, 0.35, 0.45, 0.55}
  - pruning: {0.05, 0.15, 0.25, 0.35, 0.45, 0.55}
  - pre-padding: {0.015, 0.03, 0.045, 0.06, 0.075, 0.09, 0.105, 0.120, 0.135, 0.150}
  - post-padding: {0.015, 0.03, 0.045, 0.06, 0.075, 0.09, 0.105, 0.120, 0.135, 0.150, 0.165, 0.180, 0.195, 0.210}
3. for each 4-tuple above, score against  $\Upsilon_T$  and compute
  - the false positive rate,  $FPR = FA0 + FA1 + FA2 + FA3$
  - the true positive rate,  $TPR = 1 - MS1 - MS2 - MS3$
4. for each  $(FPR, TPR)$  pair, eliminate the pair if another pair exists whose  $FPR$  is lower and whose  $TPR$  is higher

The remaining  $(FPR, TPR)$  pairs constitute a convex ROC hull.

In all experiments which follow, we produce two ROC curves. The first is as explained above, while the second is identical except that  $FPR = FA1 + FA2 + FA3$ , ie. we do not penalize for false alarms during intervals when all participants are silent. As argued above, these false alarms are less likely to lead to word insertions than other false alarms are, as the latter may be mistaking farfield vocalization (crosstalk) as nearfield vocalization.

## 5.1. Unsupervised versus supervised AMs

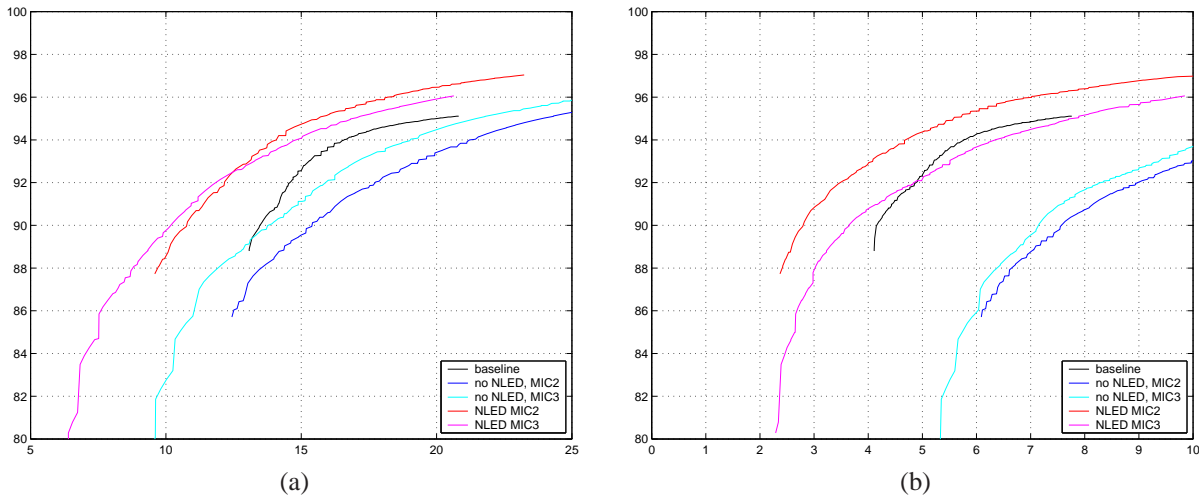
In a first experiment, we compare the performance of the baseline unsupervised AM with that of a supervised factorial AM consisting of the product of  $K$  single-participant AMs. As can be seen in Figure 5 (a), the unsupervised AM outperforms the supervised AM when only standard MFCC front-end features are used. We note that, although standard energy and MFCC

features (and their delta and delta-delta features) do not explicitly attempt to account for crosstalk, the VAD system architecture discards a significant amount of crosstalk since it allows a maximum of only two simultaneously vocalizing participants; the performance of standard MFCC front-end features would be significantly worse without this constraint.

Figure 5 (a) also shows that when standard MFCC front-end features are augmented with NLED features, the supervised AM is better than the unsupervised AM.

In Figure 5 (b), we show a similar ROC curve for the same systems, in which performance is not penalized by false alarms during intervals when *all* participants are silent. This is an alternative characterization of performance, motivated by the assumption that in the absence of farfield speech, false alarms in segmentation will not lead to word insertions in ASR. As the figure indicated, supervised AMs without NLED features have particularly low FA0 errors, which makes the former appear more competitive overall (Figure 5 (a)); when FA0 errors are not penalized, supervised AM systems without NLED features perform significantly more poorly than either unsupervised AM systems or supervised AM systems employing on NLED features.

Figure 5 also shows that in general, supervised AM systems using a 3-substate microphone space yield inferior performance to systems using a 2-substate microphone space. This aspect is currently under investigation.



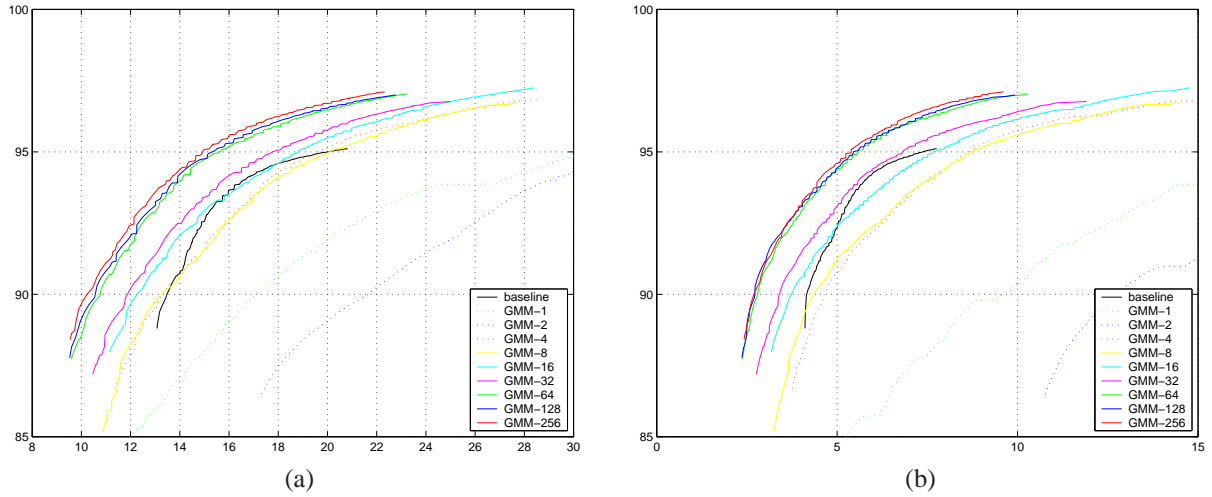
**Fig. 5.** ROC discrimination of the unsupervised AM baseline system, and of supervised AM systems with 2-substate or 3-substate microphone spaces, and either with or without NLED features, with 64 Gaussians per mixture. The maximum allowed number of simultaneously vocalizing participants is 2 for all curves.  $TPR = 1 - (MS1 + MS2 + MS3)$ ;  $FPR = FA0 + FA1 + FA2 + FA3$  in (a), and  $FPR = FA1 + FA2 + FA3$  in (b).

### 5.2. Supervised AM performance as a function of model complexity

We are also interested in how the sensitivity of supervised AM systems varies as the number of Gaussian components per mixture increases. These results are shown in Figures 6 (a) & (b). It can be seen that performance improves monotonically over the entire ROC range, for GMMs consisting of at least 8 Gaussians. There is a significant performance improvement between 32 and 64 components, after which smaller improvements continue to be visible (but at the cost of significantly longer training times). This trend is apparent for both measures of the false positive rate (FPR), when FA0 is either included (Figure 6 (a)) or excluded (Figure 6 (b)); in the latter case, performance differences between systems employing GMMs of 64 or more Gaussians are less conclusive.

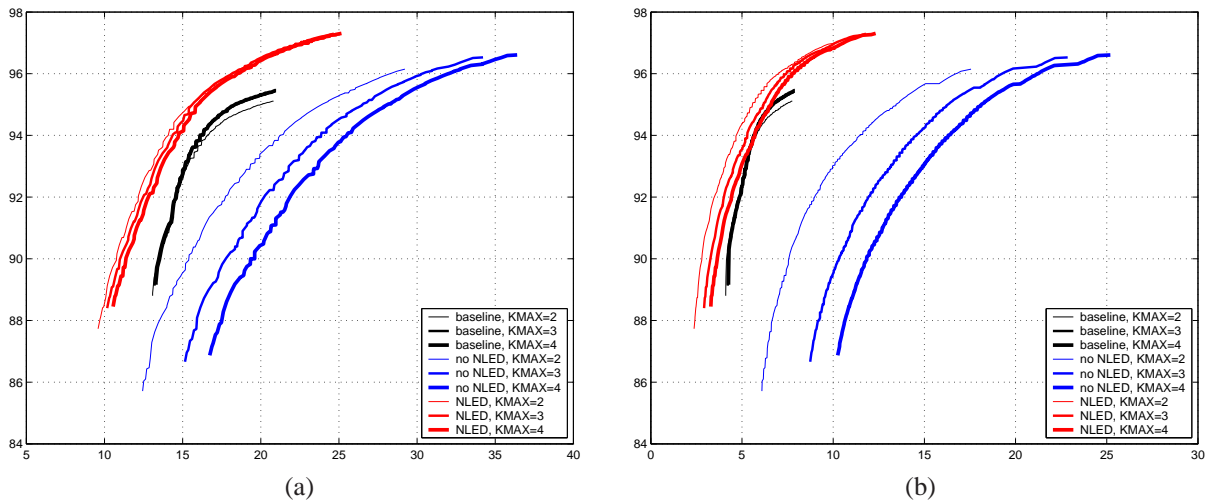
### 5.3. Supervised AM performance as a function of the maximum number of simultaneously vocalizing participants

Next, we explore the performance of the unsupervised AM system, together with that of the supervised AM systems with and without NLED features, as a function of the maximum allowed number of simultaneously vocalizing participants. As mentioned above, limiting this number,  $K_{max}$ , prevents systems for detecting all speech during intervals of high overlap (which are rare), but eliminates the opportunity for mistaking crosstalk for nearfield speech.



**Fig. 6.** ROC discrimination of the unsupervised AM baseline system, and of supervised AM systems with a 2-substate microphone space and with NLED features, as a function of the number of Gaussians per mixture. The maximum allowed number of simultaneously vocalizing participants is 2 for all curves.  $TPR = 1 - (MS1 + MS2 + MS3)$ ;  $FPR = FA0 + FA1 + FA2 + FA3$  in (a), and  $FPR = FA1 + FA2 + FA3$  in (b).

Figure 7 gives a comparison. It can be seen that for supervised AM systems without NLED features, increasing  $K_{max}$  effectively shifts the ROC curves to the right (more false alarms) by as much as 5% absolute, without significantly affecting the miss rate. When NLED features are employed, a similar trend is observed; however, the increase in the false alarm rate is only approximately 1% absolute. In contrast, the unsupervised AM system appears not to suffer from an increase in false alarms, while at the same time showing a slight reduction of misses. It can therefore be concluded that, although the overall performance of the unsupervised AM system is lower than that of the supervised AM system with NLED features, the former is more robust to an increase in  $K_{max}$ . This may make it more competitive for conversations, or intervals thereof, with high vocalization overlap.



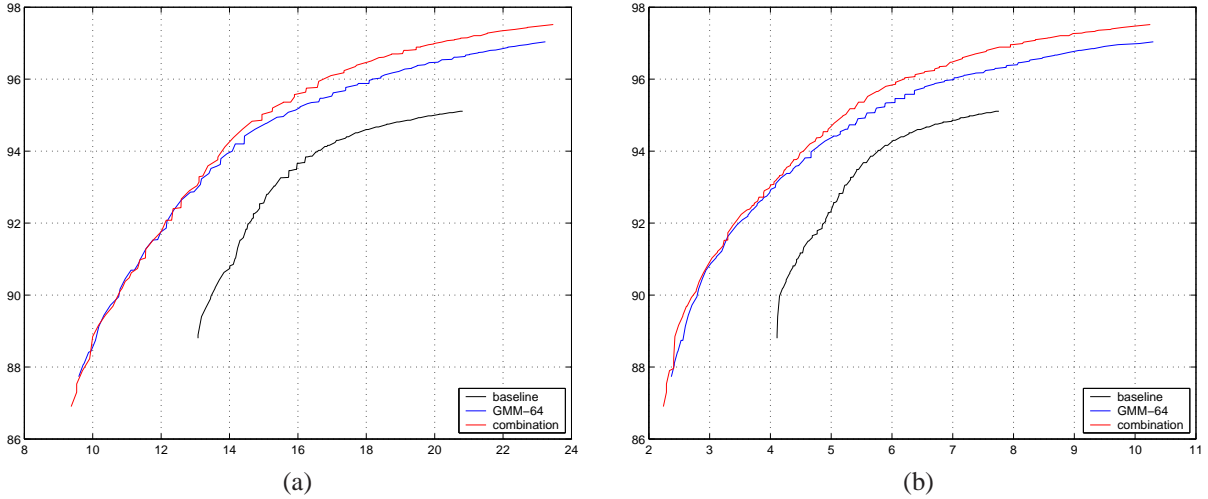
**Fig. 7.** ROC discrimination of the unsupervised AM baseline system, and of supervised AM systems with a 2-substate microphone space and either with or without NLED features, with 64 Gaussians per mixture, as a function of the maximum allowed number of simultaneously vocalizing participants.  $TPR = 1 - (MS1 + MS2 + MS3)$ ;  $FPR = FA0 + FA1 + FA2 + FA3$  in (a), and  $FPR = FA1 + FA2 + FA3$  in (b).

#### 5.4. Supervised and unsupervised AM combination

As a result of the above findings, we attempt to combine the unsupervised and supervised AMs in one system, using

$$P(\mathbf{X}_t | \mathbf{S}_i) = P_{uns}(\mathbf{X}_t | \mathbf{S}_i) \cdot P_{sup}(\mathbf{X}_t | \mathbf{S}_i) \quad (8)$$

The results of decoding using this combined AM are shown in Figure 8 (a) & (b). It can be seen that the resulting combination generally outperforms both systems; in the low FPR region, the combined system is not significantly different from the supervised system alone, while in the high TPR region the combined system yields higher recall, by approximately 0.5% absolute, than the supervised AM system alone.



**Fig. 8.** ROC discrimination of the unsupervised AM baseline system, of the supervised AM system with a 2-substate microphone space with NLED features, with 64 Gaussians per mixture and a maximum allowed number of 2 simultaneously vocalizing participants, and of an AM consisting of the product of the unsupervised and supervised AMs.  $TPR = 1 - (MS1 + MS2 + MS3)$ ;  $FPR = FA0 + FA1 + FA2 + FA3$  in (a), and  $FPR = FA1 + FA2 + FA3$  in (b).

#### 5.5. Supervised AM performance as a function of the frame step

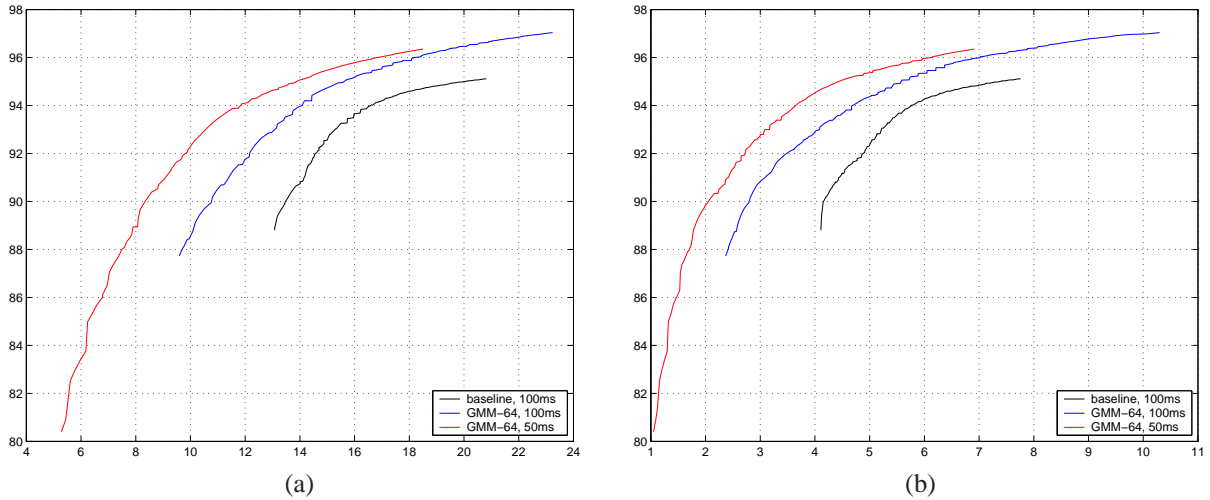
Finally, for completion, we run the supervised AM system at a smaller frame step. While all other experiments in the current work use a frame size of 100ms and a frame step of 100ms, here we train a transition model for a frame size of 100ms but a frame step of 50ms (ie. 50% overlapping frames). In [8] we noted that reductions to the frame step and frame size for the unsupervised AM system lead to overtraining and poorer NT-Norm estimation, respectively, making a direct comparison with an unsupervised AM system using a 100ms frame size and a 50ms frame rate impossible without further development.

We show the performance of this supervised 100ms/50ms AM system in Figure 9 (a) & (b). It can be seen that, using both FPR measures (with and without FA0), the halved frame step and resulting 100ms frame overlap reduce false alarms by as much as 3% absolute, at equal TPRs. Optimization of the framing parameters for supervised acoustic modeling within our architecture is currently under investigation.

#### 5.6. Generalization of supervised AM performance to unseen data

Finally, we analyze the performance of select supervised AM systems, alongside that of the unsupervised systems, for unseen evaluation data. The results are shown in Figure 10. The top panels in the figure show the development set performance of: (1) the baseline unsupervised AM system; (2) the supervised AM system with NLED features, 64 Gaussians per mixture, a maximum allowed number of 2 simultaneously vocalizing participants, and a frame step of 100ms; (3) a combination of (1) and (2) as in Subsection 5.4; and (4) the same system as in (2) with a frame step of 50ms, as in Subsection 5.5.

The observed trends for the development data in panels (a) and (b) are also observed in panels (c) and (d) for the evaluation data. The supervised AM system (2) outperforms the baseline (1) over the entire ROC range explored. A combination (3) of



**Fig. 9.** ROC discrimination of the unsupervised AM baseline system, of the supervised AM system with a 2-substate microphone space with NLED features, with 64 Gaussians per mixture and a maximum allowed number of 2 simultaneously vocalizing participants, and of an identical supervised AM system with a frame step of 50ms instead of 100ms. The frame size for all systems shown is 100 ms.  $TPR = 1 - (MS1 + MS2 + MS3)$ ;  $FPR = FA0 + FA1 + FA2 + FA3$  in (a), and  $FPR = FA1 + FA2 + FA3$  in (b).

supervised and unsupervised AMs leads to performance which is similar to that of the supervised AM system in the low FPR region but to better performance in the high TPR region. Finally, halving the frame step (4) outperforms all other systems almost everywhere in the ROC range by a significant amount.

## 6. CONCLUSIONS

We have implemented a supervised acoustic model for VAD in conversations with an arbitrary number of participants, and analyzed its performance with respect to the unsupervised AM baseline. Analysis consisted of a broad exploration of several parameters, two of which (inclusion of NLED features and decoding constraints on the maximum allowed number of simultaneously vocalizing participants) are explicitly intended to limit the deleterious effect of crosstalk. Additional parameters whose effect was analyzed included the number of Gaussians per mixture and the effect of the frame step.

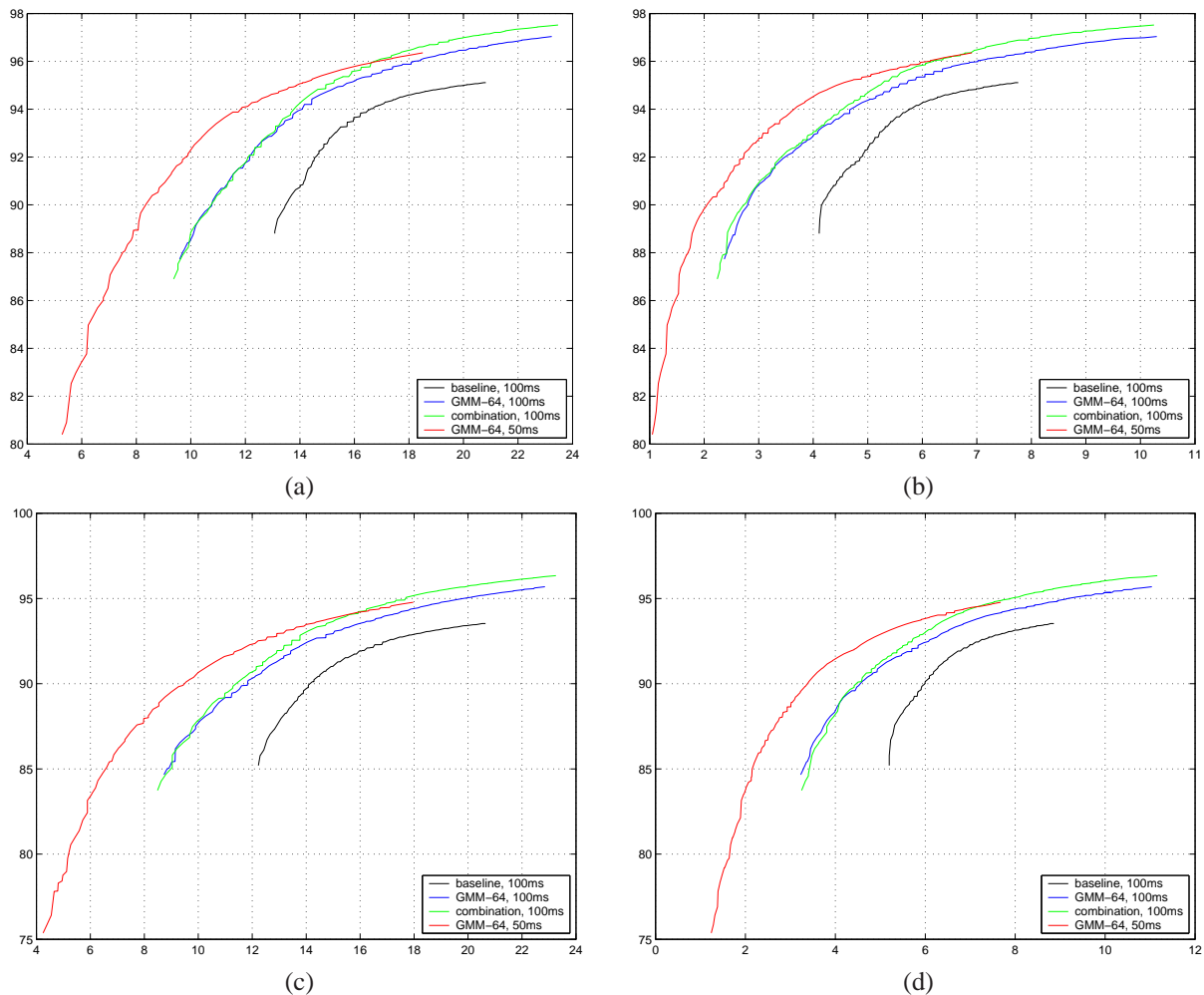
Our findings show that the unsupervised AM baseline outperforms a supervised AM system which uses standard MFCC front-end features, but that this effect is reversed when NLED features are included. An increase in the maximum allowed number of simultaneously vocalizing participants leads to higher FPRs for supervised systems, with negligible impact on TPRs; for unsupervised systems, the same increase leads to higher TPRs, with negligible impact on FPRs. Combining supervised and unsupervised systems leads to performance with higher TPRs, and similar FPRs, to those achieved with the supervised system alone. Finally, reducing the frame step from 100ms to 50ms, and thereby achieving a 50% overlap when a 100ms frame size is used, leads to significantly lower FPRs for supervised systems.

## 7. ACKNOWLEDGMENTS

We would like to thank Mari Ostendorf for useful discussion during our initial MFCC modeling efforts.

## 8. REFERENCES

- [1] K. Boakye and A. Stolcke, “Improved speech activity detection using cross-channel features for recognition of multiparty meetings,” in *Proceedings of INTERSPEECH*, Pittsburgh PA, USA, September 2006, pp. 1962–1965.
- [2] Z. Huang and M. Harper, “Speech activity detection on multichannels of meetings recordings,” in *Proceedings of MLMI*, Edinburgh, UK, 2005, vol. 3869 of *Lecture Notes in Computer Science*, pp. 415–427, Springer Berlin/Heidelberg.



**Fig. 10.** ROC discrimination of the unsupervised AM baseline system, and 3 select systems from Figures 5, 8, and 9, for the development data (a) & (b) and the evaluation data (c) & (d).  $TPR = 1 - (MS1 + MS2 + MS3)$ ;  $FPR = FA0 + FA1 + FA2 + FA3$  in (a) & (c), and  $FPR = FA1 + FA2 + FA3$  in (b) & (d).

- [3] S. Wrigley, G. Brown, V. Wan, , and S. Renals, “Speech and crosstalk detection in multichannel audio,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 84–91, 2005.
- [4] K. Laskowski and T. Schultz, “Unsupervised learning of overlapped speech model parameters for multichannel speech activity detection in meetings,” in *Proceedings of ICASSP*, Toulouse, France, May 2006, pp. 993–996.
- [5] K. Laskowski and T. Schultz, “Simultaneous multispeaker segmentation for automatic meeting recognition,” in *Proceedings of EUSIPCO*, Poznań, Poland, September 2007, pp. 1294–1298.
- [6] Ö. Çetin and E. Shriberg, “Overlap in meetings: ASR effects and analysis by dialog factors, speakers, and collection site,” in *Proceedings of MLMI*, Washington DC, USA, May 2006, vol. 4299 of *Lecture Notes in Computer Science*, pp. 200–211, Springer Berlin/Heidelberg.
- [7] K. Laskowski and T. Schultz, “Modeling vocal interaction for segmentation in meeting recognition,” in *Proceedings of MLMI*, Brno, Czech Republic, 2007, vol. 4892 of *Lecture Notes in Computer Science*, pp. 259–270, Springer Berlin/Heidelberg.
- [8] K. Laskowski and T. Schultz, “A geometric interpretation of non-target normalized maximum cross-channel correlation

for vocal activity detection in meetings,” in *Proceedings of HLT-NAACL; Companion Volume, Short Papers*, Rochester NY, USA, April 2007, pp. 89–92.

- [9] A. Janin et al., “The ICSI Meeting Corpus,” in *Proc. ICASSP*, Hong Kong, China, 2003, pp. 364–367.
- [10] E. Shriberg et al., “The ICSI Meeting Recorder Dialog Act (MRDA) Corpus,” in *Proc. SIGdial*, Cambridge MA, USA, 2004, pp. 97–100.
- [11] S. Burger, V. MacLaren, and H. Yu, “The ISL Meeting Corpus,” in *Proceedings of ICSLP*, Denver CO, USA, September 2002, pp. 301–304.
- [12] C. Fügen, S. Ikbal, F. Kraft, K. Kumatani, K. Laskowski, J. McDonough, M. Ostendorf, S. Stüker, and M. Wölfel, “The ISL RT-06S Speech-to-Text System,” in *Proceedings of MLMI*, Washington DC, USA, May 2006, vol. 4299 of *Lecture Notes in Computer Science*, pp. 407–418, Springer Berlin/Heidelberg.