CONTRASTING EMOTION-BEARING LAUGHTER TYPES IN MULTIPARTICIPANT VOCAL ACTIVITY DETECTION FOR MEETINGS

Kornel Laskowski

Language Technologies Institute, Carnegie Mellon University, Pittsburgh PA, USA
Cognitive Systems Lab, Universität Karlsruhe, Karlsruhe, Germany

ABSTRACT

The detection of laughter in conversational interaction presents an important challenge in meeting understanding, important primarily because laughter is predictive of the emotional state of participants. We present evidence which suggests that ignoring unvoiced laughter improves the prediction of emotional involvement in collocated speech, making a case for the distinction between voiced and unvoiced laughter during laughter detection. Our experiments show that the exclusion of unvoiced laughter during laughter model training as well as its explicit modeling lead to detection scores for voiced laughter which are much higher than those otherwise obtained for all laughter. Furthermore, duration modeling is shown to be a more effective means of improving precision than interaction modeling through joint-participant decoding. Taken together, the final detection F-scores we present for voiced laughter on our development set comprise a 20% reduction of error, relative to F-scores for all laughter reported in previous work, and 6% and 22% relative reductions in error on two larger datasets unseen during development.

Index Terms— Laughter detection, Speech detection, Vocal interaction, Meetings.

1. INTRODUCTION

Laughter occurs surprisingly frequently in meetings; analysis has demonstrated that it accounts for almost 10% of vocalization effort by time [1]. Its detection in conversational interaction presents an important challenge in meeting understanding, as laughter has been shown to be predictive of both emotional valence [2] and activation/involvement [3, 4].

Group laughter detection was first explored in [5], but its detection on nearfield channels and its correct attribution to specific participants has only recently been attempted [6, 7]. Authors of the latter reported that clearly audible laughter, sufficiently long in duration and temporally distant from the laugher's speech, can be detected with equal error rates below 10% when a priori channel activity knowledge is available. Although this represents a significant milestone, it is not clear how predictive of higher-level phenomena this subset of laughter is, relative to all laughter present.

The aim of the current work is the detection and participant-attribution of all *voiced* laughter on close-talk microphone channels in meetings, without reliance on prior knowledge of channel activity. For our purposes, voiced laughter is that which involves vocal fold excitation at any time during its production. It has been shown [8, 9] that voiced and unvoiced laughs are deployed contrastively under naturally occurring conditions, and there is some recent evidence [4] for meetings that voiced laughter, which accounts for the majority of laughter by time and count, is more predictive of emotional involvement in speech than is all laughter. By way of motivation, we explore this claim further in Section 3.

The remainder of this paper is organized as follows. We first describe in Section 2 the data; it is the same as that in [5, 6, 7, 10]. Section 4 describes our baseline laughter detector, whose performance is analyzed in Section 5. Experiments and a discussion of the results are presented in Sections 6 and 7, respectively. Section 8 summarizes this contribution.

2. DATA

As in other work on laughter detection in naturally occurring meetings [5, 6, 7, 10], we use the ICSI Meeting Corpus [11]. We retain the same division of the `Bmr` meetings into TRAINSET and DEVSET as proposed therein; we also report numbers for unseen EVALSET data, consisting of all of the `Bed` and `Bro` meetings.

The reference speech segmentation used in this work comes from the ICSI MRDA Corpus [12]; inter-word gaps shorter than 0.3 s were bridged. The reference segmentation of *laugh bouts* [9] comes from [13, 1]. Intervals during which a participant both speaks and laughs, known as "speech-laughs" [14], have been mapped to speech only, such that the categories of silence $\mathcal{N}$, speech $\mathcal{S}$, and laughter $\mathcal{L}$ are mutually exclusive.

3. RELATING LAUGHTER TYPE TO EMOTION

In this section, we aim to motivate the need to treat voiced and unvoiced laughter separately. The presence of voicing makes these two types of laughter acoustically distinct [9], but the two have also been shown to incur different degrees of simultaneous vocalization from multiple participants and to occur in different locales relative to the laugher's speech [1]. In particular, in [4], it was shown that voiced laughter appears to be more relevant to the inference of emotional involvement in speech than does all laughter. We explore this correlation further in Table 1, which shows experiments in which 60-second intervals of ICSI meetings were classified as either containing emotionally involved speech or not containing emotionally involved speech. The details of that classifier can be found in [4].

Panel ❶ in Table 1 lists accuracies obtained by guessing; in the first line, we show the accuracy obtained by guessing randomly, according to training set priors. These accuracies are provided to enable comparison with [3], where *chance-corrected accuracies* were provided for a similar task. The second line lists the much higher accuracies obtained by always guessing the majority class, to which all other numbers in the table should be compared.

In panel ❷, we reproduce the numbers found in [4], and additionally show accuracies obtained using a laughter segmentation from which simultaneously produced speech is excluded ($\mathcal{L} - \mathcal{S}$).

	Segmentation	Accuracy, %		
		train	dev	test
❶	guess, priors	61.3	60.9	61.2
	guess, majority	73.7	72.9	73.7
❷	$\mathcal{S}$	72.7	74.8	*75.2*
	$\mathcal{L} \cup \mathcal{S}$	75.5	78.0	*77.7*
	$\mathcal{L}$	79.2	80.4	*80.6*
	$\mathcal{L} \cap \mathcal{S}$	84.3	82.7	*83.0*
	$\mathcal{L} - \mathcal{S}$	79.5	80.4	80.8
❸	$\mathcal{L}_U$	77.7	76.4	77.4
	$\mathcal{L}_V \cup \mathcal{S}$	74.7	76.3	76.5
	$\mathcal{L}_V$	80.2	81.2	*81.4*
	$\mathcal{L}_V \cap \mathcal{S}$	84.4	82.9	85.6
	$\mathcal{L}_V - \mathcal{S}$	80.3	81.5	81.4

Table 1. Accuracy of SVM classification of 60-second meeting intervals as either containing or not containing involved speech, based on features drawn from several logical combinations of the speech and laughter reference segmentations. Numbers in italics are taken from [4], Table 1. "train", "dev", and "test" represent a different split of the corpus from that used for laughter detection in this work.

The results show that speech-laughs are particularly informative of temporally proximate involved speech, but other laughter also yields accuracies significantly exceeding majority-class guessing.

Panel ❸ shows several numbers not presented in [4]. In particular, we show classification accuracies for unvoiced laughter ($\mathcal{L}_U$); these are significantly lower than those for voiced laughter ($\mathcal{L}_V$, in panel ❷), as well as for all laughter ($\mathcal{L} \equiv \mathcal{L}_U \cup \mathcal{L}_V$). Also shown are several logical combinations of voiced laughter with speech, and all except $\mathcal{L}_V \cup \mathcal{S}$ exhibit accuracies which are higher than those of the corresponding combinations involving all laughter ($\mathcal{L}$, in panel ❷).

Taken together, these observations indicate that voiced laughter, as opposed to all laughter, is more informative of the emotional involvement exhibited in temporally proximate verbal production.

4. BASELINE LAUGHTER DETECTION SYSTEM

For the current work, we produce a baseline independent-participant vocal activity detector (VAD) based on our previous work on detecting the vocal activity state of all participants jointly [10]. The decoder operates at a frame step and a frame size of 100 ms [15]. The state of each participant evolves independently, along licensed trajectories of the hidden Markov model topology in Figure 1. Minimum duration constraints $\mathbf{T}_{(min)} = \{T^{\mathcal{S}}_{(min)}, T^{\mathcal{L}}_{(min)}, T^{\mathcal{N}}_{(min)}\}$ as shown in the figure were tuned to DEVSET in an earlier effort [10].

Each of $\mathcal{S}$, $\mathcal{L}$, and $\mathcal{N}$ is modeled by a 64-component Gaussian mixture model (GMM), defined over a feature vector of 41 elements. These include the first 13 Mel-frequency cepstral coefficients, channel-normalized using cepstral mean subtraction, their first- and second-order differences, and two features, known as minimum and maximum normalized log-energy differences (NLEDs) [16], used to mitigate the effects of crosstalk [17]. Transitions are governed by bigram probabilities, learned from the best forced-alignment Viterbi pass over all meetings in TRAINSET.

For completeness, in what follows we contrast the baseline independent-participant decoder performance with that of the joint-participant decoder introduced in [10]. In that system, the HMM topology was the Cartesian product of the topology shown in Figure 1, and decoding proceeded for all participants simultaneously,

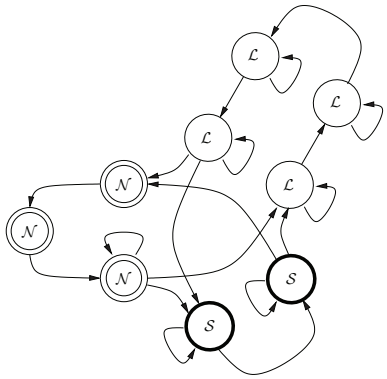

Fig. 1. Baseline HMM topology for 3-state VAD detection. $\mathcal{S}$, $\mathcal{L}$, and $\mathcal{N}$ represent speech, laughter, and silence, respectively.

allowing for inter-participant constraints to be directly modeled. To render search tractable, we enforced maximum simultaneous vocalization constraints such that e.g. at no time could more than 2 participants be speaking, or more than 3 participants be laughing.

5. ANALYSIS OF BASELINE PERFORMANCE

The performance of the baseline independent-participant and joint-participant decoders on DEVSET is shown in Table 2. For both systems, the majority of errors by absolute time is due to the misclassification of silence $\mathcal{N}$ as laughter $\mathcal{L}$. For the independent-participant decoder, the proportion of hypothesized laughter which is neither laughter nor speech is 70.8%; it is 62.6% for the joint-participant decoder. We note that a main difference between the two baseline decoders is that the independent-participant decoder hypothesizes more laughter and speech overall; we suspect this is largely due to crosstalk, more of which is eliminated in joint-participant decoding.

	Indep Decoder Hypos			Joint Decoder Hypos			Σ
	$\mathcal{N}$	$\mathcal{L}$	$\mathcal{S}$	$\mathcal{N}$	$\mathcal{L}$	$\mathcal{S}$	
$\mathcal{N}$	**655.8**	46.5	13.9	**685.4**	22.9	7.8	*716.2*
$\mathcal{L}'_U$	0.9	**4.2**	0.2	2.8	**2.4**	0.2	*5.4*
$\mathcal{L}'_V$	1.1	**8.9**	0.4	3.6	**6.5**	0.3	*10.4*
$\mathcal{L}_S$	0.0	**0.3**	0.5	0.1	**0.2**	0.5	*0.8*
$\mathcal{S}'$	3.4	5.8	**85.3**	11.9	4.5	**79.0**	*94.4*
Σ	*661.2*	*65.7*	*100.4*	*702.9*	*36.6*	*87.8*	*827.2*

Table 2. Confusion matrices for baseline independent-participant and joint-participant decoders on DEVSET. Reference labels, with laughter $\mathcal{L}$ broken down into unvoiced laughter excluding speech-laughs ($\mathcal{L}'_U$), voiced laughter excluding speech-laughs ($\mathcal{L}'_V$), and speech-laughs ($\mathcal{L}_S \equiv \mathcal{L} \cap \mathcal{S}$) are shown in rows; $\mathcal{S}'$ denotes speech excluding speech-laughs. Correctly classified time is shown in bold. All quantities are in minutes.

6. EXPERIMENTS

As analysis of the baseline independent-participant decoder shows, the majority of hypothesized laughter is actually silence. We suspect

that there are two reasons for this. First, laugh bouts in the training data (as in the test data) contain silent inter-call intervals [9]. A single acoustic model for laughter, as in our decoder, can therefore be expected to also consume silence during decoding. Second, and perhaps more importantly, because unvoiced laughter is acoustically similar to various breathing sounds which appear throughout the meeting recordings [17], we expect the laughter models to consume a significant amount of audio which correctly contains breathing, but is actually marked as silence in the references.

The experiments in this subsection involve two modifications to the independent-participant baseline, both aimed at improving precision. The baseline decoder performance, in terms of F-score for different vocalization types, is shown as IA1 in Table 3; we also show the performance of an identical system but with an ergodic HMM topology as IA0. The latter makes possible qualitative comparison with numbers published in [6].

6.1. Modeling unvoiced laughter together with silence

Systems IB0 and IB1 in Table 3 are identical in topology to systems IA0 and IA1, respectively; however, for these systems, frames marked as unvoiced laughter in TRAINSET have been removed from the training data for the $\mathcal{L}$ model and instead added to the training data for the $\mathcal{N}$ model. As a result, their $\mathcal{L}$ hypotheses are hypotheses of voiced laughter ($\mathcal{L}_V$) only. As the table shows, the resulting F-scores for $\mathcal{L}_V$ retrieval are higher than those for $\mathcal{L}$ retrieval by 3.7-3.8% absolute. We also show a third system, IB2, whose minimum duration constraints have been tuned to DEVSET; extending the minimum duration of voiced laughter improves $\mathcal{L}_V$ retrieval by 9.6% absolute. This is due to much higher precision stemming from the elimination of spurious frames, but also results in the elimination of short laugh bouts and therefore lower recall.

6.2. Explicit modeling of unvoiced laughter

Second, Table 3 shows the performance of the corresponding systems in which unvoiced laughter $\mathcal{L}_U$ is allowed its own model. It can be seen that the ergodic system IC0, and systems IC1 and IC2 which correspond in minimum duration constraints to systems IB1 and IB2, respectively, yield lower F-scores for voiced laughter, by 0.6-2.0% absolute. In these systems, both voiced laughter and unvoiced laughter are subject to the same minimum duration constraints. In contrast, for system IC3, the minimum duration constraints for the two laughter types are untied and tuned to DEVSET, resulting in a 1.7% absolute increase in F-score for voiced laughter over IB2.

We note that when voiced and unvoiced laughter are not subject to contrasting minimum duration constraints, as for systems IC0, IC1, and IC2, $\mathcal{L}_V$ detection is slightly lower than for the corresponding IB0, IB1, and IB2 systems, respectively. We believe this is due to increased model and task complexity.

6.3. Comparison with joint-participant decoding

We compare the performance of the participant-independent decoder with that of the joint-participant decoder, in Table 4, noting that system JA1 is the same as in [10].

As can be seen, $\mathcal{L}$ detection for systems JA0 and JA1 is higher than for the participant-independent systems IA0 and IA1; the same is true of system JB1 relative to IB1, with respect to $\mathcal{L}_V$ detection. In general, when unvoiced laughter is not explicitly modeled, joint-participant decoding appears to have a 0.8-2.9% absolute advantage over independent-participant decoding. However, once an explicit $\mathcal{L}_U$ model is introduced, the drop in $\mathcal{L}_V$ F-score of the JC systems relative to the JB systems is much higher than for their participant-independent counterparts. Furthermore, and most importantly, improvements due to longer minimum duration times are not directly possible for joint-participant decoding, due to the exponential growth of multi-participant topologies.

Sys	T_{min}, s				F, %				
	$\mathcal{S}$	$\mathcal{L}_V$	$\mathcal{L}_U$	$\mathcal{N}$	$\mathcal{S} \cup \mathcal{L}$	$\mathcal{S} \cup \mathcal{L}_V$	$\mathcal{S}$	$\mathcal{L}$	$\mathcal{L}_V$
IA0	0.1	0.1		0.1	75.4	—	87.4	30.9	—
IA1	0.2	0.4		0.3	76.3	—	87.6	32.6	—
IB0	0.1	0.1		0.1	—	78.3	86.6	—	34.6
IB1	0.2	0.4		0.3	—	79.0	86.9	—	36.4
IB2	0.1	2.5		0.4	—	81.7	86.6	—	46.0
IC0	0.1	0.1	0.1	0.1	71.6	83.5	87.3	25.9	32.9
IC1	0.2	0.4	0.4	0.3	72.6	83.9	87.6	27.6	34.4
IC2	0.1	2.5	2.5	0.4	76.3	84.7	87.2	35.4	45.4
IC3	0.1	3.2	1.4	0.4	74.6	85.2	87.5	32.4	47.7

Table 3. DEVSET F-scores of detecting vocalization ($\mathcal{S} \cup \mathcal{L}$), speech ($\mathcal{S}$), and laughter ($\mathcal{L}$) by VAD systems in which participants are decoded independently of one another; symbols as in the text.

Sys	T_{min}, s				F, %				
	$\mathcal{S}$	$\mathcal{L}_V$	$\mathcal{L}_U$	$\mathcal{N}$	$\mathcal{S} \cup \mathcal{L}$	$\mathcal{S} \cup \mathcal{L}_V$	$\mathcal{S}$	$\mathcal{L}$	$\mathcal{L}_V$
JA0	0.1	0.1		0.1	78.1	—	86.0	31.7	—
JA1	0.2	0.4		0.3	79.5	—	86.7	34.5	—
JB0	0.1	0.1		0.1	—	79.5	84.9	—	34.2
JB1	0.2	0.4		0.3	—	80.9	85.6	—	37.3
JC0	0.1	0.1	0.1	0.1	76.0	81.7	83.7	26.2	27.3
JC1	0.2	0.4	0.4	0.3	78.9	83.3	84.5	30.4	31.2

Table 4. DEVSET F-scores of detecting vocalization ($\mathcal{S} \cup \mathcal{L}$), speech ($\mathcal{S}$), and laughter ($\mathcal{L}$) by VAD systems in which participants are decoded jointly; symbols as in the text.

6.4. Generalization to unseen data

The performance of the IC3 independent-participant decoder on EVALSET, consisting of all the `Bed` and `Bro` meetings, is shown in Table 5, together with that of the IA1 and JA1 baselines.

Data		$p_{\mathcal{V}}(\mathcal{L})$, %	System			
			IA1	JA1	IC3	
			$\mathcal{L}$	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{L}_V$
DEVSET	`Bmr(3)`	14.94	32.6	34.5	32.4	47.7
EVALSET	`Bed`	7.53	16.7	17.0	14.5	22.0
	`Bro`	5.94	19.1	19.0	16.3	37.1

Table 5. Laughter ($\mathcal{L}$) and voiced laughter ($\mathcal{L}_V$) detection F-scores on several datasets using three different VAD systems. Also shown is the proportion $p_{\mathcal{V}}(\mathcal{L})$ of vocalization time spent in laughter.

As noted in [10], detection scores are closely correlated with the proportion of vocalization time spent in laughter. Second, they are

higher for Bro than for Bed, most likely due to the larger number of Bro participants also found in TRAINSET (cf. [10]). However, for both meetings types, system IC3 exhibits detection improvements which are commensurate with those observed for DEVSET, ie. an F-score for voiced laughter which is 29-95% relative higher than the F-score for laughter reported in [10]. We also note that the difference between the IA1 and JA1 baseline decoders is negligible for both Bed and Bro meeting types.

7. DISCUSSION

The specific errors committed by the IC3 system are shown in Table 6. As in the confusion matrix for the IA1 baseline in Table 2, the amount of time which is both classified as laughter ($\mathcal{L}_V \cup \mathcal{L}_U$) and transcribed as laughter is approximately the same (13.1-13.6 minutes). However, the IC3 system yields an $\mathcal{L}_V$ precision of 45.2%, compared to an $\mathcal{L}$ precision of 20.4% in the independent-participant baseline. A reverse trend can be observed for unvoiced laughter, where the precision for the IC3 system is only 6.7%. This shows that the separation of voiced and unvoiced laughter leads to increased precision in the detection of that subset of laughter which we have argued is more relevant to upstream conversation processing systems.

	$\mathcal{N}$	$\mathcal{L}_U$	$\mathcal{L}_V$	$\mathcal{S}$	Σ
$\mathcal{N}$	**649.4**	45.1	5.0	16.7	*716.2*
$\mathcal{L}_U$	0.8	**3.6**	0.7	0.4	*5.4*
$\mathcal{L}_V$	0.9	3.7	**5.6**	1.0	*11.2*
$\mathcal{S}$	3.6	2.2	1.1	**87.6**	*94.4*
Σ	*654.5*	*54.6*	*12.4*	105.6	

Table 6. Confusion matrix for the IC3 independent-participant decoder on DEVSET; format and symbols as in Table 2.

The confusion matrix in Table 6 also shows that the improved IC3 system still hypothesizes much voiced laughter as unvoiced laughter, and much silence as voiced laughter. An avenue of future research consists of constructing topologies for voiced laughter which explicitly model embedded unvoiced inter-call intervals.

8. CONCLUSIONS

We have presented the first attempt to detect and correctly attribute voiced laughter in conversational interaction using a large corpus of multi-party meetings. We described preliminary evidence which suggests that detection of voiced laughter, as opposed to that of all laughter, may be more useful to upstream conversation processing tasks. Our detection results on DEVSET show that modeling only voiced laughter leads to detection F-score improvements of 3.8% absolute over modeling all laughter. Furthermore, the experiments indicate that extending the allowed minimum duration of voiced laughter (to 2.5 seconds) is a more effective means than interaction modeling through joint-participant decoding towards eliminating false alarms; the observed increase in F-score is an additional 9.6% absolute. Finally, explicitly modeling unvoiced laughter, and imposing contrastive minimum duration constraints on the two laughter types, yields an additional 1.7% absolute improvement. Together, these modifications comprise a 46.3% relative increase in F-score, or a 22.4% relative reduction of DEVSET error, on this difficult unbalanced-prior task. Similar improvements were observed for the much larger and unseen EVALSET data.

9. REFERENCES

[1] K. Laskowski and S. Burger, "Analysis of the occurrence of laughter in meetings," in *Proc. INTERSPEECH*, Antwerpen, Belgium, 2007, ISCA, pp. 1258–1261.

[2] K. Laskowski and S. Burger, "Annotation and analysis of emotionally relevant behavior in the ISL Meeting Corpus," in *Proc. LREC*, Genoa, Italy, 2006.

[3] B. Wrede and E. Shriberg, "The relationship between dialogue acts and hot spots in meetings," in *Proc. ASRU*, St. Thomas, US Virgin Islands, 2003, pp. 180–185.

[4] K. Laskowski, "Modeling vocal interaction for text-independent detection of involvement hotspots in multi-party meetings," in *Proc. SLT*, Goa, India, 2008, pp. 81–84.

[5] L. Kennedy and D. Ellis, "Laughter detection in meetings," in *Proc. ICASSP Meeting Recognition Workshop*, Montreal, Canada, 2004, pp. 118–121.

[6] K. Truong and D. van Leeuwen, "Evaluating automatic laughter segmentation in meetings using acoustic and acoustic-phonetic features," in *Proc. ICPhS WS on Phonetics of Laughter*, Saarbrücken, Germany, 2007, pp. 49–53.

[7] M. Knox, N. Morgan, and N. Mirghafori, "Getting the last laugh: Automatic laughter segmentation in meetings," in *Proc. INTERSPEECH*, Brisbane, Australia, 2008, pp. 797–800.

[8] K. Grammer and I. Eibl-Eibesfeldt, "The ritualization of laughter," in *Naturlichkeit der Sprache und der Kultur: Acta colloquii*, W. Koch, Ed., Bochum, Germany, 1990, pp. 192–214, Brockmeyer.

[9] J.-A. Bachorowski, M. Smoski, and M. Owren, "The acoustic features of human laughter," *J. of Acoustical Society of America*, vol. 110, no. 3, pp. 1581–1597, 2001.

[10] K. Laskowski and T. Schultz, "Detection of laughter-in-interaction in multichannel close-talk microphone recordings of meetings," in *Proc. MLMI*, Utrecht, The Netherlands, 2008, Springer LNCS **5237**, pp. 149–160.

[11] A. Janin et al., "The ICSI Meeting Corpus," in *Proc. ICASSP*, Hong Kong, China, 2003, pp. 364–367.

[12] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI Meeting Recorder Dialog Act (MRDA) Corpus," in *Proc. SIGdial*, Cambridge MA, USA, 2004, pp. 97–100.

[13] K. Laskowski and S. Burger, "On the correlation between perceptual and contextual aspects of laughter in meetings," in *Proc. ICPhS WS on Phonetics of Laughter*, Saarbrücken, Germany, 2007, pp. 55–60.

[14] E. Nwokah, H.-C. Hsu, P. Davies, and A. Fogel, "The integration of laughter and speech in vocal communication: A dynamic systems perspective," *J. of Speech, Language & Hearing Research*, vol. 42, pp. 880–894, 1999.

[15] K. Laskowski and T. Schultz, "Modeling vocal interaction for segmentation in meeting recognition," in *Proc. MLMI*, Brno, Czech Republic, 2007, Springer LNCS **4892**, pp. 259–270.

[16] K. Boakye and A. Stolcke, "Improved speech activity detection using cross-channel features for recognition of multiparty meetings," in *Proc. INTERSPEECH*, Pittsburgh PA, USA, 2006, pp. 1962–1965.

[17] T. Pfau, D. Ellis, and A. Stolcke, "Multispeaker speech activity detection for the ICSI Meeting Recorder," in *Proc. ASRU*, Madonna di Campiglio, Italy, 2001, pp. 107–110.