MODELING VOCAL INTERACTION FOR TEXT-INDEPENDENT DETECTION OF INVOLVEMENT HOTSPOTS IN MULTI-PARTY MEETINGS

Kornel Laskowski

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

ABSTRACT

Indexing, retrieval, and summarization in recordings of meetings have, to date, focused largely on the propositional content of what participants say. Although objectively relevant, such content may not be the sole or even the main aim of potential system users. Instead, users may be interested in information bearing on conversation flow. We explore the automatic detection of one example of such information, namely that of hotspots defined in terms of participant involvement. Our proposed system relies exclusively on low-level vocal activity features, and yields a classification accuracy of 84%, representing a 39% reduction of error relative to a baseline which selects the majority class.

Index Terms— Speech processing, Meetings, Pattern classification, Information retrieval.

1. INTRODUCTION

Indexing, retrieval, and summarization in recordings of meetings and conversations have, to date, focused largely on the propositional content of what participants say. Although objectively relevant, such content may not be the sole or even the main aim of potential system users. Instead, users may be interested in information bearing on meeting or conversation flow. An example of this type of information is the varying degree of *involvement* exhibited by participants, individually or as a group.

Involvement in meetings has been defined as a meeting- and speaker-specific, prosodically marked, utterance-level characteristic [1], and shown to correlate with automatically computed acoustic cues. Human agreement on the perception of involvement was shown to be significantly above chance ($\kappa = 0.59$). Furthermore, involvement was determined to associate significantly with several dialog act subclasses, such as suggestions, jokes, and floor grabbers [2]. The authors of [1, 2] showed that such detailed dialog act knowledge yields chance-corrected accuracies of almost 40% for per-utterance classification of involvement; it was not stated whether this outperforms guessing the majority class.

In a practical information retrieval setting, single utterances may be too short a unit to remain interpretable without a broader temporal neighborhood. In [1], the authors proposed the notion of *hotspots*, or "intervals of about half a minute to one minute", exhibiting "high involvement on the part of two or more participants". However, under this definition (henceforth VERSION1), high involvement is neither a necessary nor a sufficient condition for the inclusion of an utterance in a hotspot. It is therefore not known to what extent VERSION1 hotspots can be detected automatically.

In more recent work [3], an alternate definition of hotspots was proposed (henceforth VERSION2) as "parts in conversations" in which "participants are more involved" [and/or] "there is a higher degree of interaction by participants who are trying to get the floor". Although it was shown in [4] that VERSION2 hotspots, whose temporal extent is a function of involved utterance duration, are associated with the degree of simultaneous vocalization from multiple participants (overlap), no evidence was presented to show that the observed differences are discriminative.

Our objective in the current work is to present a baseline hotspot detector. Using the extensive annotation of VERSION2 hotspot involvement (described in Section 2), but with the less subjective temporal support of VERSION1 hotspots, we propose a system which classifies 60-second intervals of meetings as either containing involved speech ($\mathcal{I}$) or not containing involved speech ($\neg\mathcal{I}$), and which relies only on very low-level vocal interaction features as might be available from a vocal activity detector. These features are described in Section 3, and the experiments presented in Sections 4 and 5 demonstrate that laughter is almost solely responsible for our reduction in error of 39.2% relative to a majority class baseline. Section 6 compares automatic versus human performance, and the impact of our results is briefly discussed in Section 7 and summarized in Section 8.

2. DATA

The current work makes use of the ICSI Meeting Corpus [5], which consists of 75 naturally occurring meetings with between 3 and $K_{max} = 9$ participants per meeting. The corpus is accompanied by orthographic transcription, as well as lexical forced time-alignment, dialog act annotation, and VERSION2 hotspot annotation [6]. Also available for the entire corpus is a segmentation of laughter [7], and an annotation of laughter as either voiced or unvoiced [8].

For each meeting, we have excluded the unconversational calibration intervals known as the `Digits` task and retained only the single longest contiguous portion of each meeting, amounting to 63 hours of multichannel audio. Division of the 75 meetings into a TRAINSET, a DEVSET, and an EVALSET was motivated as follows. There are two sets of manually produced hot spot labels available for excerpts from 11 meetings. To enable comparison with human performance (cf. Section 6), we placed these 11 meetings, which we subsequently refer to as EVALSUBSET, in EVALSET; we further augment EVALSET with 4 meetings of groups which are under-represented in EVALSUBSET relative to the rest of the corpus (`Bro010`, `Bro012`, `Bro016`, and `Bns002`). Of the remaining meetings, those whose numerical identifier ends with 1, 3, 5 or 7 were placed in DEVSET, and the remainder in TRAINSET.

VERSION2 hotspots differ from VERSION1 hotspots in duration; the former have an approximately log-normal distribution, with a most likely duration of 7 seconds and only one hotspot as long as 30 seconds. As suggested in the introduction, we appeal to the

VERSION1 hotspot duration guidelines (of 30-60 s) and detect not whether a dialog act contains involved speech, but whether a 60 second interval of meeting time does so. Reference labels for each interval are produced from the VERSION2 hotspot utterance tag [3]; an interval is given the label $\mathcal{I}$ only when it contains lexical productions marked as involved (with `b` or `b-`). Intervals are extracted from each meeting every 15 seconds. The resulting total number of intervals in the corpus is 15649; of these, 26.6% contain involved speech. The priors across the two labels $\mathcal{I}$ and $\neg\mathcal{I}$ are near-identical for all three of TrainSet, DevSet, and EvalSet.

3. VOCAL INTERACTION FEATURES

We propose to perform classification using features extracted from the vocal interaction record [9] of each meeting, ie. the time-aligned record of vocal activity from all K meeting participants; for this work, we use reference as opposed to automatic segmentations. We define vocal interaction $\mathbf{Q} \equiv \{\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_N\}$, where each column $\mathbf{q}_t$ is a K-element vector in $\{0,1\}$ representing the binary vocalization state of each participant k, $1 \leq k \leq K$, and N is the number of frames. $\mathbf{Q}$ is obtained by discretizing a continuous segmentation, such as implied by forced-alignment start and end times of spoken lexical items. The discretization process is described in [10]; here, we use a frame step and frame size of 10 ms and 20 ms, respectively.

We consider 5 different binary reference segmentation types:

- $\mathcal{S}$: speech (words and word fragments) versus non-verbal vocalization and silence, obtained from the forced-alignments in the ICSI MRDA Corpus [6];
- $\mathcal{L}$: laughter versus verbal and other non-verbal vocalization and silence, as in [7];
- $\mathcal{L}_V$: voiced laughter versus unvoiced laughter, verbal and other non-verbal vocalization, and silence, as in [8];
- $\mathcal{S} \cup \mathcal{L}$: speech or laughter versus other vocalization and silence, computed from $\mathcal{S}$ and $\mathcal{L}$; and
- $\mathcal{S} \cap \mathcal{L}$: "laughed speech" [11] versus other speech, non-verbal vocalization, and silence, computed from $\mathcal{S}$ and $\mathcal{L}$.

For a particular segmentation type, and a particular 60 s interval, we extract both static and dynamic features. Our static features are of two types:

- $\{p_j^{\mathcal{V}}\}$: the proportion of interval duration for which each participant vocalizes, sorted by decreasing magnitude and padded with zeros to $K_{max} = 9$; and
- $\{o_j^{\mathcal{V}}\}$: the proportion of interval duration for which at least j participants vocalize simultaneously, for $1 \leq j \leq 9$.

Zero-padding allows for length-consistent feature vectors when meetings contain fewer than $K_{max} = 9$ participants.

In computing dynamic features, we wish to consider variations in the transition probabilities governing entry and egress out of various multiparticipant vocal activity states. We measure this variation using a parametric *spin glass* [12] model $\mathcal{M}_{SG}$ of vocal interaction, proposed in [13]. The model estimates the conditional probabilities that each particular participant k, $1 \leq k \leq K$, vocalizes at time t given $\mathbf{q}_{t-1}$, the vocalization state of all participants at the previous instant:

$$P_k\left(\mathbf{q}_t\left[k\right] = \mathcal{V} \,|\, \mathbf{q}_{t-1}\right) \quad = \quad \frac{1}{1 + e^{-\left(b_k + \sum_{l=1}^{K} w_{kl}\,\mathbf{q}_{t-1}[l]\right)/T_k}} \;. \quad (1)$$

The model is thus entirely governed by the parameters $\{\mathbf{b} = \{b_k\}, \mathbf{W} = \{w_{kl}\}, \mathbf{T} = \{T_k\}\}$, which are estimated as follows.

First, $\mathbf{b}$ and $\mathbf{W}$ are estimated from the entirety of the duration of $\mathbf{Q}$. During this pass, the *pseudo-temperatures* T_k are clamped to an arbitrary but fixed constant $T_{ref} \equiv 1$. The estimated values of $\mathbf{b}$ and $\mathbf{W}$ can be said to represent long-observation-time norms of vocalization timing; they are specific to both the meeting, and to the individual participants. In a second pass, $\mathbf{b}$ and $\mathbf{W}$ are kept fixed at their long-observation-time values and, for a particular 60 s interval of interest, the T_k are allowed to vary to account for (transient) departure from long-observation-time norms. As can be seen from Equation 1, T_k represents a non-linear interpolation parameter of the probabilities $P_k\left(\cdot\right)$ with $1/2$ (i.e., randomness). For $T > T_{ref}$, each probability in Equation 1 is pulled towards $1/2$, whereas for $T < T_{ref}$ each is pulled away. In the current work, we use the inferred T_k, $1 \leq k \leq K$, as our dynamic features. In both passes, we infer parameters using gradient descent in the negative logarithm of the likelihood of $\mathbf{Q}$ given $\mathcal{M}_{SG}$, where

$$P\left(\mathbf{Q}|\mathcal{M}_{SG}\right) \quad \doteq \quad \prod_{t=1}^{T}\prod_{k=1}^{T} P_k\left(\mathbf{q}_t\left[k\right] | \, \mathbf{q}_{t-1}, \mathcal{M}_{SG}\right) \;. \quad (2)$$

ie. we assume that participants are conditionally independent and that the conversation is first-order Markovian. We note that the problem of parameter inference can be reformulated as *logistic regression* and the parameters of $\mathcal{M}_{SG}$ can be identically inferred using the *reweighted least squares* algorithm [14].

In the current setting, we explore 2 types of T_k features:

- $\{T_k^{PI}\}$: K participant-specific measures of departure from a participant-independent (PI) norm (ie. $b_k = b$, $w_{kk} = w_+$, and $w_{kl} = w_-$, $\forall l \neq k$, for 3 degrees of freedom), sorted by decreasing magnitude and padded with zeros to $K_{max} = 9$; and
- $\{T_k^{PS}\}$: K participant-specific measures of departure from participant-specific (PS) norms (ie. untied $\mathbf{b}$ and $\mathbf{W}$, for $K + K^2$ degrees of freedom), sorted by decreasing magnitude and padded with zeros to $K_{max} = 9$.

Both dynamic feature types are meeting-specific, as they are inferred from long-observation-time norms of each meeting separately.

4. EXPERIMENTS

Since no previous experimental results are available, we select as our baseline simply choosing the majority class (i.e., no involvement). On EvalSet, the resulting accuracy is 73.67%. We note that chance guessing, informed by the TrainSet majority class prior, yields an EvalSet accuracy of 61.23%.

Our factorial experiment is shown in Table 1. All 4 feature types (of 9 features each) are drawn from all 5 segmentation types, resulting in 20 cells. For each cell, we train a support vector machine[1] (SVM) on TrainSet, and perform single feature forward selection to maximize accuracy on DevSet; all feature values are Z-normalized to facilitate SVM learning. Table 1 shows EvalSet accuracies only, together with the number of features selected for that cell.

We make 4 broad observations regarding these results. First, we note that, except for one cell ($\{T_j^{PS}\}$ for $\mathcal{S}$), all feature and segmentation types yield accuracies which exceed majority class guessing, and all outperform chance guessing by 31.7% to 59.3% relative.

[1] We use SVM^{light}, available from Thorsten Joachims at `http://svmlight.joachims.org/` (downloaded on 5 August 2008 at 1430hrs GMT). Only a linear kernel with a biased hyperplane was explored; all other toolkit parameters were left at their default values to enable subsequent trend analysis.

Second, looking at the static feature types only, there exists a clear progression in accuracy towards increasingly smaller subsets of the laughter segmentation $\mathcal{L}$. We note that $(\mathcal{L} \cap \mathcal{S}) \subseteq \mathcal{L}_V \subseteq \mathcal{L} \subseteq (\mathcal{L} \cup \mathcal{S}) \supseteq \mathcal{S}$. The best single-cell accuracy of 83.0% can be found for just two features in $\{p_j^{\mathcal{V}}\}$ from $\mathcal{L} \cap \mathcal{S}$, and this accuracy decreases as supersets of the $\mathcal{L} \cap \mathcal{S}$ segmentation are considered.

Third, dynamic features outperform static features only infrequently, and only by small amounts. In particular, the $\{T_j^{PS}\}$ features appear to be uncompetitive as a whole. This is due to their poor generalization to EvalSet. However, as we will show in the next section, dynamic features appear complementary even alongside the best-performing static features.

Finally, feature combination frequently results in improved EvalSet performance. Table 1 shows the effect of combining across feature types in the rightmost column, and across segmentation types in the bottom row; feature selection in these cells is performed on 36 and 45 features, respectively. Cases in which feature combination results in a degradation of EvalSet accuracy represent a mismatch of feature relevance between DevSet and EvalSet. Feature selection performed on all 180 features, resulting in 84.0%, is only 0.2% absolute lower than the best accuracy observed anywhere in the table. We treat 84.0% as the final performance measure on unseen data; it represents a 39.2% relative reduction of error over guessing the majority class, and a 58% relative error reduction over chance informed by TrainSet priors.

Segm. Type	Feature Type				
	Static		Dynamic		all
	$\{p_j^{\mathcal{V}}\}$	$\{o_j^{\mathcal{V}}\}$	$\{T_j^{PI}\}$	$\{T_j^{PS}\}$	
$\mathcal{S}$	75.2 (3)	73.9 (3)	75.3 (1)	73.5 (4)	75.5 (7)
$\mathcal{L} \cup \mathcal{S}$	77.7 (4)	80.1 (9)	77.1 (1)	76.5 (1)	80.0 (3)
$\mathcal{L}$	80.6 (1)	81.2 (6)	80.8 (1)	75.5 (1)	80.0 (5)
$\mathcal{L}_V$	81.4 (2)	82.1 (6)	81.6 (1)	75.9 (6)	81.9 (8)
$\mathcal{L} \cap \mathcal{S}$	83.0 (2)	82.1 (6)	78.1 (1)	79.0 (4)	84.2 (7)
all	83.4 (9)	82.6 (2)	82.7 (8)	75.4 (3)	84.0 (5)

Table 1. Classification accuracy on EvalSet using a linear-kernel SVM, for static and dynamic feature types (in columns) computed from different segmentation types ("Segm.", in rows). Each cell shows the accuracy achieved in % by an optimal feature subset identified using DevSet; the number of selected features, out of a total of $K_{max} = 9$ available in each non-"all" cell, is shown in parentheses.

5. FEATURE ANALYSIS

We briefly explore the relative merits of the 5 best features responsible for our final EvalSet accuracy of 84.0%, in Table 2. Features are ranked according to the sequence in which they were incrementally selected. Because selection is not based on TrainSet accuracy, each feature's rank reflects to some extent how well it generalizes. For comparison, we also show the magnitude of the weighted sum of all learned support vectors in column 2 (and its rank in column 3), a function of TrainSet only.

The table shows that the $\mathcal{L} \cap \mathcal{S}$ segmentation offers the most discriminating feature, namely the vocalization proportion of the most $(\mathcal{L} \cap \mathcal{S})$-vocalizing participant, $p_1^{\mathcal{V}}$. This feature alone is responsible for 95% of the absolute accuracy improvement of all five features over majority class guessing. Although many other features show individual accuracies which are in the same range, most of those appear to be redundant given the first feature, and are not selected. Only one other selected feature yields an individual accuracy of 81.58%; all subsequently selected features have individual accuracies ≤75.0%.

#	SVM		Segm.	Feat.	j	Acc, %abs	
	Weight	Rank	Type	Type		alone	cum.
1	0.39	3	$\mathcal{L} \cap \mathcal{S}$	$p_j^{\mathcal{V}}$	1	83.5	83.5
2	0.16	18	$\mathcal{L}_V$	T_j^{PI}	3	81.6	83.7
3	0.03	105	$\mathcal{L}$	$p_j^{\mathcal{V}}$	7	75.0	84.5
4	0.07	56	$\mathcal{L}_V$	T_j^{PS}	8	73.8	83.9
5	0.04	92	$\mathcal{S}_S$	T_j^{PS}	8	73.7	84.0

Table 2. Feature ranking for the 5 features judged as optimal, and SVM weights and weight ranks. EvalSet accuracies ("Acc") for each feature alone and in combination with more relevant features ("cum.") are given in % absolute. Features are identified by their (sorted) position j in the 9-element vector representing features of "Feat. Type" drawn from segmentation of "Segm. Type".

We note also that features ranked 3 through 5 in the table are the 7th or 8th largest features in their respective 9-element vectors, indicating that they are useful for meetings with 7 or more participants. This suggests that accuracies may be improved when classification decisions are conditioned on meeting group size.

6. COMPARISON WITH HUMAN PERFORMANCE

The detection of involvement is known to be a difficult and subjective task, as shown in an analysis of 13 meetings in which the majority of speech was contributed by 6 same participants [1]. Utterance-level agreement between any two native English-speaking labelers (out of 6) who were familiar with the meeting participants was shown to be $\kappa = 0.63$; non-native labelers, also familiar with the participants, appeared to agree at only $\kappa = 0.52$.

Subsequent analysis on EvalSubSet (a more varied subset of the corpus than used in [1]) using two labelers showed that per-utterance agreement on involvement is $\kappa = 0.63$, while that for "grown" hotspot intervals [15] is $\kappa = 0.67$. In this section, we explore the agreement of the same two labelers (here, A and B) and on the same data as [15], on whether a 60 s interval is $\mathcal{I}$ or $\neg\mathcal{I}$. For each interval in EvalSubSet, we extract A and B labels as described in Section 2 for the final consensus labels; we also compute $A \cup B$ and $A \cap B$ to gain insight into consensus creation on this task. All four sets of labels, the final consensus labels, and those produced by our final system are shown in Table 3.

	B	$A \cup B$	$A \cap B$	`ref`	`hyp`
A	0.68	0.91	0.77	0.84	0.59
B		0.78	0.90	0.83	0.57
$A \cup B$			0.69	0.85	0.58
$A \cap B$				0.81	0.57
`ref`					0.54

Table 3. Pair-wise inter-labeler agreement measures (κ) on EvalSubSet between two human judges (A and B), their logical combinations ($A \cup B$ and $A \cap B$), the final consensus labels (`ref`) used as reference, and our automatic system labels (`hyp`).

As Table 3 shows, inter-labeler agreement on our task is 0.68, similar to that for utterance-level involvement. We note that because agreement between B and $A \cap B$ is near unity, and that between A and $A \cup B$ is near unity, B's involvement judgments appear to be a subset of A's. However, comparison between $A \cup B$, $A \cap B$, and the consensus labels `ref` indicates that the latter are a relatively complex combination of the two annotators' labels.

Table 3 also shows that agreement between our automatic system and the human-produced consensus labels is $\kappa = 0.54$, and that between our system and either human taken alone is $\kappa \in [0.57, 0.59]$, slighly higher. We note that 54% is also the chance-corrected accuracy measure used in [2], where the maximum attained using detailed dialog act knowledge was shown to be just below 40%.

7. DISCUSSION

The experiments presented here, using reference speech and laughter segmentations, indicate that laughter is temporally collocated with prosodic involvement and thus important for its detection. This presents a strong motivating case for technological advancement in laughter detection for meetings [16, 17, 18, 19]. At the current time, meeting laughter detectors contrast between speech and laughter, rendering $\mathcal{L} \cap \mathcal{S} = \emptyset$. However, as we have shown, $\mathcal{L}_V$ yields features with quite similar hotspot detection performance. Cursory analysis in [19] suggests that voiced laughter is easier to detect than all laughter, primarily because unvoiced laughter is frequently confused with breath and contact noise.

For vocal activity systems which do not discriminate between laughter and speech, but do discriminate between vocalization and silence (cf. results for $\mathcal{L} \cup \mathcal{S}$ in Table 1), our vocal interaction features yield an accuracy of 80.0% on unseen data, representing a 24% reduction of error over majority class guessing. This makes them more informative than detailed dialog act tagging; preliminary accuracies for our features drawn from an $\mathcal{S}$ segmentation containing only speech found in hotspot-correlated dialog act types [2] are 1-2% absolute above majority class guessing. However, current state-of-the-art dialog act taggers consider only coarse dialog act classes.

8. CONCLUSIONS

We have presented a system for the classification of 60 s intervals as either containing or not containing involved speech. The system is suitable for real-time deployment and relies on only low-level features, as may be extracted from the output of a vocal activity detector. The most informative features are those pertaining to "laughed speech", voiced laughter, and laughter in general, in descending order. On 12.5 hours of unseen meeting data, the system yields an accuracy of 84.0%, representing a relative reduction in error of 39.2% over a majority class baseline. Chance-corrected agreement between our automatic labels and labels produced by human annotators is 10% absolute lower than that between annotators, and 6% absolute higher than agreement among non-native annotators on the corresponding per-utterance task with a similar range of agreement.

9. ACKNOWLEDGMENTS

We would like to thank Liz Shriberg for access to the ICSI MRDA Corpus and to internal ICSI documentation, and for several helpful suggestions during the preparation of this manuscript. We also thank Tanja Schultz for useful discussion.

10. REFERENCES

[1] B. Wrede and E. Shriberg, "Spotting "hot spots" in meetings: Human judgments and prosodic cues," in *Proc. EUROSPEECH*, 2003, pp. 2805–2808.

[2] B. Wrede and E. Shriberg, "The relationship between dialogue acts and hot spots in meetings," in *Proc. ASRU*, 2003, pp. 180–185.

[3] B. Wrede, S. Bhagat, R. Dhillon, and E. Shriberg, "Meeting Recorder project: Hot spot labeling guide," Tech. Rep. TR-05-004, ICSI, Berkeley CA, USA, 12 May 2005.

[4] Ö. Çetin and E. Shriberg, "Overlap in meetings: ASR effects and analysis by dialog factors, speakers, and collection site," in *Proc. MLMI*, 2006, vol. 4299 of *Springer LNCS*.

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

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

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

[8] 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*, 2007, pp. 55–60.

[9] J. Dabbs and R. Ruback, "Dimensions of group process: Amount and structure of vocal interaction," *Advances in Experimental Psychology*, vol. 20, pp. 123–169, 1987.

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

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

[12] K. Fischer and J. Hertz, *Spin Glasses*, Cambridge University Press, Cambridge, UK, 1991.

[13] K. Laskowski, "Quantifying transient departure from conversation- and participant-specific norms of talkspurt deployment timing," *in preparation*.

[14] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*, Springer, 2004.

[15] B. Wrede and E. Shriberg, "Reliability analysis for hot spot annotations in the MRDA Corpus," *Internal document*, ICSI, Berkeley CA, USA, 11 April 2005.

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

[17] K. Truong and D. van Leeuwen, "Automatic discrimination between laughter and speech," *Speech Communication*, vol. 49, no. 2, pp. 144–158, 2007.

[18] M. Knox and N. Mirghafori, "Automatic laughter detection using neural networks," in *Proc. INTERSPEECH*, 2007, pp. 2973–2976.

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