Using Linguistic Information to Detect Overlapping Speech

Jürgen T. Geiger[1], *Florian Eyben*[1], *Nicholas Evans*[2], *Björn Schuller*[3,1] *and Gerhard Rigoll*[1]

[1]Institute for Human-Machine Communication, Technische Universität München, Germany
[2]Multimedia Communications Department, EURECOM, Sophia Antipolis, France
[3]Institute for Sensor Systems, University of Passau, Germany

{geiger,eyben,rigoll}@tum.de, evans@eurecom.fr, bjoern.schuller@uni-passau.de

Abstract

Overlapping speech is still a major cause of error in many speech processing applications, currently without any satisfactory solution. This paper considers the problem of detecting segments of overlapping speech within meeting recordings. Using an HMM-based framework recordings are segmented into intervals containing non-speech, speech and overlapping speech. New to this contribution is the use of linguistic information, where spoken content is used to improve overlap detection. Using language models for speech and overlap, an overlap score is created for every spoken word and used as an additional feature within the HMM framework. Experiments conducted on the AMI corpus demonstrate the potential of the proposed linguistic features.

Index Terms: Speech Overlap Detection, Spontaneous Speech, Speaker Diarization, Language Modelling

1. Introduction

Overlapping speech, i. e., segments where two or more speakers are simultaneously active, remains a major source of error in many speech processing applications, e. g., speech recognition or speaker diarization [1, 2, 3, 4]. Particularly in spontaneous, conversational speech, overlap occurs at speaker turn points or as backchannel utterances or interruptions, for example. Overlap can degrade the performance of speech processing systems which assume only one active speaker. With speaker diarization, for example, overlap can lead to speaker model impurities which indirectly contribute to diarization error through degraded clustering performance. Furthermore, overlap directly provokes increases in the missed speaker rate. The successful detection of overlap thus has the potential to improve the robustness of speech processing applications under realistic conditions.

Overlap detection has attracted growing attention in the recent past, especially in the context of speaker diarization. The earliest prior work analysed the general influence of overlapping speech on diarization performance [3, 5, 6]. The first hidden Markov model (HMM)-based overlap detection system using mainly spectral features (MFCCs, RMS energy, LPC residual energy, and diarization posterior entropy) was reported in 2008 [7]. The work showed how speaker diarization performance can be improved by excluding overlapping speech segments from those used in speaker modelling and then by attributing overlapping segments to at most two speakers. This work was extended in [8] and [9] which assessed the use of new features including spectral flatness, the harmonic energy ratio, modulation spectrogram features, kurtosis, zero-crossing rate and harmonicity. More recently, spatial [10, 11] and prosodic [12] features have been investigated. Our own previous work introduced the use of convolutive non-negative sparse coding (CNSC) and other spectral, energy and voicing-related features [13, 14, 15]. An alternative approach which uses the output of a voice activity detection component and the silence distribution to detect overlap was reported in [16]. This work was extended by exploiting long-term conversational features for overlap detection [17]. Finally, there is related, prior work in speaker recognition [18], where overlap detection is used as a preprocessing step for speaker recognition. In [19, 20], overlapping speech was analyzed beyond acoustic properties. Despite the level of recent interest, none of the above approaches gives satisfactory performance; overlap detection remains an unsolved problem.

Almost all of this prior work focuses on the use of acoustic cues to detect overlap. While backchannel utterances auch as "yeah" or "mm-hmm" occur frequently in spontaneous, overlapping speech [21], they do not necessarily overlap acoustically with competing speech. Our work in [14] also shows that especially short overlap segments, which typify instances of overlap such as those from backchannel, are particularly difficult to detect using acoustic features on their own. New approaches, exploiting different cues are thus required.

The linguistic contents of a speech signal has previously been used for example for speech emotion recognition [22] or for speaker diarization [23]. This contribution considers the use of higher-level information for overlap detection. The spoken content of the audio signal is one such source of information. Central to the idea is the use of language models to detect backchannel words and other language characteristics which typify instances of overlap. Thus this paper presents a new approach to overlap detection using linguistic features, where language models are used to characterise the spoken content in overlapping speech and speech with only a single active speaker. The language models are then used to assign scores to each word in a dictionary and thus to estimate the probability that the linguistic content reflects overlapping speech. Using the output of an automatic speech recognition (ASR) system, such scores can be used within a conventional HMM framework to detect overlap. Experiments conducted on the AMI Corpus show that the proposed linguistic features lead to improved performance.

The remainder of this paper is structured as follows: In Section 2 we present the overlap detection system and the employed energy, spectral, voicing related and CNSC-based audio features. Section 3 gives an insight into our proposed linguistic features. Experiments and results are described in Section 4 before conclusions are given in Section 5.

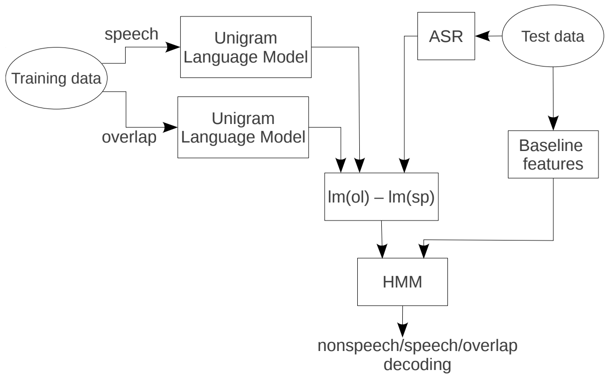

Figure 1: System overview for the overlap detection system

2. System Overview

An overview of the proposed system is illustrated in Figure 1. It shows the integration of the new linguistic feature into an HMM-based overlap detection system with baseline features.

2.1. Overlap detection system

We use an HMM-based overlap detection system similar to that in [7]. Speech, nonspeech and overlapping speech are each modeled by a three-state HMM. Observations are modeled by a multivariate Gaussian Mixture Model (GMM) with diagonal covariance matrices. Due to unbalanced training data, mixtures in the speech model have 256 components, while those in the nonspeech and overlap models have 64 components. The models are trained with an iterative mixture splitting technique with successive re-estimation. During decoding, transitions between nonspeech and overlap are forbidden, as are self-transitions, e. g., from overlap to overlap. The log-likelihood transition penalty from speech to overlap, also referred to as the overlap insertion penalty (OIP), is tuned to control the trade-off between precision and recall performance.

2.2. Baseline features

Through previous work reported in [15], we performed feature selection on a large set of candidate features to identify those best suited to overlap detection. The candidate feature set is derived from the baseline feature set we provided for the first audio-visual emotion challenge (AVEC) in 2011 [24]. As shown in Table 1, a total of 23 features were selected. They can be categorised into energy and spectral features, voicing-related features and features based on CNSC.

Conventional mel-frequency cepstral coefficients (MFCCs) have been used for overlap detection in prior work [7]. Since overlap contains speech from multiple speakers, and is thus often of higher volume than speech from a single person, energy features are natural indicators. Jitter and shimmer are measures of fluctuations in fundamental frequency and amplitude respectively, and are thus also ideally suited.

CNSC [25] is an approach to represent non-negative, multivariate data as a linear combination of lower rank bases. Their use for overlap detection was first reported in [13]. A non-negative matrix $D \in \mathbb{R}^{\geq 0}_{M \times N}$ is represented as:

$$D \approx \sum_{p=0}^{P-1} W_p \overset{p\rightarrow}{H}, \tag{1}$$

Energy & spectral features (18)
MFCC 1-12
loudness (auditory model based)
energy in band 250 - 650 Hz
energy in band 1 kHz - 4 kHz
spectral flux
spectral kurtosis
spectral harmonicity
Voicing-related features (3)
probability of voicing
jitter
shimmer (local)
CNSC-based features (2)
CNSC energy ratio
CNSC total energy

Table 1: Baseline energy, spectral, voicing-related and CNSC-based features

where $W_p \in \mathbb{R}^{\geq 0}_{M \times R}$ and $H \in \mathbb{R}^{\geq 0}_{R \times N}$ are the bases and base activations, respectively. P is the convolutional range. The column shift operator $\overset{p\rightarrow}{\cdot}$ shifts p columns of H to the right. The bases and activations are learned such that the regularised least square error between the original matrix D and the recomposition WH is minimised. In all work reported here, we used an approach proposed in [26, 27] to learn bases. Bases W are learned for each speaker in an audio document using spectral magnitude features extracted from segments of preferably pure (non-overlapping) speech. Their detection, however, is the very goal of this work and thus, in practice, they are identified using speaker diarization.

The base patterns of each speaker are concatenated to create a global basis. When the spectral magnitude features of a recording are decomposed or projected onto each speaker basis, the resulting activations H reflect each speaker's activity. Summing over all activations for a given speaker s leads to an estimate of the speaker energy $E_j(s)$ for frame j. The first CNSC-based feature is the CNSC energy ratio

$$ER_j = \frac{E_j(\hat{s}_2)}{E_j(\hat{s}_1)} \tag{2}$$

which reflects the difference in activation energy between the two most active speakers. The second CNSC-based feature is the CNSC total energy

$$ET_j = \sum_{s \in S} E_j(s) - \frac{f}{|J_{sp}|} \sum_{j \in J_{sp}} \sum_{s \in S} E_j(s) \tag{3}$$

which is the sum of all speaker energies, normalised by the mean over all the speech frames J_{sp}. Here, f is a regularization factor tuned on held-out development data. Full details of the CNSC feature extraction are reported in [15].

Finally, the feature set is augmented with first order regression coefficients and normalised using the statistics of the training set to have zero mean and unity variance.

3. Linguistic Cues for Overlap Detection

The motivation to use linguistic features for overlap detection stems from the hypothesis that some words are more likely during overlap than others and thus that spoken words can be used to detect overlap. This is instinctively the case for floor grabbers, backchannel and interruptions, for example. We use language models to characterise the distribution of words used during non-overlapping and overlapping speech.

We now turn to the left-hand side of Figure 1. Unigram language models are learned for single-speaker and overlapping speech using independent training data and ground-truth, word-level transcriptions. Test data is processed with an ASR system to produce a comparable word-level transcription. It is used together with the two language models to estimate a score which reflects the relative likelihood that the signal contains speech from a single, or more than a single speaker. The score is combined with the baseline features and used in an HMM detection system which classifies the signal as either non-speech, speech (from a single speaker), or overlap.

Generally, a language model is used to assign a probability to a sequence of words $p(w_1, ..., w_m)$. Of practical use are N-gram language models, where the probability for a word depends on the last $N-1$ words. In speech recognition, it is common to use bigram or trigram language models. In contrast, a unigram language model describes only the probability of a single word $p(w)$. Using training data for single-speaker speech and overlapping speech, we compute such unigram language models for speech and overlap denoted $p(w|sp)$ and $p(w|ol)$, respectively. In practice, however, to allow for the automatic recognition of only a single word at a time, only the longest of the spoken words during an interval of overlap is taken into account.

The detection of overlap using linguistic content is therefore equivalent to determining the probability of overlap $p(ol|w)$ for any given word. With Bayes' theorem, this can be expressed as:

$$p(ol|w) = \frac{p(w|ol) \cdot p(ol)}{p(w)}. \tag{4}$$

Since $p(ol)$ is independent of the word and $p(w)$ is approximated by the language model probability for single-speaker speech $p(w|sp)$, Eq. (4) can be reduced to

$$p(ol|w) \approx \frac{p(w|ol)}{p(w|sp)}. \tag{5}$$

Using log-likelihoods, the probability of overlap is finally expressed as:

$$s(w) = \log(p(ol|w)) \approx \log(p(w|ol)) - \log(p(w|sp)). \tag{6}$$

Eq. (6) reflects the relative likelihood of speech from a single or multiple speakers. The value of $s(w)$ can be computed for every word in the training set. A ground-truth, speaker-level annotation is then used to construct a word-level, look-up table and to identify those words which are most and least indicative of overlap. Words which occur more often in overlapping speech and least often in non-overlapping speech are assigned higher scores whereas those which occur most often in non-overlapping speech and least often in overlapping speech are assigned smaller scores. The score forms the new feature used for overlap detection.

Words typically used in overlap segments (e.g., in backchannel utterances) like "mh-hmm", "uh-huh","um", "yeah", "yep", "okay", "nope", "but", "wait" were all shown to be among those with the highest scores. These observations support the hypothesis that linguistic cues have potential for overlap detection.

For the training data, the reference transcriptions are used to determine $s(w)$ on a frame-by-frame basis according to Eq. (6). The value of $s(w)$ is then added to the baseline feature set.

Test set			
EN2003a	EN2009b	ES2008a	ES2015d
IN1008	IN1012	IS1002c	IS1003b
IS1008b	TS3009c		

Table 2: Meetings from the AMI evaluation dataset used for the tests

As with all other features, first order regression coefficients are added and the features are normalised. Recognised words in test data are assigned the corresponding value of $s(w)$ from the look-up table determined with training data. For both training and test data, a value of $s(w) = 0$ is assigned in the abscence of any recognised words.

To assess the potential of linguistic cues for overlap detection independently from the performance of an ASR system, we assessed the performance of the proposed system using a so-called *oracle-style* ASR, or ground-truth transcripts. Accordingly, i. e., to simulate the output of a more realistic ASR system, the transcripts are purged of overlapping words so as to retain only those with the largest duration. The output of such an *oracle-style* ASR system thus corresponds to that of a perfect ASR system, but capable only of single-word recognition. This is the same strategy as applied for language model training.

4. Experiments

4.1. Experimental Setup

Experiments were conducted using the AMI Corpus [28]. We used a subset of 40 meeting recordings for HMM training, 6 for tuning and 10 for testing. Table 2 lists all meeting recordings contained in the test set. Language models are estimated using a larger training set of 161 meeting recordings. The length of the recordings in the test set varies between 17 and 57 minutes and in total, the length of the test set is more than 6 hours. All are single-channel, far-field microphone recordings – the most challenging scenario. On average the amount of overlapping speech is in the order of 20 %.

We apply the following system parameters, based on our previous experience: Energy, spectral and voicing-related features are computed every 20 ms. A window size of 60 ms is applied for MFCC and voicing-related features, whereas other energy and spectral features are determined using a window size of 25 ms. CNSC is applied using magnitude spectra computed from 40 ms windows with a window shift of 20 ms. We used $R = 35$ bases per speaker, a convolutional range of $P = 4$ and a sparseness parameter $\lambda = 0.05$. The regularisation factor in Eq. (3) is set to $f = 1.2$. Speaker bases are learned using speaker-specific training data obtained with the LIA-Eurecom speaker diarization system [29].

System performance is measured in terms of frame-wise precision, recall and detection error, which is equivalent to the total duration of missed and false alarm overlap time divided by the reference overlap time. Note that, since overlap makes up only around 20 % of the recordings, false positive detections can result in an overlap detection error above 100 %. For a typical application such as overlap handling for speaker diarization, the detection error is the most meaningful metric.

4.2. Results

We report results for four different combinations of MFCC features, AVEC features, CNSC-based features, and the newly pro-

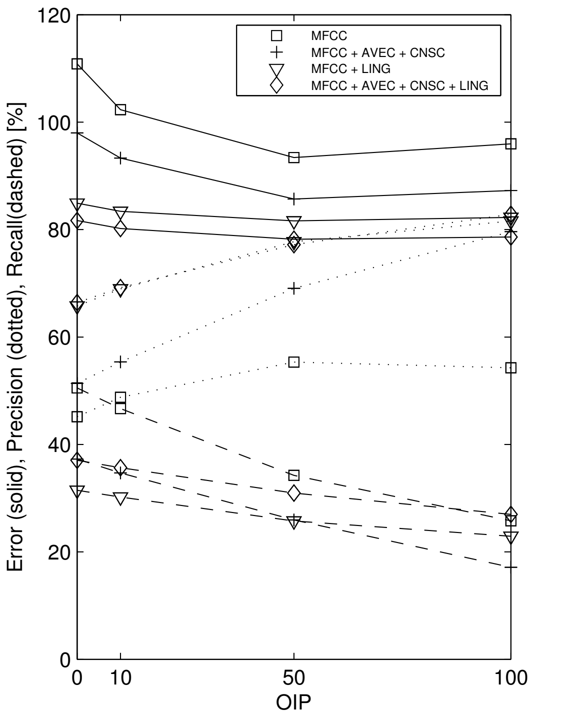

Figure 2: Overlap detection performance as a function of the OIP for five different feature combinations. Performance illustrated in terms of detection error (solid line), precision (dotted line) and recall (dashed line).

Features	OIP	Prec.	Rec.	Err.
MFCC	50	55.3	34.2	93.4
MFCC + AVEC + CNSC	50	77.8	25.8	81.6
MFCC + Ling.	65	72.6	23.2	85.5
MFCC+AVEC+CNSC+Ling.	85	81.7	28.0	78.3

Table 3: Precision (Prec.), recall (Rec.) and overlap detection error (Err.) on the test set for the four tested feature combinations. Operating points are tuned (by varying OIP) to achieve minimum overlap detection error on the tuning set.

posed linguistic features, using an oracle-style ASR system. Results are illustrated in Fig. 2 for each feature combination as function of OIP. In addition, Table 3 lists test set results for all four systems for one operating point. This operating point is determined by varing OIP and evaluating on the tuning set to achieve a minimum overlap detection error.

The lowest detection error achieved with MFCC features alone is 93.4 % achieved with an OIP of 50. When combined with the new linguistic feature, the error drops to 85.5 % as a result of significant improvements in precision. For other system operating points (other values for OIP), the addition of linguistic information to the MFCC feature set also helps to decrease the overlap detection error. The best feature set without linguistic features combines MFCCs, energy, spectral and voicing-related features from the AVEC set and CNSC features. With this feature set the minimum detection error is 81.6 %, again with an OIP of 50. When linguistic features are added to this feature set, the error drops to 78.3 %. This time, however, the drop in error is attributed to an increase in both, precision and recall.

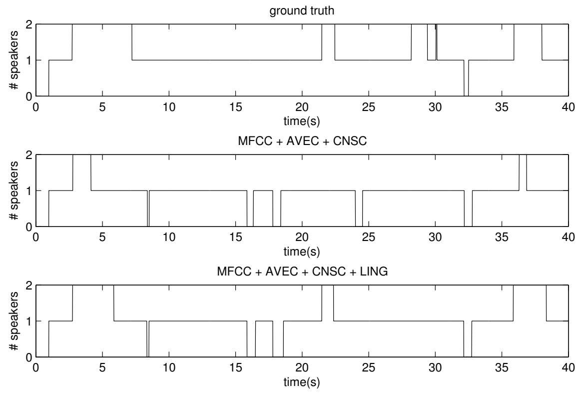

Figure 3: Illustration of overlap detection, showing annotations for the ground truth, the baseline system and the baseline combined with the linguistic feature (from top to bottom) for a 40-second excerpt of a recording from the test set.

Here again, the addition of the new linguistic feature leads to a constant performance gain for all tested values of OIP. The increased recall in the case of using linguistic features can be attributed to a better detection of small overlap segments, for example those containing backchannel utterances.

Figure 3 illustrates overlap detection performance for a 40 second long excerpt of a recording from the test set. The three plots show the ground truth annotation (0, 1 or 2 active speakers), the output of the baseline system and that when combined with the new linguistic feature. Together they show how the new feature has the potential not only to improve overlap detection accuracy of those segments already detected with the baseline approach, but also smaller segments which the baseline system otherwise fails to identify.

5. Conclusions

This paper presents our latest work in overlap detection. Meeting recordings are segmented into intervals of nonspeech, speech and overlapping speech using an HMM-based framework and a diverse set of features. The new contribution relates to the use of unigram language models to distinguish between speech and overlap. Especially short overlap segments (backchannel utterances or interruptions) are expected to exhibit different word distributions than single speaker intervals. Experiments confirm the hypothesis that linguistic features can help to improve overlap detection; results show improved precision and recall performance and reduced overlap detection error. Linguistic features were derived from an oracle-style ASR system. It has to be tested, how these results transfer to the case when a real more error-prone ASR system is used.

To extend the idea further, future work should consider the use of not only language model scores, but also those from acoustic models. In particular, further work should study in greater detail the nature of overlapping speech so that new insights can stimulate the develop of future overlap detection systems.

6. Acknowledgements

This research was supported by the ALIAS project (AAL-2009-2-049) co-funded by the EC, the French ANR and the German BMBF. Thanks to Ravichander Vipperla for his contributions to the CNSC features.

7. References

[1] E. Shriberg, A. Stolcke, and D. Baron, "Observations on Overlap: Findings and Implications for Automatic Processing of Multi-Party Conversation," in *Proc. Eurospeech*, Aalborg, Denmark, 2001, pp. 1359–1362.

[2] E. Shriberg, "Spontaneous Speech: How People Really Talk and Why Engineers Should Care," in *Proc. Eurospeech*, Lisbon, Portugal, 2005, pp. 1781–1784.

[3] M. Huijbregts and C. Wooters, "The Blame Game: Performance Analysis of Speaker Diarization System Components," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 1857–1860.

[4] F. Brugnara, D. Falavigna, D. Giuliani, and R. Gretter, "Analysis of the Characteristics of Talk-show TV Programs," in *Proc. Interspeech*, Portland, OR, USA, 2012.

[5] C. Fredouille and N. Evans, "The influence of speech activity detection and overlap on the speaker diarization for meeting room recordings," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 2953–2956.

[6] S. Otterson and M. Ostendorf, "Efficient use of overlap information in speaker diarization," in *Proc. ASRU*, Kyoto, Japan, 2007, pp. 683–686.

[7] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped Speech Detection for Improved Diarization in Multi-Party Meetings," in *Proc. ICASSP*, Las Vegas, NV, USA, 2008, pp. 4353–4356.

[8] K. Boakye, O. Vinyals, and G. Friedland, "Two's a crowd: improving speaker diarization by automatically identifyingand excluding overlapped speech," in *Proc. Interspeech*, Brisbane, Australia, 2008, vol. 1, pp. 32–35.

[9] K. Boakye, O. Vinyals, and G. Friedland, "Improved Overlapped Speech Handling for Speaker Diarization," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 941–944.

[10] M. Zelenák, C. Segura, and J. Hernando, "Overlap detection for speaker diarization by fusing spectral and spatial features," in *Proc. Interspeech*, Makuhari, Japan, 2010, pp. 2302–2305.

[11] M. Zelenak, C. Segura, J. Luque, and J. Hernando, "Simultaneous speech detection with spatial features for speaker diarization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 436–446, 2012.

[12] M. Zelenak and J. Hernando, "The Detection of Overlapping Speech with Prosodic Features for Speaker Diarization," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 1041–1044.

[13] R. Vipperla, J. Geiger, S. Bozonnet, D. Wang, N. Evans, B. Schuller, and G. Rigoll, "Speech Overlap Detection and Attribution Using Convolutive Non-Negative Sparse Coding," in *Proc. ICASSP*, Kyoto, Japan, 2012, pp. 4181–4184.

[14] J. Geiger, R. Vipperla, N. Evans, B. Schuller, and G. Rigoll, "Speech Overlap Detection and Attribution Using Convolutive Non-Negative Sparse Coding: New Improvements and Insights," in *Proc. EUSIPCO*, Bucharest, Romania, 2012, pp. 340–344.

[15] J. Geiger, R. Vipperla, S. Bozonnet, N. Evans, B. Schuller, and G. Rigoll, "Convolutive Non-Negative Sparse Coding and New Features for Speech Overlap Handling in Speaker Diarization," in *Proc. Interspeech*, Portland, OR, USA, 2012.

[16] S. H. Yella and F. Valente, "Speaker Diarization of Overlapping Speech based on Silence Distribution in Meeting Recordings," in *Proc. Interspeech*, Portland, OR, USA, 2012.

[17] S. H. Yella and H. Bourlard, "Improved Overlap Speech Diarization of Meeting Recordings using Long-Term Conversational Features," in *Proc. ICASSP*, Vancouver, Canada, 2013.

[18] H. Sun and B. Ma, "Study of Overlapped Speech Detection for NIST SRE Summed Channel Speaker Recognition," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 2345–2348.

[19] M. Wlodarczak, J. Simko, and P. Wagner, "Temporal Entrainment in Overlapped Speech: Cross-linguistic Study," in *Proc. Interspeech*, Portland, OR, USA, 2012.

[20] K. Laskowski, M. Heldner, and J. Edlund, "On the Dynamics of Overlap in Multi-Party Conversation," in *Proc. Interspeech*, Portland, OR, USA, 2012.

[21] A. Gravano, S. Benus, J. Hirschberg, S. Mitchell, and I. Vovsha, "Classification of discourse functions of affirmative words in spoken dialogue," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 1613–1616.

[22] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *Proc. ICASSP*, Montreal, Canada, 2004, pp. 574–577.

[23] S. Bozonnet, R. Vipperla, and N. Evans, "Phone Adaptive Training for Speaker Diarization," in *Proc. Interspeech*, Portland, OR, USA, 2012.

[24] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011–The First International Audio/Visual Emotion Challenge," in *1st Int. Audio/Visual Emotion Challenge and Workshop*, Memphis, TN, USA, 2011, pp. 415–424.

[25] P. Smaragdis, "Convolutive Speech Bases and Their Application to Supervised Speech Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[26] D. Wang, R. Vipperla, and N. Evans, "Online pattern learning for non-negative convolutive sparse coding," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 65–68.

[27] D. Wang, R. Vipperla, N. Evans, and T. F. Zheng, "Online non-negative convolutive pattern learning for speech signals," *to appear in IEEE Transactions on Signal Processing, Oct. 2012*, 2012.

[28] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, et al., "The AMI meeting corpus: A pre-announcement," *Machine Learning for Multimodal Interaction*, pp. 28–39, 2006.

[29] S. Bozonnet, N. Evans, and C. Fredouille, "The LIA-Eurecom RT09 Speaker Diarization System: Enhancements in Speaker Modelling and Cluster Purification," in *Proc. ICASSP*, Dallas, TX, USA, 2010, pp. 4958–4961.