OVERLAPPED SPEECH DETECTION USING LONG-TERM SPECTRO-TEMPORAL SIMILARITY IN STEREO RECORDING

Bo Xiao, Prasanta Kumar Ghosh, Panayiotis Georgiou, and Shrikanth S. Narayanan

Signal Analysis and Interpretation Laboratory, Department of Electrical Engineering,
University of Southern California, Los Angeles, CA 90089
boxiao@usc.edu, prasantg@usc.edu, georgiou@sipi.usc.edu, shri@sipi.usc.edu

ABSTRACT

The problem of detecting overlapped speech in stereo recordings using close-talk microphones is important for a variety of applications including the identification of back-channels, interruptions *etc.* in a dyadic or multi-party interactions. For detecting overlapped speech, we propose a feature derived using the spectral similarity of two channels over a range of acoustic frames. During overlapped speech frames the proposed spectro-temporal similarity-based feature values decrease and during non-overlapped speech frames the feature values increase due to the presence of cross-talk. Thus the proposed feature helps to discriminate the overlapped speech frames from the non-overlapped ones. Using overlapped speech detection experiments on a dyadic interaction corpus, it is shown that the proposed feature provides a significant improvement, ∼26% absolute, in the accuracy of detecting the overlapped speech frames when used as an additional feature to the baseline feature obtained from the two channels' intensity profiles.

Index Terms— overlapped speech, spectrogram, correlation coefficient, stereo recording

1. INTRODUCTION

Overlapped speech detection (OSD) aims to identify the segments of a speech signal that contain multiple simultaneously-active speakers. Since the acoustic characteristics of overlapped speech are different from that of a single talker, performance of most automated processing components such as voice activity detection, automatic speech recognition, and speaker diarization in the overlapped speech regions is significantly degraded. Also, OSD serves an important behavioral cue in the analysis of both interpersonal interactions (e.g., meetings, doctor-patient interactions *etc.*) and in mediated interactions (e.g., interpreted speech-to-speech translation). It can also be a useful feature for characterizing back-channels and interruption patterns. Thus, the ability to detect overlapped speech is highly desirable and can serve as an important element for a range of analyses of interactions.

Robust speech features are critical so that overlapped speech can be efficiently discriminated from non-overlapped speech and silence. Since overlapped speech typically occurs less frequently than speech from a single talker in most spoken interactive settings, the feature must be designed not only to increase the overlapped speech detection accuracy (i.e., recall) but also to decrease the false alarm rate (i.e., improve precision).

Features available for OSD are fundamentally dependent on the type and number of sensors, i.e., microphones. For instance, in the scenario where only a single-channel speech signal is available, i.e., all speakers are recorded from a single microphone, an early work [1] proposed the "Spectral Autocorrelation Peak Valley Ratio" (SAPVR) criterion, with the intuition that overlapped speech would have a lower value of such ratio. In [2] about 40 features were tested and those most effective in detecting overlapped speech were reported, including notably *Mel-Frequency Cepstral Coefficients* (MFCCs), root-mean-squared energy, and linear predictive coding residual energy. In addition, the frame level entropy measure has been suggested to be effective because overlapped speech in general has higher entropy than others [3]. In the case of audio acquisition through a microphone array the direction of arrival of the sound, and hence the active speaker direction, can be estimated [4]. If multiple prominent directions are present, it is likely that multiple speakers are simultaneously active.

In our work, we consider the case where close-talk microphones are used for each speaker. The use of multiple microphones while may improve OSD performance compared to single microphone recording, may cause a critical side effect, namely cross-talk. Cross-talk happens when one speaker's sound is also picked by the other speaker's microphone. Processing each channel independently would fail since the inactive speaker may be detected as active. Hence, investigation of the relationship among different channels is essential. In [5] it was first suggested that the lag between two channels' signals can be estimated though the location of the peak in the cross-correlation (XC) spectrum. A thresholding on the smoothed maxima of the XC in time domain was used to reject "false overlap". Based on the XC, in [6] two algorithms were proposed called "Inter-microphone Time Differences" and "Joint Maximum Crosscorrelation". Yet, in the recording setup with low signal-to-noise ratio (SNR), the peak detection in the correlation function might be prone to more error. A more extensive work was done in [7], where a selected group of features were used for better modeling the overlapped speech, including MFCC, energy, kurtosis, and fundamentalness. The activity of each speaker was modeled independently into 4 states, namely not speaking, not speaking but picking cross-talk, speaking, and both speaking and picking cross-talk. The result showed that features derived from XC are the most effective for modeling single attribute states, but energy provided a consistently good result for all states. Inspired by this study, energy based features are adopted as baseline method in our work.

In this paper, we focus on the OSD from stereo recordings using close-talk microphones of two speakers engaged in a conversation. In an earlier work [8], a similar scenario was adopted, and it was shown that the long term information of the audio was effective for channel selection in the presence of cross-talk. However, overlapped speech was assumed to be absent. To better detect overlapped speech, we introduce a new feature — spectro-temporal similarity of two channels' acoustics, which is effective in separating overlapped speech regions from the non-overlapped ones. In general, during cross-talk, the similarity between the time-varying spectra of the original signal and the cross-talk signal would be high as the later contains mainly the leakage from the active speaker. For overlapped case the similarity of the time-varying spectra in two channels would be low. Through overlapped/non-overlapped speech classification experiments, we demonstrate the usefulness of the proposed feature for OSD. We obtain significant improvements in both recall

[†] This work was supported in part by funds from DARPA and NSF.

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

and precision of detecting overlapped speech frames when the proposed feature is used in addition to the baseline feature. For classification, we use the Gaussian Mixture Model (GMM) with a *Maximum Likelihood* (ML) decision.

We begin with the description of the dyadic interaction dataset in Section 2. In Section 3, we describe the proposed long-term spectro-temporal similarity feature between two channels and its characteristics during overlapped and non-overlapped speech. We report and discuss the experimental results in Section 4 with conclusions following in Section 5.

2. DATA SET

For the experiments in this paper, we have used the multi-modal dyadic (two person) interaction database recently collected at USC [9]. Among the multiple available microphones in the corpus, we have only used the two lapel microphones, one for each speaker. During the recording, the two interlocuters were seated side by side on a couch in a normal office environment. The recording was done at 48kHz sampling frequency and 24 bits PCM quantization; we down-sampled it to 8kHz for computing the proposed feature. The dataset used for our experiments contains 15 sessions of total duration 83 minutes. The speaker activity information was manually labeled using the Transcriber software and one of the four labels are given to each frame (every 10 msec): 1) Silence (S_N), 2) Right channel speaker active (S_R), 3) Left channel speaker active (S_L), and 4) Overlapped speech (S_{RL}). The data set is divided into a training set and a test set; the details are given in Table 1. The training sessions are those having high proportion of overlapped speech such that models built on training data are generalizable.

Part	Ses.	Length	S_N	S_R	S_L	S_{RL}
Train/Dev	5	34 min	21%	43%	29%	7%
Test	10	49 min	25%	30%	42%	3%

Table 1. Details of the training and test split of the dataset. We used the training set also as a dev set for any parameter tuning.

3. LONG-TERM SPECTRO-TEMPORAL SIMILARITY

We exploit the similarity between the signal characteristics in two channels to detect the overlapped speech regions. The basic idea is simple: when only one subject (say, corresponding to the right channel) talks then the signals in both channels are similar due to cross-talk; however, when both subjects speak at the same time (i.e., during overlapped speech), the signals in the two channels are dissimilar. Thus, a measure of similarity between signals in the two channels is expected to indicate whether there is overlap. We use the correlation between the two channels' signal spectra as a measure of similarity. Below we provide an explanation of how such measure is different for the non-overlapped speech regions compared to the overlapped speech regions.

Without loss of generality, let us assume that only the subject corresponding to the right channel is speaking. Then the signals of the right and left channels, i.e., $x_R[n]$ and $x_L[n]$ respectively, can be written as

$$\begin{aligned} x_R[n] &= s[n] + w_R[n] \\ x_L[n] &= \alpha s[n - n_0] + w_L[n], \end{aligned} \tag{1}$$

where $s[n]$ is the speech signal spoken by the right channel's subject and α is the attenuation factor ($0 < \alpha < 1$) from the right channel to the left channel. n_0 is the delay between the left and right channel signals; n_0 is usually 16–24 samples (for the 8kHz sampling frequency) for our dataset. $w_R[n]$ and $w_L[n]$ are the additive noises in the right and the left channels, assumed to be zero-mean white Gaussian distributed with variance σ^2. We also assume that the noises are independent of the speech signal. Thus, the spectra of the signals from the two channels are as follows:

$$\begin{aligned} S_{x_R}(\omega) &= S_s(\omega) + \sigma^2 \\ S_{x_L}(\omega) &= \alpha^2 S_s(\omega) + \sigma^2, \end{aligned} \tag{2}$$

where $S_z(\omega) = |Z(\omega)|^2$ is the spectrum of signal $z[n]$ and $Z(\omega)$ represents the Fourier transform of signal $z[n]$. Note that we can write $S_{x_L}(\omega)$ as follows:

$$S_{x_L}(\omega) = \alpha^2 S_{x_R}(\omega) + \sigma^2(1 - \alpha^2) \tag{3}$$

This means that the left channel spectrum is an affine function of the right channel spectrum. Thus, one channel's spectrum is linearly correlated to the other channel's spectrum and, hence, the correlation among them will be high. This correlation (similarity) is illustrated in Fig. 1(a) and (b) [left column] using a spectrogram over a duration of 1 second for two channels, when the subject corresponding to the first channel is speaking.

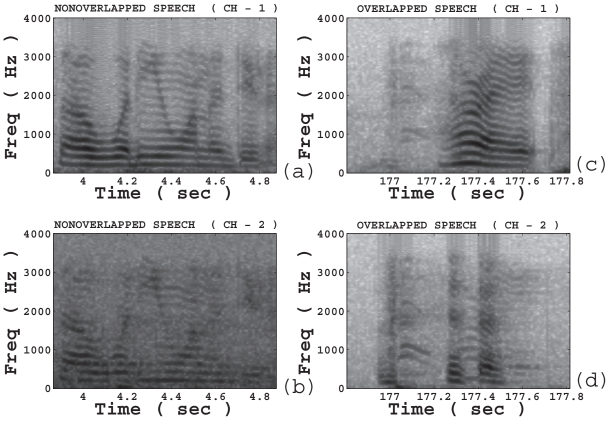


Fig. 1. Illustrative examples of the spectrograms of two channels during non-overlapped (left) and overlapped (right) speech.

On the other hand, when both the subjects are speaking, the signals of the right and left channels are

$$\begin{aligned} x_R[n] &= s_R[n] + \alpha_R s_L[n - n_0] + w_R[n] \\ x_L[n] &= s_L[n] + \alpha_L s_R[n - n_0] + w_L[n], \end{aligned} \tag{4}$$

where $s_R[n]$ and $s_L[n]$ are the speech signal from the right and left channel subjects respectively. We assume that $s_R[n]$ and $s_L[n]$ are independent of each other because they are produced by two different subjects. α_L is the attenuation ($0 < \alpha_L < 1$) of $s_R[n]$ when it is received by the left channel and, similarly, α_R is the attenuation ($0 < \alpha_R < 1$) of $s_L[n]$ when it is received by the right channel. Thus, the spectra of the two channels' signals can be written in the following vector-matrix form:

$$\begin{bmatrix} S_{x_R}(\omega) \\ S_{x_L}(\omega) \end{bmatrix} = \begin{bmatrix} 1 & \alpha_R^2 \\ \alpha_L^2 & 1 \end{bmatrix} \begin{bmatrix} S_{S_R}(\omega) \\ S_{S_L}(\omega) \end{bmatrix} + \begin{bmatrix} \sigma^2 \\ \sigma^2 \end{bmatrix} \tag{5}$$

We can assume that the speech spectra from the right and the left channel subjects, i.e., $S_{s_R}(\omega)$ and $S_{s_L}(\omega)$ are uncorrelated to one another. Then, for the no noise condition (i.e., $\sigma^2 = 0$), it is easy to show that the correlation between $S_{x_R}(\omega)$ and $S_{x_L}(\omega)$ will be low provided that $\alpha_R \ll 1$ and $\alpha_L \ll 1$. This is illustrated in Fig. 1(c) and (d) where the spectrograms in Fig. 1(c) and 1(d) appear dissimilar and, hence, a low correlation between them is expected.

Based on the above explanation, we define a similarity measure using correlation coefficient between the time-varying spectra of two channels. Let the target frame for the OSD be denoted by index m.

Let $\mathbf{S}_{x_R}[n,\omega_k],\ \mathbf{S}_{x_L}[n,\omega_k],\ m-P \le n \le m+P,\ 1 \le k \le K$ be the magnitudes of the time-varying spectra of the right and the left channels respectively at $2P+1$ frames around the target frame computed at K frequency points between 0Hz and $\frac{1}{2}F_s$Hz, where F_s is the sampling frequency. The analysis frame length and frame shift is N_w and N_{sh} samples. We use the Pearson's correlation coefficient between $\mathbf{S}_{x_R}[n,\omega_k]$ and $\mathbf{S}_{x_L}[n,\omega_k]$ to define the similarity $\rho_{RL}(m)$ at frame m to be

$$\rho_{RL}(m) = \frac{V_{RL} - M_R M_L}{\sqrt{V_R - M_R^2}\sqrt{V_L - M_L^2}}, \qquad (6)$$

where $M_R = \frac{1}{T}\sum_{n,k} \mathbf{S}_{x_R}[n,\omega_k],\ M_L = \frac{1}{T}\sum_{n,k} \mathbf{S}_{x_L}[n,\omega_k],$

$$V_R = \frac{1}{T}\sum_{n,k} \left(\mathbf{S}_{x_R}[n,\omega_k]\right)^2,\ V_L = \frac{1}{T}\sum_{n,k} \left(\mathbf{S}_{x_L}[n,\omega_k]\right)^2,$$

$$V_{RL} = \frac{1}{T}\sum_{n,k} \mathbf{S}_{x_R}[n,\omega_k]\mathbf{S}_{x_L}[n,\omega_k],\ \text{and}\ T = K(2P+1)$$

$\rho_{RL}(m)$ is the Pearson's correlation coefficient between two $K(2P+1)$ dimensional vectors corresponding to the right and left channels whose elements correspond to the elements of $\mathbf{S}_{x_R}[n,\omega_k]$ and $\mathbf{S}_{x_L}[n,\omega_k]$, respectively. Note that we assume the effect of n_0 on the fixed frame based analysis is negligible. For our analysis this is a reasonable assumption given that n_0 is significantly less than the frame shift. A histogram of $\rho_{RL}(m)$ for overlapped and non-overlapped frames obtained from five sessions (a total of 34 minutes stereo recording) is shown in Fig. 2. To compute $\rho_{RL}(m)$, we use the following parameter values: F_s=8kHz, N_w=160, N_{sh}=80, P=25, K=252, and the K frequency points $\omega_1,\ \ldots,\ \omega_K$ are uniformly spaced from 50Hz to 4000Hz. As we explained above, $\rho_{RL}(m)$ is low for overlapped speech and high for non-overlapped speech. This histogram indicates that $\rho_{RL}(m)$ provides useful information in disambiguating the two classes and can be used as a feature for such purpose. In the next section, we perform experiments to detect overlapped speech frames using $\rho_{RL}(m)$.

4. EXPERIMENTAL RESULTS

In our experiments, the goal is to classify each acoustic speech frame as either overlapped speech or non-overlapped speech. We use short-time energy of each channel (denoted by E_R and E_L) as the feature for the baseline experiment (following the results of [7]), estimated from the intensity values of the original two channels' signals provided by the Praat software [10]. During overlapped speech (S_{RL}), the short-time intensities of both channels are expected to be high, but in a non-overlapped speech frame there can be three types of short-time energy distributions — either the energy of one channel is higher than the other (when only one subject is speaking [S_R or S_L]), or the energy of both channels are lower (during silence [S_N]). We use Gaussian mixture model (GMM) to model the feature space of overlapped and non-overlapped speech frames. We found that the classification accuracy on the training data improves when four GMMs are trained separately on the features from S_{RL}, S_R, S_L, and S_N compared to a 2-way GMM setting with classes (S_{RL}) and (S_R, S_L, S_N). Thus, our 4-way classification problem will result in overlap detection with the detection of S_{RL} and non-overlap in any of the other three cases. We investigate both *Maximum Likelihood* (ML – equal prior probability for each class) and *Maximum A-Posteriori* (MAP – class dependent prior probabilities) classification using GMM. When a test frame is classified as either S_R or S_L, or S_N, it is declared as non-overlapped speech frames otherwise it is declared as an overlapped speech frame. In the case of E_R and E_L being features, we optimize the number of components in GMM by maximizing the two-class classification accuracy on the training set (15297 overlapped frames and 191840 non-overlapped frames). In our experiment, 4-component and 5-component GMM turned out to be the best for the ML and MAP classification schemes respectively. We report the overlapped speech hit-rate (OHR) [defined as the ratio of the number of correctly classified frames among all overlapped speech frames], non-overlapped speech hit-rate (NOHR), and the overall accuracy for the optimized GMM on the training set in the first row of Table 2. Since the non-overlapped frames greatly outnumber the overlapped frames, OHR for MAP classification is worse compared to ML classification although the best classification accuracies using MAP is higher than that using ML.

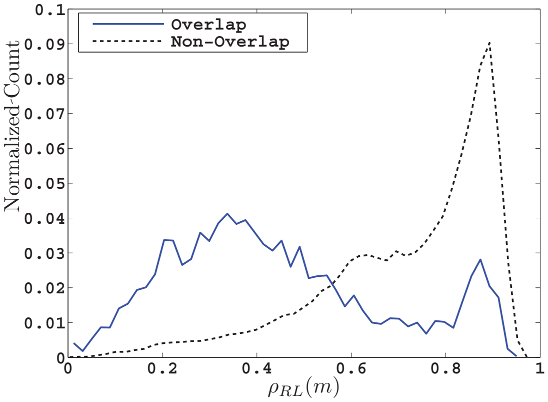


Fig. 2. Illustrative histogram of $\rho_{RL}(m)$ for the non-overlapped and overlapped speech.

The optimized GMM based (using E_R and E_L as features) ML and MAP classifier result on the test set is also shown in the first row of Table 2. The relative performances of ML and MAP classifiers on the test set are consistent with those on the training set. Although the overall accuracies of GMM classifiers using E_R and E_L as features are high, the respective OHRs are poor. To improve the OHR, we examine the utility of the proposed long-term spectro-temporal similarity feature ρ_{RL} for the task of OSD.

We optimize the related parameters P, K for computing ρ_{RL} on the training set using *Receiver Operating Characteristics* (ROC) curve to achieve highest accuracy on the two-class classification problem. The parameter combination corresponding to the *Equal Error Rate* (EER) is finally selected. ρ_{RL} computed from the narrowband spectrogram yielded better accuracy compared with that from the wideband spectrogram. P is chosen from the set of {1, 5, 10, 15, 20, 25}. The time-varying spectrum is computed at 256 uniformly spaced frequency points between 0Hz and 4000Hz and K frequency points are selected by choosing 12 different frequency ranges with lower frequency values {50, 500, 1000} and higher frequency values {2500, 3000, 3500, 4000}. An EER corresponding to a threshold of 0.599 was obtained for P=25, and frequency range 50-4000Hz with K=252. The corresponding OHR, NOHR, and overall accuracies are shown in second row of Table 2. It is important to note that the OHR and NOHR using ρ_{RL} on the test set is complementary to the OHR and NOHR using energy-based GMM classifier, i.e., energy-based GMM classifier yields more NOHR than OHR, while it is opposite for ρ_{RL}. This observation motivates us to use E_R, E_L, and ρ_{RL} altogether as a feature vector for OSD.

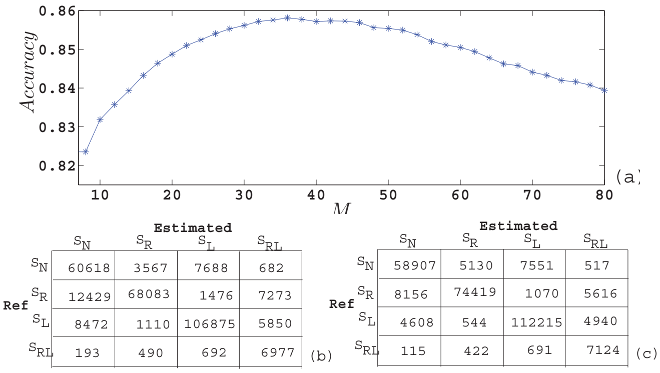
We use E_R, E_L, and ρ_{RL} as 3-dimensional feature vector and both ML and MAP classification using GMM to model the distribution of S_N, S_R, S_L, and S_{RL} similar to what was done for GMM using E_R and E_L as features. 5-component GMMs yielded the highest accuracy of 92.87% and 95.19% on the training set for the ML and MAP classifications, respectively. The corresponding OHR and NOHR are shown in the third row of Table 2. It is important to observe that after adding ρ_{RL} as features, the OHR (75.54%) for ML classification has improved compared to the the OHR (54.02%) obtained using E_R and E_L features only. The OHR and NOHR on the

Features	Classifier		Train Set			Test Set		
			OHR	NOHR	Accuracy	OHR	NOHR	Accuracy
E_R E_L	GMM	ML MAP	54.02% 33.61%	95.38% 99.25%	92.32% 94.41%	**57.38**% 21.13%	**94.80**% 99.74%	93.73% 97.49%
ρ_{RL}	ROC		76.75%	76.75%	76.75%	89.73%	79.41%	80.17%
E_R E_L ρ_{RL}	GMM	ML MAP	75.54% 56.99%	94.25% 98.24%	92.87% 95.19%	**83.54**% 42.35%	**95.13**% 98.97%	94.80% 97.35%

Table 2. OHR and NOHR on the training and the test set set for various feature and different optimized classifier combinations.

test set using the 3-dimensional feature based optimized GMM are also shown in the third row of Table 2. It is interesting to note that both OHR and NOHR are better on the test set compared to the training set; this indicates that the GMMs are generalized enough to work well on the unseen data. It is also important to note that the OHR on the test set improves to 83.54% from 57.38% (significant at p=0.002) for using ρ_{RL} as feature in addition to E_R and E_L for the ML classification; a similar improvement is observed for MAP classification too. This improvement in OHR demonstrates the efficacy of ρ_{RL} for OSD.

The MAP classifier performs worse compared to ML in terms of OHR due to imbalance between the number of realizations of overlapped and non-overlapped speech frames; this is consistent for both two and three dimensional features. Overall, the ML classifier using 3-dimensional feature based GMM achieves the best performance on the test set. To improve the performance further, we perform a median filtering with length M, where M is optimized on the training set. Note that this median filtering was done on a sequence of four-class labels (S_N, S_R, S_L, and S_{RL}); median filtering on two class label sequence (i.e., overlapped and non-overlapped frames) was poor in terms of OHR. Fig. 3 shows the four-class classification accuracy on the training set for varying median filter length M. M=37 was picked since it resulted in the highest accuracy. We performed a median filtering of length 37 on the four-class labels of the test set and the overlap/non-overlap accuracy improved from 94.80% to 95.79% (OHR improved from 83.54% to 85.30% and NOHR improved from 95.13% to 96.10%). The corresponding four-class confusion matrices on test set before and after median filtering is shown in Fig. 3 (b) and (c) respectively. Since many of the detected overlapped speech frames are false alarms, we compute F-score of OSD to incorporate both precision and recall to one metric. We found that the F-score $\left(2\frac{Precision \times Recall}{Precision + Recall}\right)$ on the test set using 2-dimensional energy feature based ML GMM classifier was 0.34; when ρ_{RL} is used as additional feature the F-score improved to 0.48 (significant at p=0.002) and after median filtering it further improved to 0.54.



		Estimated			
		S_N	S_R	S_L	S_{RL}
Ref	S_N	60618	3567	7688	682
	S_R	12429	68083	1476	7273
	S_L	8472	1110	106875	5850
	S_{RL}	193	490	692	6977

		Estimated			
		S_N	S_R	S_L	S_{RL}
Ref	S_N	58907	5130	7551	517
	S_R	8156	74419	1070	5616
	S_L	4608	544	112215	4940
	S_{RL}	115	422	691	7124

Fig. 3. (a) Accuracy on the training set with median filtering with different length M, (b) and (c) The four-class confusion matrix on the test set before and after median filtering respectively.

5. CONCLUSIONS

We have shown that the use of the correlation coefficient between the two channels' spectra as a feature significantly improves the overlapped speech detection (OSD) performance compared to that obtained by two channel intensity profile in a channel-per-speaker recording setup. It was also found that the correlation of spectra over a larger number of consecutive frames (P=25) are more effective for OSD. In spite of these improvements, we observed that there are false alarms for overlap and further investigation indicates that they occur spuriously or in an isolated manner, i.e., not in a contiguous number of frames. Thus, in applications such as interruption or back-channel detection, where OSD can be used as a pre-processing step, a temporal contiguity measure or a probabilistic model of state transitions may help to further pinpoint the actual locations of the overlapped speech regions.

6. REFERENCES

[1] R. E. Yantorno, K. R. Krishnamachari, J. M. Lovekin, D. S. Benincasa, and S. J. Wenndt, "The Spectral Autocorrelation Peak Valley Ratio (SAPVR)-a Usable Speech Measure Employed as a Co-channel Detection System," in *Proceedings of IEEE International Workshop on Intelligent Signal Processing (WISP)*, 2001.

[2] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4353–4356.

[3] O. Ben-Harush, H. Guterman, and I. Lapidot, "Frame level entropy based overlapped speech detection as a pre-processing stage for speaker diarization," in *IEEE International Workshop on Machine Learning for Signal Processing*. IEEE, 2009, pp. 1–6.

[4] V. Rozgic, K. J. Han, P. G. Georgiou, and S. S. Narayanan, "Multimodal speaker segmentation and identification in presence of overlapped speech segments," *Journal of Multimedia*, vol. 5, no. 4, pp. 322–331, 2010.

[5] T. Pfau, D. P. W. Ellis, and A. Stolcke, "Multispeaker speech activity detection for the ICSI meeting recorder," in *IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2001, pp. 107–110.

[6] K. Laskowski, Q. Jin, and T. Schultz, "Crosscorrelation-based multispeaker speech activity detection," in *Eighth International Conference on Spoken Language Processing*, 2004.

[7] S. N. Wrigley, G. J. 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, 2004.

[8] P. K. Ghosh, A. Tsiartas, P. G. Georgiou, and S. S. Narayanan, "Robust voice activity detection in stereo recording with crosstalk," in *In Proceedings of InterSpeech 2010*, Makuhari, Japan, Sep 2010.

[9] V. Rozgic, B. Xiao, A. Katsamanis, B. Baucom, P. G. Georgiou, and S. S. Narayanan, "A new multichannel multimodal dyadic interaction database," in *In Proceedings of InterSpeech*, Makuhari, Japan, Sep 2010.

[10] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 4.3.01) [computer program]," in *Retrieved from http://www.praat.org/*, 2005.