OVERLAPPED-SPEECH DETECTION WITH APPLICATIONS TO DRIVER ASSESSMENT FOR IN-VEHICLE ACTIVE SAFETY SYSTEMS

Navid Shokouhi, Amardeep Sathyanarayana, Seyed Omid Sadjadi, John H.L. Hansen*

Center for Robust Speech Systems (CRSS)
The University of Texas at Dallas, Richardson, TX 75080-3021, USA
{navid.shokouhi,amardeep,sadjadi,john.hansen}@utdallas.edu

ABSTRACT

In this study we propose a system for overlapped-speech detection. Spectral harmonicity and envelope features are extracted to represent overlapped and single-speaker speech using Gaussian mixture models (GMM). The system is shown to effectively discriminate the single and overlapped speech classes. We further increase the discrimination by proposing a phoneme selection scheme to generate more reliable artificial overlapped data for model training. Evaluations on artificially generated co-channel data show that the novelty in feature selection and phoneme omission results in a relative improvement of 10% in the detection accuracy compared to baseline. As an example application, we evaluate the effectiveness of overlapped-speech detection for vehicular environments and its potential in assessing driver alertness. Results indicate a good correlation between driver performance and the amount and location of overlapped-speech segments.

***Index Terms*—** Active safety, co-channel speech, overlapped speech detection

1. INTRODUCTION

Co-channel speech is referred to a monophonic audio recording in which at least two speakers are present. The presence of co-channel speech in audio recordings poses a great challenge for many speech applications such as automatic speech recognition (ASR) and speaker identification (SID). There have been two major approaches in the literature towards alleviating the co-channel speech problem. One approach is to separate the target or interfering speech signals by enhancing one or suppressing the other [1, 2, 3]. The second approach is to detect the presence of more than one speaker at every time instance, which is mainly referred to as overlapped-speech detection [4, 5, 6]. In many applications, including the problem investigated in this study, the latter suffices in mitigating the degradations caused by co-channel speech (for a review see [7]). An example is reducing errors in speaker diarization by omitting overlapping speech regions [8]. In [9], Yantorno investigated the impact of overlapping speech segments in SID. Another advantage of detecting overlapping speech segments in co-channel scenarios is that it enables one to extract the contextual information of a conversation by determining the locations and amount of overlapped-speech in that conversation [10]. The main focus of the present study is to determine the location of overlapped-speech and showcase the correlation of these locations with speakers' attention.

As an example application, we evaluate the effectiveness of overlapped-speech detection for in-vehicular environments and its potential in assessing driver alertness. Human error contributes to over 95% of the accidents on the road, causing not only an economic impact but also human loss and suffering [11]. There are several factors contributing to human error in driving scenarios, one of which is driver distraction. Many regulations in the form of new laws and guidelines are enforced on drivers as well as auto-manufacturers to minimize distractions within the vehicular environment, particularly targeting visual distractions [12]. As the use of audio/speech based systems is dramatically increasing for in-vehicular technologies, it is important to understand how they impact driver performance. Previously in [13] the impact of in-vehicle audio/speech activity on driving performance was investigated. This study further analyzes the driver involvement in in-vehicular speech activity and its influence on driving performance. The objective of our study is to employ overlapped-speech detection as a means to identify highly competitive conversations in which the driver is involved. By sensing any variations in driving performance, the co-channel analysis along with driving performance evaluation could help recognize speech-related distractions, which are hypothetically highly correlated with overlapping speech [10]. Driver performance is evaluated using sensor information extracted from a *smart* portable device (e.g., Tablet), and developing statistical models for the purpose of maneuver recognition and analysis. The evaluation is accomplished by computing deviations of current maneuvers from the general trend (see [14] for more details).

The paper is organized as follows: we first describe our method of detecting overlapped speech segments in a given co-channel speech signal; Section 2 provides information on audio features used in our framework. The back-end or the likelihood ratio based detection system is explained in Section 3, followed by experimental setup using the TIMIT corpora in Section 4. In Section 5, we continue by applying the proposed system to a real-world in-vehicle scenario thereby demonstrating the correlation between overlapped-speech and instances of driver distraction. At the end, conclusions are drawn in Section 6.

2. FEATURES

There have been a number of studies discussing the pros and cons of different features in the context of overlapped speech detection. In [4], Boakye evaluated the commonly used features in overlapped speech detection, and identified the best performing features for the task of speaker diarization. In this study, a subset of features used in [4] and [5] are adopted for our system development. An important characteristic of overlapped-speech compared to single-speaker speech is the presence of two relatively different fundamental frequencies [15]. The interference of the overlapping speaker results

*This project was funded by AFRL under contract FA8750-12-1-0188 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

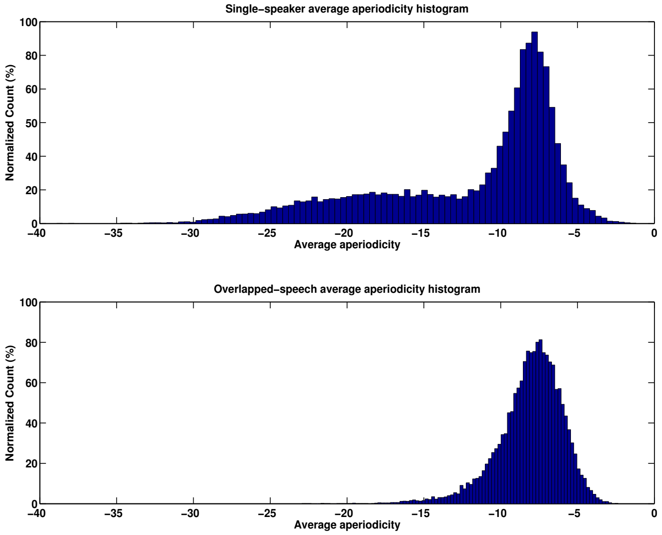


Fig. 1. Comparison of overlapped-speech average aperiodicity(top) and single-speeker speech average aperiodicity (bottom). The single-speaker features have a wider distribution at lower aperiodicities which belongs to voiced speech. The presence of the interfering speech in voiced regions increases the aperiodicity.

in less harmonic structure in the spectrum of overlapped-speech segments. There have been attempts to track the fundamental frequency of multi-speaker speech signals [16], but the accuracy of these methods drops significantly in noisy conditions, even when dealing with a single speaker. Other methods of capturing the amount of harmonicity of speech segments exploited in overlapped-speech detection are the Spectral Autocorrelation Peak-Valley Ratio [17] and the fundamentalness of speech segments[5]. In this study, we incorporate multiple features to effectively capture complementary harmonic and envelope information from speech spectra. The features chosen are 1) the aperiodicity measure introduced in [18], 2) the kurtosis of speech, 3) the spectral flatness measure (SFM), and 4) mel-frequency cepstral coefficients (MFCC) and its first order time derivative. The logic based on which these features are selected is discussed in more detail in the remaining of this section.

2.1. Average Spectral Aperiodicity

Aperiodicity is a measure for the amount of non-harmonic structure in the speech spectrum of a given frame. It is the normalized energy of non-harmonic frequency bins to the total energy of the spectrum [18]. The decrease in the harmonic structure due to the presence of a competing talker is the motivation to use such a measure in detecting overlapped-speech. Fig. 1 compares the distributions of the average aperiodicity measure for each frame of overlapped and single-speaker speech. The difference in the distributions is in the low-aperiodicity portion. A lower aperiodicity typically indicates "voiced" phonemes such as vowels that have a harmonic spectrum. This figure shows that the aperiodicity measure can be adopted as an effective feature for overlapped-speech detection. The detection is more accurate for vowels which is desirable for many speech applications (e.g., SID).

2.2. Kurtosis

Kurtosis has been reported as an effective measure for detecting the presence of multiple-speakers in co-channel signals in numerous studies [5, 4, 19]. It is well-known that overlapped-speech signals exhibit lower kurtosis compared to single-speaker speech signals [20].

2.3. Spectral Flatness Measure

The SFM is a feature that represents the amount of harmonicity of the short term speech spectrum in a single value. It is the ratio of the geometric to arithmetic mean of the spectral magnitudes in all frequency bins. The closer these two means are, the less harmonic the spectrum. The presence of two interfering fundamental frequencies in overlapped-speech would result in a lower SFM, i.e., the harmonicity is reduced.

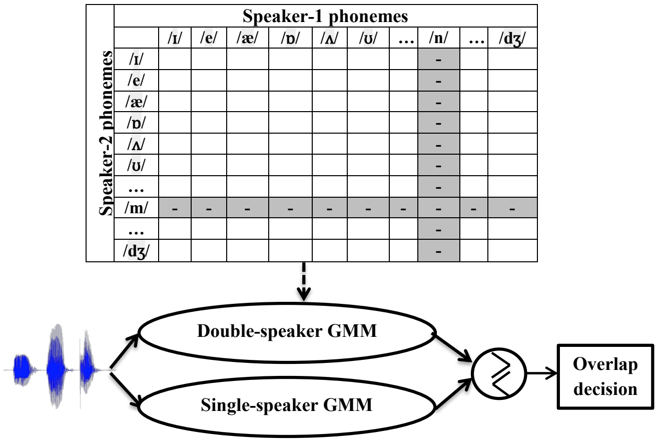
2.4. Mel-Frequency Cepstral Coefficients

MFCCs are chosen to capture differences in the spectral envelope of single-speaker to overlapped-speech. Our experiments show that using the first order cepstral derivative (ΔMFCC) provides valuable information for overlapped-speech detection. This is not surprising since one would expect the variation of spectral envelopes to be different in the presence of competing speakers. Here, 12-dimensional MFCC features (excluding C_0) are extracted and appended with ΔMFCC to form 24-dimensional vectors.

3. SYSTEM DESCRIPTION

The detection system, in which the above noted features are incorporated, is based on Gaussian mixture modeling (GMM) of single versus overlapped speech. To the best of our knowledge, GMM based feature space modeling for overlapped-speech detection has not been effectively utilized. Here, we propose a technique to create proper training input to represent overlapped-speech. Our pilot experiments show that all combinations of different phonemes are not equally informative in training the double-speaker model (see Section 4). More importantly, using a certain set of phonemes in generating artificial overlapped speech signals may also have a negative impact on the system performance. In order to examine this hypothesis, a table of all English phonemes is constructed and in different attempts different subsets of phonemes are selected to create overlapped-speech segments. Fig. 2 shows how the phoneme-pairs are chosen in our framework. In this figure, as an example, the nasal phonemes /m/ and /n/ are removed (shaded in the table). This implies that these phonemes are not used in generating the overlapped-speech data. Each small segment of the overlapped-speech data is created by the summation of the phonemes corresponding to each element in a table similar to Fig. 2. The phonemes are normalized and temporally aligned by truncating the longer phoneme. Each of the phonemes from the pair belongs to one of two speakers. This procedure is repeated for different pairs of speakers (from a set of M speakers) in order to capture the speaker variability. Some phonemes, such as stop consonants, are not used in creating the overlapped-speech data, because summing these (i.e., stop consonants) with high energy and harmonically structured phonemes (such as vowels) will result in signals that more or less belong to the single-speaker class. A major challenge in overlapped-speech detection is that single-speaker speech can be easily mistaken as overlapped-speech, and the proposed phoneme-omission scheme is expected to solve this problem to a great extent, thereby resulting in better performance.

The detection system uses two GMMs, one for single-speaker speech and one for overlapped-speech. The models are trained using various speech files from different speakers. Individual phonemes are summed with 0dB average Signal to Interference Ratio (SIR). The types of phonemes which are used to create overlapped-speech segments are experimentally selected, although some of the omitted phoneme-pairs can be inferred from basic understanding of phoneme



Speaker-2 phonemes	Speaker-1 phonemes									
	/ɪ/	/e/	/æ/	/ɒ/	/ʌ/	/ʊ/	…	/n/	…	/dʒ/
/ɪ/								-		
/e/								-		
/æ/								-		
/ɒ/								-		
/ʌ/								-		
/ʊ/								-		
…								-		
/m/	-	-	-	-	-	-	-	-	-	-
…								-		
/dʒ/								-		

Fig. 2. Overlapped speech detection system description. Shaded areas in the table indicate which phonemes are omitted in generating phoneme-pairs.

characteristics. Table 1 shows the improvements obtained by omitting different phoneme sets from the overlapped-speech training set.

4. EXPERIMENTS

The single-speaker and double-speaker models are trained using phonemes from TIMIT data. The individual phonemes are extracted using TIMIT data transcriptions. All 10 TIMIT utterances are used for 10 speakers to generate the single-speaker model. The double-speaker model is trained using the procedure explained in Section 3. A total of 10 different speakers are selected to build the double-speaker training data, of which half are used as speaker 1 and the other half as speaker 2 (refer to Fig. 2 for the definition of speaker 1 and 2). The features are extracted from 25 ms frames with a frame-shift of 10 ms. Two sets of test data are used: (i) Artificially generated co-channel speech signals by adding TIMIT utterances. This type of co-channel data might be unrealistic to represent conversational co-channel speech; however such co-channel speech signals do exist and are the result of cross-talk between two independent channels. (ii) A set of realistic conversational co-channel speech signals collected from the UTDrive vehicle (see Section 5). In all cases experiments are carried out only for male speakers due to the fact that the UTDrive data consists solely of male speakers.

The measures used to evaluate the overlap detection performance in the TIMIT test set are as follows:

Recall Rate (R_r): Number of Correctly Detected Overlaps to Total Number of Overlaps Present

Precision Rate (R_p): Number of Correctly Detected Overlaps to Total Number of Overlaps Detected

False-Alarm Rate (R_f): Number of Incorrectly Detected Overlaps to Total Number of Overlaps Detected

F-measure (F): $F = \dfrac{2R_rR_p}{R_r + R_p}$

F-measure takes on values between 0 and 100, with higher values indicating better performance. The performance of the detection system is shown in table 1. It is seen that the phoneme selection scheme results in higher F-measure. In addition, utilizing the aperiodicity measure provides further gain in performance.

Table 1. Results obtained by omitting misleading phonemes in generating the artificial overlapped-speech data.

Removed Phonemes	**F(%)**	**R_f(%)**	**R_r(%)**	**R_p(%)**
No phonemes removed (baseline)	63.24	37.12	63.60	62.88
Nasals	63.30	36.20	62.80	63.80
Nasals and stops	63.42	36.33	63.19	63.67
Nasals, stops, and glides w/o aperiodicity	**65.49**	34.42	65.41	65.58
Nasals, stops, and glides w/ aperiodicity	**69.49**	34.31	73.77	65.69

5. REAL-WORLD IN-VEHICLE APPLICATION

In addition to artificial co-channel signals, we investigate the proposed system's performance in a real-world scenario where overlapped speech is part of the conversations between the parties. Based on social sciences studies [10], overlapped-speech is known to be a sign of the amount of competitiveness of the speakers and we believe that there exists a high correlation between the amount of competitiveness in a conversation and the amount of "focus" the driver can spend on his/her driving task. Detecting "unsafe" conversations by measuring the amount of overlapped-speech as a signature for highly competitive behavior in conversations could be a preliminary step towards developing speech based warning systems to alert the driver about the driving safety, without requiring sophisticated speech systems such as ASR to analyze the content of the conversations. An initial analysis is performed measuring the variations in driving performance while the driver is engaged in a conversation. For the analysis, the noisy in-vehicle speech data is enhanced using the optimally modified log-spectral amplitude (OM-LSA) estimation algorithm [21], and speech regions are automatically segmented using speech activity detection (SAD) [22].

A previous study in [14] demonstrated that in-vehicle speech activity affects driving performance. This study incorporates maneuver recognition and analysis algorithms presented in [23, 14]. However, no further analysis was made in [14] on the contextual information of the speech taken place. Using the described method for overlapped-speech detection, not only the speech segments, but also the competitive involvement can now be extracted and analyzed. The conversational speech data collection with UTDrive, [13], is designed with the focus to examine speech and particularly overlapped-speech as parameters that are known to partially determine driver behavior. The tasks in which the drivers have been asked to participate are designed to activate their competitiveness and involvement in order to increase the amount of overlapped-speech segments in the conversations [10]. The driving route is divided into four segments that are repeated in two phases. In the first phase the driver drives through the route without performing any secondary task to become familiar with the route and the vehicle. The second phase demands some extra activities each belonging to one segment of the route. The tasks are described below:

Segment 1: In this segment of the road the passengers will initiate a conversation by asking the driver questions about casual topics such as the weather (other topics may be chosen).

Segment 2: In this segment the driver will participate in a game called "I spy" that requires him/her to spot and name an object such as a billboard, traffic sign, etc. and call out its name before the other passengers. Whoever gets the maximum "I spy" correctly is the winner. A wrong "I spy" will result in a negative point.

Segment 3: A set of TIMIT sentences are played through a portable tablet and the driver is required to repeat each sentence before the next sentence is played.

Segment 4: A conversation is initiated by one of the passengers and the driver is asked to give his/her opinion on the subject and debate on their agreement or disagreement with the passenger.

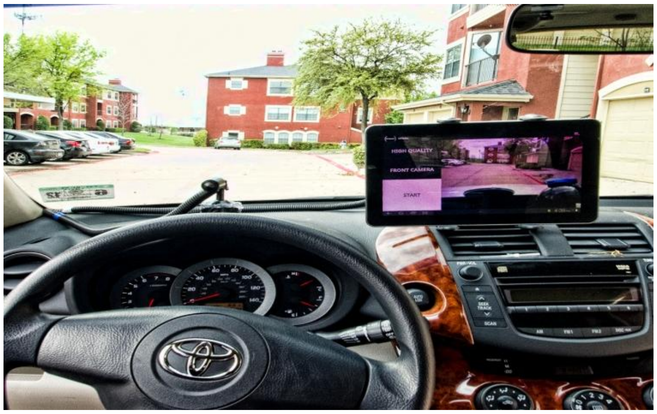


Fig. 3. A snapshot of inside the UTDrive data collection vehicle. The portable device is mounted on the wind-shield.

5.1. Driving maneuver recognition and analysis

In order to measure the driver's performance, maneuver recognition using CAN-bus signals has been previously utilized [13]. However, as shown in [23, 14], sensor information extracted from low cost portable devices can provide more accurate maneuver recognition. In this study, to better understand the influence of in-vehicle speech on driving performance, we collect in-vehicle speech along with all the available sensor information on the smart portable device (i.e., Tablet) in the UTDrive setup. Fig. 3 shows the inside of the UTDrive vehicle and the mounted Tablet. For more information about the maneuver recognition system see [23, 14].

Fig. 4 shows the driving performance for a segment of the route. The regions where the driving performance is normal are marked as green, while the yellow and red regions indicate moderate and risky driving, respectively. The time instances at which the driver is engaged in a conversation are shown as blue circles where the filled circles mark overlapped-speech. It is seen from the figure that there are at least 3-4 segments in the route where the driver performed a moderate/risky maneuver. Although these instances are short (20–40 s), they are likely to result in a crash of varying intensity. The same segment of the route was driven by the same driver in the first phase with no abnormalities in his normal driving pattern. The only change in the second phase is the engagement of the driver in a conversation. The driver's speech involvement and co-channel information shown in Fig. 4 follow the expected trend. In patch 2, the driver continues speaking even though he was recently involved in a competitive speech, hence compromising severely his driving performance.

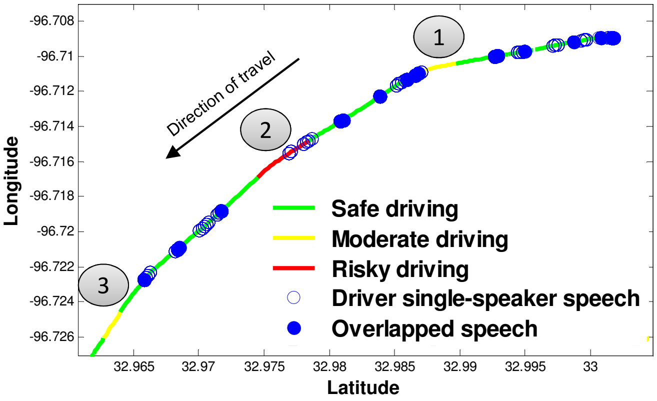


Fig. 4. The impact of overlapped speech on driver performance for a 5 min segment of the route. Competitive speech increases the driver's cognitive load which adversely impacts his driving performance. The driver is aware of this and predicts his inability to maintain normal driving and stops speaking in patches 1-3.

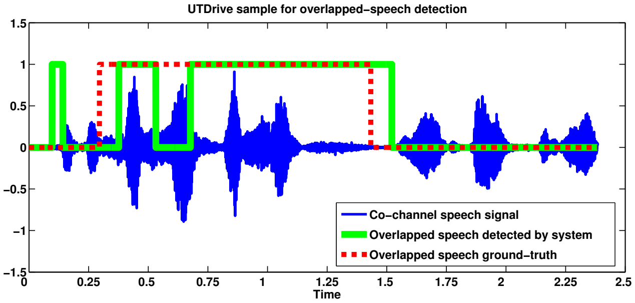


Fig. 5. Sample waveform to demonstrate overlapped-speech detection in UTDrive data. The ground-truth of overlapped speech regions(red dashed line) is manually transcribed. Green solid lines mark the regions obtained from the detection system.

5.2. Overlapped-speech detection for UTDrive data

The UTDrive conversational data consists of low SNR speech data in car noise. This noise decreases the performance of the overlapped-speech detection system to a great extent. Moreover, the channel difference between the train data (i.e., TIMIT) and the UTDrive data also degrades performance. To solve this problem, a small sample of single-speaker data is manually extracted from the UTDrive data to adapt the two models (i.e., single and overlapped) using maximum a posteriori (MAP) adaptation [24]. This MAP-adaptation technique is not impractical since it is reasonable to assume that a small sample of speech data can be extracted from the speech file. The extracted single-speaker speech data is divided into two segments and the speech data is summed to create a training set to be used in adapting the double-speaker model. Pilot experiments show that this technique improves the accuracy of the overlap detection system, making it possible to have a rough estimation of the amount of overlapped-speech data in a realistic co-channel speech recording (see Fig. 5). This figure shows that the detection system can be reliably used for in-vehicular environments.

6. CONCLUSION

The present study has proposed a system for the detection of overlapped-speech for simulated and actual scenarios. It was shown that the concatenation of spectral harmonicity and envelope information can effectively discriminate the single and overlapped speech models. The discrimination is further increased by selecting a subset of individual phonemes to be used in generating artificial overlapped data for model training. As an example application, we evaluated the effectiveness of the proposed framework in real in-vehicle data with the purpose of showing the correlation between the location and amount of overlapping speech segments and driver performance. We showed that the presented system alongside the features used in the experiments result in a reasonably accurate overlapped-speech detection.

7. ACKNOWLEDGEMENT

The authors would like to acknowledge the contributions of the CRSS lab manager James Hoover for his diligent work on the UTDrive data collection which made this study possible. We would also like to thank Dr. Kawahara, Wakayama University, for generously providing us with his implementation of STRAIGHT.

8. REFERENCES

[1] T. Quatieri and R. Danisewicz, “An approach to co-channel talker interference suppression using a sinusoidal model for speech,” *IEEE Trans. Acoustics Speech and Signal Process.*, vol. 3X, no. I, pp. 56–69, January 1990.

[2] D. Chazan, Y. Stettiner, and D. Malah, “Optimal multipitch estimation using the em algorithm for co-channel speech separation,” in *Proc. IEEE ICASSP*, April 1993, pp. 728–731.

[3] D. P. Morgan, E. B. George, L. T. Lee, and S. M. Kay, “Co-channel speaker separation by harmonic enhancement and suppression,” *IEEE Trans. on Speech and Audio Process.*, vol. 5, no. 5, pp. 407–424, September 1997.

[4] K. Boakye, “Audio segmentation for meeting speech processing,” Ph.D. dissertation, Fall 2008.

[5] S. N. Wrigley, G. J. Brown, W. Vincent, and S. Renals, “Speech and crosstalk detection in multichannel audio,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 13, no. 1, pp. 84–91, January 2005.

[6] R. E. Yantorno, “Cochannel speech study,” Electrical and Computer Engineering Department Temple University, Tech. Rep., September 1999.

[7] B. Smolenski and R. Ramachandran, “Usable speech processing: A filterless approach in the presence of interference,” *Circuits and Systems Magazine, IEEE*, vol. 11, no. 2, pp. 8 –22, 2011.

[8] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, “Overlapped speech detection for improved diarization in multiparty meetings,” in *Proc. ICASSP*, Las Vegas, Nevada, 2008, pp. 4353–4356.

[9] R. E. Yantorno, D. S. Benincasa, and S. J. Wenndt, “Effects of co-channel speech on speaker identification,” in *SPIE Intl. Symp. on Tech. for Law Enforcement*, November 2000.

[10] E. A. Schegloff, *Accounts of Conduct in Interaction: Interruption, Overlap and turn-taking, in J. H. Turner (ed.), Handbook of Sociolinguistics*. New York: Plenum, 2002, pp. 287–321.

[11] N. guidelines, “Visual-manual nhtsa driver distraction guidelines for in-vehicle electronic devices,” Tech. Rep., Febuary 2012.

[12] J. R. Treat, N. S. Tumbas, S. T. McDonald, D. Shinar, Hume, R. D., R. E. Mayer, R. L. Stanisfer, and N. J. Castellan, “Tri-level study of the causes of traffic accidents,” Tech. Rep., 1977.

[13] A. Sathyanarayana, S. O. Sadjadi, and J. H. L. Hansen, “Leveraging speech-active regions towards active safety in vehicles,” in *IEEE Intl. Conf. Emerging Signal Processing Applications, ESPA 2012*, Las Vegas, January 2012, pp. 48–51.

[14] ——, “Automatic driving maneuver recognition and analysis using cost effective portable devices,” in *SAE World Congress abd Exhibition*, Detroit, April 2013.

[15] P. F. Assmann, “Fundamental frequency and the intelligibility of competing voices,” in *Proc. of the Intl. Congress of Phonetic Sciences*, San Francisco, August 1999, pp. 179–182.

[16] Y. Shao and D. L. Wang, “Co-channel speaker identification using usable speech extraction based on multi-pitch tracking,” in *Proc. ICASSP*, Hong Kong, 2003, pp. 205–208.

[17] K. Krishnamachari, R. E. Yantorno, D. S. Benincasa, and S. J. Wenndt, “Spectral autocorrelation ratio as a usability measure of speech segments under co-channel conditions,” in *IEEE Intl. Symp. on Intelligent Signal Processing and Communication Systems, ISPACS*, November 2000, pp. 710–713.

[18] H. Kawahara, J. Estill, and O. Fujimura, “Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modication and synthesis system straight,” in *Proc. 2nd MAVEBA*, Firenze, Italy, 2001.

[19] K. Krishnamachari, R. E. Yantorno, J. M. Lovekin, D. S. Benincasa, and S. J. Wenndt, “Use of local kurtosis measure for spotting usable speech segments in co-channel speech,” in *Proc. ICASSP*, Salt Lake City, Utah, 2001, pp. 649–652.

[20] J. LeBlanc and P. de Leon, “Speech separation by kurtosis maximization,” in *Proc. ICASSP*, Seatle, Washington, 1998, pp. 1029–1032.

[21] I. Cohen, “Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator,” *IEEE Signal Process. Lett.*, vol. 9, pp. 113–116, April 2002.

[22] S. O. Sadjadi and J. H. L. Hansen, “Unsupervised speech activity detection using voicing measures and perceptual spectral flux,” *IEEE Signal Process. Lett.*, vol. 20, pp. 197–200, March.

[23] A. Sathyanarayana, S. O. Sadjadi, and J. H. L. Hansen, “Levaraging sensor information from portable devices towards automatic driving maneuver recognition,” in *IEEE 15th Intl. Conference on Intelligent Transportation Systems*, Anchorage, AK, September 2012.

[24] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital Signal Processing 10*, pp. 19–41, September 2000.