A Comparative Analysis of Overlapping Speech Detection Techniques That Utilize Machine Learning for Meeting-type Audio Recordings

Candy Espulgar[1], Neil Romblon[2] and Ronald Pascual[1,*]
[1] College of Computer Studies
De La Salle University
Manila, Philippines
**Corresponding Author: ronald.pascual@dlsu.edu.ph*

Abstract: Overlapping speech poses an obstacle in speech analysis, especially for speaker diarization. It is often desirable to either separate or disregard sections in audio recordings containing overlapping speech, as this process has been observed to improve the performance of speech analysis models. The primary process of automatically identifying the overlapping speech segments is known in the literature as overlapped speech detection or OSD (Xiao et al., 2011). This study aims to compare existing approaches used to detect overlapping speech in audio recordings. Although existing works were trained to handle different types of recordings, this study focused on the subset of works that were tested on meeting-type audio recordings. This study also attempts to address the issue of mismatched evaluation metrics through a novel approach called relative estimation. The resulting comparative analysis showed that for meeting-type audio recordings, a GMM-based detection approach trained with phoneme omission relatively gave the best results. However, insights and observations from this study reveal the need for a universal evaluation metric in order to reliably compare existing approaches. The need for a dedicated overlapping speech database to aid in the implementation of such a metric is also recommended.

Key Words: overlapping speech detection; machine learning

1. INTRODUCTION

Automatic Speech Recognition (ASR) research goes in-line with the ideas of the upcoming 4th Industrial Revolution. The 4th Industrial Revolution or **Industry 4.0** is characterized by smart decision-making through relevant data (Marr, 2018). Industry 4.0 can benefit greatly from ASR applications as ASR allows computers to capture information from conversations, specifically in the meeting conference scenario wherein people discuss a specific agenda. Future applications, such as AI-based meeting assistants can summarize and suggest actions based from the meeting's transcripts. Such assistants can aid the users to make better discussion and possibly better decisions. However, one of the current hurdles that automatic speech recognition faces is the inaccuracy caused by overlapping speech.

Overlapping speech occurs when multiple speakers speak simultaneously. It is a common occurrence in conversational speech and poses a great problem for speech analysis, especially towards speaker diarization (Ryant et al., 2018). Speaker diarization is a speech-to-text transcription task that solves the problem of "who spoke when" (Anguerra et al., 2012). Studies have shown that the presence of overlapping speech may increase diarization error rates up to 27% (Ryant et al., 2018). To address this, one may either separate the speakers in an overlapped speech segment, or completely exclude these segments in the data processing altogether. Regardless of the approach, overlapping speech segments must first be detected in order to be addressed; this task is called overlapping speech detection. There are numerous existing studies for overlapping speech detection. These approaches often involve machine learning (Zelenák & Hernando, 2011; Yella & Bourlard, 2014; Geiger et al., 2013a; Shokouhi et al., 2013) or deep learning techniques (Geiger et al., 2013b; Sajjan et al., 2018). Approaches that utilize other techniques also exist, but for the purpose of coherence, this study is limited to those that utilize machine learning. The studies included for comparison are also limited to those tested on meeting-type data. This is due to the current trend of studies being geared towards analyzing meetings (Anguerra et al., 2012).

The objective of this research is to provide a comparative analysis that aptly summarizes and compares existing overlapping speech detection approaches that were trained to handle meeting-type data. The output of this research is expressed as a table that notes key information about each approach and orders the entries based on a rough ranking. The biggest challenge of this research is the comparison of studies measured using different evaluation metrics. An attempt to resolve this is done through a method dubbed as relative estimation.

However, the output of this research is not meant to be a definite measure of performance. It simply gives an overview of the current state of the approaches in overlapping speech detection, which may be used as a reference for future works.

2. COMPARISON AND RELATIVE ESTIMATION OF OVERLAPPED SPEECH DETECTION METHODS

2.1 Overlapping Speech Detection Techniques

The overlapping speech detection techniques included in this research are described below. Studies A, B, and C propose new features for use in an HMM/GMM classifier. Studies D and E employ deep learning techniques for overlap detection. Study F proposes a preprocessing step for improving overlap detection performance.

2.1.1 Speaker Overlap Detection with Prosodic Features for Speaker Diarisation

The study (Zelenák & Hernando, 2011) explores on the use of prosodic features in addition to usual spectral features (Mel- frequency cepstral coefficients or MFCC, residual energy, spectral flatness) in speech analysis. Prosody refers to the broader aspects of speech such as stress, rhythm, and intonation. These prosodic characteristics are quantified in the study as pitch, intensity, first four formant frequencies, and as well as their long-term statistics.

The study utilized the AMI meeting dataset for training, testing, and evaluation. Evaluation was done by comparing a baseline system with the proposed system. The baseline is a Hidden Markov Model / Gaussian Mixture Model (HMM/GMM)-based system which uses spectral-based features. The proposed system is also HMM/GMM-based but uses prosodic features on top of the baseline system's spectral-based features. Both systems consider a segment as one of three classes: non-speech, single-speaker speech, and overlapping speech. Additional evaluation is performed by integrating both overlapping speech detection systems to a speaker diarization system to perform overlap exclusion + labelling. Annotations for training and testing are obtained from performing force-alignment through a recognizer.

2.1.2 Overlapping Speech Detection using Long-term Conversational Features for Speaker Diarization in Meeting Room Conversations

The study (Yella & Bourlard, 2014) explores and takes advantage of contextual information available in a conversation. Nuances in conversation such as silence patterns and speaker turn changes are shown to have a relation with the occurrence of overlapping speech. It has been found that in a time segment, silence duration is inversely proportional, and the number of speaker changes is directly proportional to the probability of an overlap occurring. Hence, these long-term statistics have been used by the study as supplementary features for overlapping speech detection.

For the experiments, the AMI meeting dataset, NIST RT ('05, '06, '07, and '09) meeting corpus, and ICSI meeting dataset were used. The baseline overlap detection system is an HMM/GMM-based system that uses the following short-term acoustic features: 12 MFCCs with log energy, spectral flatness, and 12th order LP (linear prediction) residual energy. The proposed system adds long-term conversational features on top of the baseline system. The long-term conversational features are in the form of a prior probability after an initial pass of speaker diarization is performed and its speaker change, and silence statistics are computed. Both schemes are then evaluated with its performance on a speaker diarization system.

2.1.3 Using Linguistic Information to Detect Overlapping Speech

The study (Geiger et al., 2013b) considers common words spoken during an overlap for improving overlap detection. Some words, such as backchannel utterances (e.g. "yeah", "mm-hm") are known to occur commonly in overlapping speech (Gravano et al., 2007). To take advantage of this relation, the study utilizes unigram language models trained from overlapped speech segments (and another from single speaker segments), then using the word's probability to supplement the baseline system's prediction.

The AMI meeting dataset was used for the experiments in the study. The performance was evaluated through a comparison between the proposed system and a baseline system. The baseline system is an HMM/GMM-based system that uses energy, spectral, voicing-related, and convolutive non-negative sparse coding (CNSC)-based features (all together will be referred to as ESVC). The proposed system adds an oracle-style Automatic Speech Recognition system (perfect accuracy; ground-truth) to obtain the words from the segment and determine, from the language models, its probability of being an overlapping speech segment. Only the longest word from the overlap segment was used in determining the probability of an overlap given a word. The resulting probability is used in conjunction with the baseline results to obtain the final prediction. Evaluation is conducted by comparing the system's performance with complete set and subsets of the baseline features with and without the linguistic probability.

2.1.4 Detecting Overlapping Speech with Long Short-Term Memory Recurrent Neural Networks

The study (Geiger et al., 2013a) uses a long short-term memory recurrent neural network regressor as an additional source for features and as a standalone classifier. A recurrent neural network is a deep learning technique that has been known to work well in speech analysis tasks, since it is able to exploit previous instances' events.

The experiments in the study were performed by using the AMI meeting dataset. The study uses an HMM/GMM-based classifier as the baseline, and an LSTM regressor as a supplementary feature source for the baseline as well as its own standalone classifier. The baseline features are the same as Geiger et al.'s (2013b) baseline. The LSTM is trained by assigning overlapped speech segments with 1, single-speech with 0, and non-speech with -1. The standalone LSTM classifier works by using a threshold to determine if the segment is overlapped or non-overlap. Evaluations are conducted by feature-system configuration pairs. Possible feature configuration values are: MFCC only, ESVC, MFCC+LSTM prediction, ESVC+LSTM prediction. On the other hand, system configuration values are: HMM and LSTM.

2.1.5 Leveraging LSTM Models for Overlap Detection in Multi-Party Meetings

The study (Sajjan et al., 2018) employs different deep learning approaches to overlapping speech detection and investigates which perform best. Four detection systems were discussed in the study: GMM-based (classical machine learning), deep neural networks (DNN), convolutional neural networks (CNN), and long short-term memory networks (LSTM).

The study tests on both TIMIT and AMI datasets, however, for the conciseness of this review, only the test on AMI is discussed. The baseline overlap detection system is GMM-based that has kurtosis, SFM, and MFCC+D as its features. The deep learning approaches, on the other hand, derive derives features from the mel-spectrogram. The feature set for the deep learning techniques are fbank, or the logarithm of the weighted sum of spectral energy in each bin. The feature vector is in the context of 11 frames, with 5 previous from the previous frame, 1 current, and 5 next.

2.1.6 Overlapped-speech Detection with Applications to Driver Assessment for In-vehicle Active Safety Systems

The study (Shokouhi et al., 2013) uses overlapping speech as a means to detect competitive, and potentially dangerous, conversations which may distract the driver. It uses spectral features (MFCC, aperiodicity, kurtosis, spectral flatness) to analyze phonemes. It also omits misleading phonemes during the training phase to improve performance.

Experiments in the study were conducted while using the TIMIT dataset. The system consists of two GMMs, one for a single-speaker model and another for a double-speaker model. Each model is trained using phonemes taken from the TIMIT dataset. The single-speaker model used individual phonemes while the double-speaker model used artificially combined phonemes. Different subsets of misleading phonemes (nasals, stops, and glides) were omitted when generating the artificial data.

Evaluation was done by training the system on different types of phoneme omission schemes. The baseline used was the performance of the system when trained without phoneme omission.

2.2 Comparison and Relative Estimation

The approaches are compared and ranked based on their common evaluation metrics. Studies without a common evaluation metric are compared through a novel technique to be referred to as relative estimation. Before applying relative estimation, the following must first be satisfied: (1) there must be an evaluation metric that is common to at least half of the studies, which will serve as the pivot metric; and (2) studies that do not have the pivot metric must have an evaluation metric (relative metric) common to at least one study that has the pivot metric.

The approaches with the pivot metric (called pivot approaches) are initially ranked through the pivot metric. Approaches that lack this metric (called relative approaches) are ranked by comparing their relative metric(s) against pivot approaches that have a matching relative metric. This allows the relative approaches to be inserted in the rankings based on how their relative metrics compare with those of the pivot approaches.

The overlapping speech detection techniques included in this research are shown in Table 1. Studies A, B, and C propose new features for use in an HMM/GMM classifier. Studies D and E employ deep learning techniques for overlap detection. Study F proposes a preprocessing step for improving overlap detection performance.

Table 1. Overlapping Speech Detection Techniques

Study	Approach	Author(s)
A	Prosodic Features HMM/GMM	Zelenak & Hernando (2011)
B	Long-term Conversational Features HMM/GMM	Yella & Bourlard (2014)
C	Linguistic Features HMM/GMM	Geiger et al. (2013a)
D	LSTM Regressor	Geiger et al. (2013b)
E	LSTM Classifier	Sajjan et al.(2018)
F	Phoneme Omission GMM	Shokouhi et al. (2013)

Table 2. Evaluation Metrics per Study

Evaluation Metric	Study
F-measure	A, B, C, D, F
Precision	A, C, D, F
Recall	A, C, D, F
Detection Accuracy	C, D, E
Relative DER Improvement	A, B, E

Table 3. Comparative Analysis of Overlapping Speech Detection Techniques for Meeting-type Audio Data

Study	Approach	Dataset		Audio Features	Evaluation Metric				
		Train	Test		F-Measure (pivot metric)	Detection Accuracy	Relative DER	Precision	Recall
F	Phoneme Omission GMM	TIMIT		Filtered Phonemes[1]	**0.69**	-	-	65.59	**73.77**
E	LSTM Classifier	TIMIT AMI	AMI	F-bank in 11 frame context[2]	-	**68.4**	**+21%**	-	-
A	Prosody HMM/GMM	AMI		Spectral and Prosodic features	0.51	-	+7%	76	38
D	LSTM Regressor	AMI		ESVC[3] and LSTM Prediction	0.45	23.1[4,5]	-	78.6	31.7
B	Long Conversational HMM/GMM	AMI NIST RT[6] ICSI		Acoustic and Long-term Conversational	0.44	-	+17%	-	-
C	Linguistic HMM/GMM	AMI		ESVC and Language Model Probability	0.42	21.7[5]	-	**81.7**	28

Table 2 shows the evaluation metrics present in each study. It can be seen that F-measure is the most prominent, closely followed by precision and recall. While detection accuracy and relative DER improvement are used by half of the studies. As such, F-measure was chosen as the pivot metric, and detection accuracy then relative Detection Error Rate (DER) improvement are the relative metrics of studies C, D, E, and A, B, E respectively. Precision and recall were not included as they are components of F-measure.

3. RESULTS AND DISCUSSION

Table 3 shows the resulting comparative analysis table. Study F was ranked the highest, with an F-measure value of 0.69. This is followed by Study E, whose ranking is derived through relative estimation since it did not have the pivot metric. The ranking of Study E was then derived by comparing it to studies that matched one of its relative metrics (either detection accuracy or relative DER). In this regard, Study E was ranked higher than Study D due to having a better value on its relative metric of detection accuracy. It was also ranked higher than Study A due to having a higher relative metric value for its relative DER. This resulted to the current ranking of: F, E, A, D, B, C.

4. CONCLUSIONS

The resulting comparative analysis table showed that the best approach for overlapping speech detection in meeting-type audio recordings is a Gaussian Mixture Model-based approach trained under a phoneme omission scheme. This was followed by an LSTM-based classifier approach and an HMM-GMM-based approach that utilized prosodic features.

It is important to note that, despite the attempt to address mismatched evaluation metrics through relative estimation and analysis, the resulting ranking should be taken as a rough approximation rather than a definite measurement of performance. The lack of a universal evaluation metric makes it difficult to satisfactorily compare the results of different studies. To establish a universal evaluation metric, a topic that future research can focus on is the relationship of precision and recall to overlap detection accuracy, and overlap detection accuracy to the relative improvement in DER. It can be seen in Table 3 that despite study A has the greater F-measure compared to study B, the relative

[1] Phonemes w/o nasals, stops, and glides with aperiodicity

[2] The 11 frames consist of the 5 previous, 1 current, and 5 following frames

[3] Energy & Spectral, Voicing-related, and CNSC-based features

[4] Taken using (100 – detection error rate)

[5] Value taken from the best performing LSTM scheme

[6] NIST RT '05, '06, '07 were used for training and NIST RT '09 for testing

DER improvement is greater for study B. It is possible that precision and recall might not have equal impact on the system's performance, which is not captured by the F-measure.

Furthermore, there is also a need for databases which have a focus on overlapping speech data. Existing databases are mainly aimed towards speech analysis in general and only contain a partial amount of overlapping speech data. Having such database allows for a more robust training and testing process. This can also be tied to the previously mentioned universal evaluation metric, where such a database can be used as a common basis for evaluation.

5. ACKNOWLEDGMENTS

This paper is an outcome of the CSC784M Digital Signal Processing class for Speech and Image Processing. As such, the members of the class for providing feedback during the preliminary and final presentations.

6. REFERENCES

Anguerra, X., Bozonnet, S., Evans, N.W., Fredouille, C., Friedland, G., & Vinyals, O. (2012). Speaker Diarization: A Review of Recent Research. IEEE Transactions on Audio, Speech, and Language Processing, 20, 356-370.

Geiger, J.T., Eyben, F., Schuller, B.W., & Rigoll, G. (2013a). Detecting overlapping speech with long short-term memory recurrent neural networks. INTERSPEECH.

Geiger, J.T., Eyben, F., Evans, N.W., Schuller, B.W., & Rigoll, G. (2013b). Using linguistic information to detect overlapping speech. INTERSPEECH.

Gravano, A., Benus, S., Hirschberg, J., Mitchell, S., & Vovsha, I. (2007). Classification of discourse functions of affirmative words in spoken dialogue. In *Eighth Annual Conference of the International Speech Communication Association*.

Marr, B. (2016, April 06). Why Everyone Must Get Ready For The 4th Industrial Revolution. Retrieved from: https://www.forbes.com/sites /bernardmarr /201604/05/why-everyone-must-get-ready-for-4th-industrial-revolution/#730e42ec3f90

Ryant, N., Bergelson, E., Church, K., Cristià, A., Du, J., Ganapathy, S., Khudanpur, S., Kowalski, D., Krishnamoorthy, M., Kulshreshta, R., Liberman, M., Lu, Y., Maciejewski, M., Metze, F., Profant, J., Sun, L.M., Tsao, Y., & Yu, Z. (2018). Enhancement and Analysis of Conversational Speech: JSALT 2017. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5154-5158.

Sajjan, N., Ganesh, S., Sharma, N.K., Ganapathy, S., & Ryant, N. (2018). Leveraging LSTM Models for Overlap Detection in Multi-Party Meetings. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5249-5253.

Shokouhi, N., Sathyanarayana, A., Sadjadi, S.O., & Hansen, J.H. (2013). Overlapped-speech detection with applications to driver assessment for in-vehicle active safety systems. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2834-2838.

Yella, S.H., & Bourlard, H. (2014). Overlapping Speech Detection Using Long-Term Conversational Features for Speaker Diarization in Meeting Room Conversations. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22, 1688-1700.

Xiao, B., Ghosh, P. K., Georgiou, P., & Narayanan, S. S. (2011, May). Overlapped speech detection using long-term spectro-temporal similarity in stereo recording. In 2011 *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP). (pp. 5216-5219). IEEE.

Zelenák, M., & Hernando, J. (2011). The Detection of Overlapping Speech with Prosodic Features for Speaker Diarization. INTERSPEECH