Journal of Applied Engineering and Technological Science
Vol 6(1) 2024: 355-364

CROWD SPEAKER IDENTIFICATION METHODOLOGIES, DATASETS AND FEATURES: REVIEW

Ghadeer Qasim Ali[1], Husam Ali Abdulmohsin[2*]
Computer Science Department, College of Science, University of Baghdad, Iraq[12]
ghadeer.ali2201m@sc.uobaghdad.edu.iq[1], husam.a@sc.uobaghdad.edu.iq[2*]

Received: 02 May 2024, Revised: 04 August 2024, Accepted: 18 August 2024
**Corresponding Author*

ABSTRACT
Crowded speech or Overlapping speech, occurs when multiple individuals speak simultaneously, which is a common occurrence in real-life scenarios such as telephone conversations, meetings, and debates. The critical task in these situations is to identify all the speakers rather than just one. Overlapping speech identification is a significant research domain with applications in human-machine interaction, criminal detection in airports, trains, and public spaces. Our work examines crowd speech identification from four perspectives, including the most commonly used datasets, the most effective features for crowed speaker identification, and the best methodologies employed, and the highest results gained. This study proposes a comprehensive survey of research on crowd speech identification, covering the period from 2016 to present. The survey includes ninety research papers, fifty of which, are empirical studies. Initially, statistical methods were predominant, but the current trend leans towards artificial intelligence, particularly deep learning, which has demonstrated considerable efficacy in this field.
Keywords*: Crowed Speaker Identification, Deep Learning, Speech Features, Crowed Datasets.*

1. Introduction

The difficulties experienced by air traffic controllers in the early 1950s served as a major inspiration for much of the early research in this field. At that time, controllers heard pilot communications via the control tower's loudspeakers, with the sound of multiple pilots' voices mingling over a single loudspeaker, making it a highly challenging assignment for controllers (Kantowitz & Sorkin, 1983). This situation prompted the first technical effort that specifically addressed the "cocktail party problem." Cherry suggested several elements that would facilitate the creation of a "filter" to distinguish voices: different directions represented by the voices, lip-reading and similar behaviors, variations in speaking styles, average speeds, pitches, male vs. female voices, accent variations, and probabilities of transitions based on grammar, voice dynamics, and subject matter (Mitchell et al., 1971).

Early experiments demonstrated that subjects could accept noises in one ear while rejecting information played to the other, with either ear capable of recognizing the target at any time. This ability to shadow one message easily led to an understanding of the cocktail party effect, where binaural listening helps noise from different directions less effectively obscure the intended signal (Quatieri & Danisewicz, 1990). This principle is applied in earphones for fighter pilots, which deliver an antiphase signal to one ear synchronized 180 degrees with the signal delivered to the opposite ear, aiding in separating spoken signals from the cockpit's loud noises.

In 1971, Bell Labs researchers described a signal processing technique for isolating a speech signal from a known location amidst a backdrop of other sounds. Their system used an array of four microphones and basic computational components to achieve a 3-6 dB noise suppression. However, the source had to remain exactly centered in the microphone array, making this approach somewhat impractical. Bregman later described various computational approaches for separating speakers based on tracking principles, although these systems faced limitations due to unvoiced speech sounds and difficulties in detecting the fundamental frequency as the number of speakers increased(Quatieri & Danisewicz, 1990). Weintraub discovered that distinguishing between stronger and weaker voices improved speech recognition accuracy.

Despite these advancements, current conferencing systems still face restrictions on the number of people who can speak simultaneously, making it challenging to distinguish

individual speakers. Adding spatial audio cues to conferencing environments can help distinguish speakers, particularly if video is also introduced.

Interest in crowded speech identification has grown since 2011(Sun & Ma, 2011), and there are numerous applications of speech processing(Alsalam et al., 2021)(Shakat et al., 2021)(T. S. Mohammed et al., 2021)(Z. K. Mohammed & Abdullah, 2022)(Swadi & Ali, 2019)(Svendsen & Kadry, 2024), which has gained increasing significance in telecommunications, healthcare(Abdulmohsin, Al-Khateeb, et al., 2022)(Abdulmohsin, 2022), and entertainment(Abdulmohsin, Stephan, et al., 2022). Statistical modeling techniques, particularly Hidden Markov Models (HMMs), have been pivotal in advancing the field(Gales & Young, 2008)(Rabiner, 1989). These models have enabled significant breakthroughs in speech processing research and development.

In recent years, the field of speech processing has been transformed by powerful tools, including machine learning and deep learning. These technologies have revolutionized the analysis and processing of speech signals through deep neural networks (DNNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs). These architectures have proven highly effective in various speech-processing applications, such as speech recognition, speaker recognition, and speech synthesis (Mehrish et al., 2023).

A survey paper in 2015 highlighted the need for optimizing the separation and analysis of various sounds, noting the limited understanding of how other speech features, such as speaking style, linguistic variety, and contextual information, impact auditory grouping of speech(Bronkhorst, 2015). Research has primarily focused on voice properties and spatial signals, with less attention given to other characteristics that affect target speech understandability. Freyman et al. (2001) and Iyer et al. (2010) suggested altering the features of interfering speech or adjusting various target speech property combinations to maintain intelligibility.

Further research is needed on the dynamic aspects of attention, particularly the impact of switching voices or locations, as well as changes in other speech characteristics.

This study on crowded speech identification builds on one previous literature reviews (Arons, 1992) and aims to provide a comprehensive overview of the field. Our work is structured to familiarize interested parties with previous studies, presented in table form. The main contribution of this study is to present a crowded speech identification approach, feature, number of speakers, accuracy, and dataset.

2. Literature Review

(Thakker et al., 2018)explored the complexities and methodologies involved in identifying individual speakers within environments where multiple speakers overlap. The study emphasizes the challenge of distinguishing speakers in settings akin to the "cocktail party effect," where numerous voices converge simultaneously. The authors review various approaches, including the use of Mel Frequency Cepstral Coefficients (MFCC) and Dynamic Time Warping (DTW) techniques for voice recognition, and highlight the effectiveness of neural network-based systems. Additionally, they discuss the utilization of Gaussian Mixture Models (GMM) and Back Propagation Neural Networks (BPNN) to improve speaker isolation and identification accuracy. The findings suggest that advanced machine learning algorithms and feature extraction methods are pivotal in enhancing the performance of speaker identification systems in multi-speaker environments.

(Chang et al., 2019)the authors address the challenges of speaker recognition using a sparse dataset. They propose a method leveraging Triplet Neural Networks (TNNs) to create a latent space representation for the Multi-Target Speaker Detection and Identification Challenge Evaluation (MCE 2018) dataset, which includes i-vectors from 3,631 speakers, with only three samples per speaker. The research demonstrates that TNNs significantly improve speaker identification accuracy, particularly in scenarios with limited training data. When both the training and development sets are used, the proposed TNN model outperforms the baseline by 23%. Even when the training data is reduced, the TNN approach achieves a 46% improvement over the baseline. This indicates that TNNs are highly effective in handling sparse datasets with

few instances per class, showcasing their potential for robust speaker detection and identification tasks.

(Yousefi & Hansen, 2020) the authors investigate the detection of overlapping speech in short segments (as brief as 25 ms) using convolutional neural networks (CNNs). They utilize features such as Mel Frequency Cepstral Coefficients (MFCC), Mel Filter-Bank (MFB), spectral magnitude, and pyknogram. The proposed CNN model demonstrates significant accuracy improvements, with detection accuracy ranging from 79% to 89% depending on the feature set and speaker gender combinations, indicating its efficacy in detecting overlapping speech in various conditions.

(Wu et al., 2020)introduced a novel method for enhancing speech signals in multi-speaker environments by integrating enhancement and recognition processes. The authors propose a multitask training approach that combines enhancement and recognition tasks to minimize distortion and improve recognition accuracy. An end-to-end joint training paradigm optimizes both the enhancement and acoustic models, using a hybrid deep learning framework with a convolutional, long short-term memory, and fully connected network (CLDNN) trained with the connectionist temporal classification (CTC) objective. The method demonstrates significant improvements in character error rate (CER) and perceptual evaluation of speech quality (PESQ), showing its effectiveness in enhancing multi-channel target speech and reducing distortion in overlapped speech scenarios.

(Mao et al., 2020)Recent advancements in automatic speech recognition (ASR) and speaker diarization (SD) have shifted from using separate models to joint frameworks that leverage audio-lexical interdependencies, improving word diarization performance in extended multi-speaker conversations. introduced a new benchmark using hour-long podcasts from "This American Life" to evaluate these approaches. Their research demonstrates that while separate ASR and SD models perform well with known utterance boundaries, joint models excel in handling long conversations without predefined boundaries. Key innovations include a striding attention decoding algorithm and data augmentation techniques, such as ShiftAug and AlignAug, to better manage inter-utterance dependencies and imprecise boundaries. These methods, along with model pre-training, significantly enhance ASR and SD performance, making joint models particularly effective for practical applications in long-duration, multi-speaker audio scenarios.

(Sato et al., 2021)addresses the challenge of recognizing overlapping speech by proposing a dynamic switching mechanism between enhanced and observed signals, based on signal-to-interference ratio (SIR) and signal-to-noise ratio (SNR). The study highlights that while speech enhancement (SE) can improve recognition performance, it may also introduce artifacts that degrade performance under certain conditions. To counter this, a rule-based switching mechanism was developed, allowing the system to alternate between using the raw observed mixture and the enhanced speech. Experimental results demonstrated that this switching approach improves automatic speech recognition (ASR) accuracy compared to always using enhanced signals, particularly in scenarios with varying noise types and levels. This method was tested using a dataset created from the Corpus of Spontaneous Japanese and the CHiME-4 corpus, revealing significant performance gains across different SIR and SNR.

3. Crowed speech definition and description

Crowed speech is the term used in linguistics and speech research to describe when two or more speakers produce utterances simultaneously within a shared temporal window, The time span that the utterances overlap is specified by this window. The precise duration may differ based on the subject of the talk, cultural conventions, and the delivery style of each speaker. According to estimates, up to 13% of spoken interactions may involve this phenomenon, which is common in natural conversation. While most study focuses on dyadic (two-person) interactions, the term "overlapping speech" can include scenarios in which multiple people are speaking at the same time. Even still, our brains manage an amazing feat: they analyse the overlapping melodies and tune in to the unique subtleties and rhythm of each speaker, all despite the disorganized cacophony. We can distinguish between people by using a complex

combination of auditory signals, language context, and social awareness to determine who is speaking, what they are saying, and even the subliminal messages hidden in their words.

To sum up, overlapping speech is a complicated and dynamic phenomenon with important linguistic, social, and cognitive ramifications that goes beyond just a distracting byproduct of discussion. In order to improve our understanding of human communication and develop artificial intelligence (AI) systems that are actually able to communicate with humans, it is imperative that we investigate and comprehend this phenomenon.

4. Crowed speech Identification challenges

Crowed Speech that represent a big challenge. By deciphering these convoluted talks, robots that can not only tolerate interference but also extract subtle meaning and speaker intent are being built. This is an ambitious technological endeavour Let's discuss the particular challenges in more detail:

1. Speech that overlaps: When several speakers talk at once, individual words may be obscured or distorted, making it challenging to discern the contributions of each speaker.
2. Overlapping speech frequently results in ambiguity and confusion. It may be difficult to distinguish the intended words due to misinterpretations of the combined audio, which may mimic phonemes or words from either speaker. The model finds it challenging to choose which speaker it should be listening to due to this "cocktail party effect".
3. Tracking Speakers: Another major problem is to identify and reliably track individual speakers throughout an overlapping conversation. Since most traditional ASR models assume there is only one active speaker, it is challenging to adapt to multi-voice real-world circumstances.
4. processing Cost: Complex methods such as deep learning-based systems and speaker diarization sometimes demand large amounts of processing power to handle overlapping speech. For real-time applications, where efficient processing and minimal latency are essential, this could provide problems.

5. Deep Learning and Statistics Approaches for Identifying Overlapping Speech

With their own advantages and disadvantages, statistical and deep learning techniques both provide viable answers.

5.1 Deep Learning Approaches

- Convolutional Neural Networks (CNNs): These are highly effective in identifying intricate details in audio recordings. CNNs are useful for differentiating overlapping speech from single-speaker segments because they can recognize spatiotemporal patterns in spectrograms, Mel-Frequency Cepstral Coefficients (MFCCs), and other representations(Mehrish et al., 2023).
- Recurrent Neural Networks (RNNs): RNNs are capable of capturing temporal dependencies in audio signals, particularly Long Short-Term Memory (LSTM) networks. They are therefore appropriate for simulating the dynamics of overlapping speech, in which the identities of the speakers and the active voices might fluctuate rapidly(Mehrish et al., 2023).

5.2 Statistical Approaches

- Hidden Markov Models (HMMs): HMMs have a long history of modeling speech signals. These can be used to create speaker-specific models and find state sequences that match individual and overlapped speech segments(Mehrish et al., 2023).
- Gaussian Mixture Models, or GMMs, are capable of simulating the probability distribution of various speech characteristics. Segment classification can be achieved by statistical inference by training several GMMs for single and overlapping speech(Mehrish et al., 2023).

- Support vector machines (SVMs): Using a variety of features, SVMs are able to learn the discriminative borders between single and overlapping speech. They provide interpretable models and are resilient to noisy data(Mehrish et al., 2023).

5.3 Combining Deep Learning and Statistics

The strengths of both deep learning and statistical methods can be combined in a hybrid approach. CNNs, for instance, are capable of extracting rich information, but HMMs or GMMs can offer an organized framework for modeling speech events that overlap.

6. STATE OF THE ARTWORKS SURVEY

The tables provide a comprehensive overview of various speech processing techniques, their features, classification methods, and performance results from 2017 to 2023. In 2017, the Anchored Deep Attractor Network (ADAN) using log spectral magnitude achieved a 62.8% reduction in Word Error Rate (WER). In 2018, multiple methods were explored: Utterance-level permutation invariant training with discriminative learning (uPIT-DL) using spectrogram and MFCCs resulted in a 22.0% WER reduction; Gated Convolutional Network (GCN) with MFCCs yielded a 15% improvement; Deep clustering with MFCCs showed a 13.9% reduction in Character Error Rate (CER); Recurrent Neural Network (RNN) using MFCCs achieved an 83.1% improvement; and SSN using magnitude spectrum led to a 33.3% WER reduction.

In 2019, techniques using filter bank features included Permutation invariant training (PIT), which showed a 5% WER reduction, and Bidirectional Long Short-Term Memory (BLSTM), which improved the signal-to-distortion ratio by 0.7 dB. Deep Clustering achieved a 16.5% WER reduction, the CTC attention-based encoder-decoder framework resulted in a 15% WER reduction, and a novel neural sequence-to-sequence architecture yielded a 60% WER reduction. By 2021, TDNN using MFCC features demonstrated a 96.9% improvement, as shown in table 1.

Table 1 - Previous Studies on wsj0-2mix Datase

Ref.	Year		Features	Classification	Results
(Chen et al., 2017)	2017	-	log spectral magnitude	Anchored Deep Attractor Network (ADAN)	62.8%WER reduction
(Fan et al., 2018)	2018	-	Spectrogram	Utterance-level permutation invariant training with discriminative learning (uPIT-DL)	22.0%
		-	MFCCs		
(Chang et al., 2018b)	2018	-	MFCCs	Gated Convolutional Network (GCN)	15% improvement
(Settle et al., 2018)	2018	-	MFCCs	Deep clustering	13.9% (CER)
(Seki et al., 2018)	2018	-	MFCCs	Recurrent Neural Network (RNN)	83.1% improvement
(Yoshioka et al., 2018)	2018	-	magnitude spectrum	SSN	33.3% WER reduction
(Shangguan & Yang, 2019)	2019	-	filter bank	Permutation invariant training (PIT)	5% WER reduction
(Huang et al., 2019)	2019			Bidirectional Long Short-Term Memory (BLSTM)	Reduced 0.7 dB signal-to-distortion ratio (SDR)
(Menne et al., 2019)	2019	-	filter bank	Deep Clustering	16.5%WER reduction
(Zhang et al., 2019)	2019	-	filter bank	CTC attention-based encoder-decoder framework	15%WER reduction
(Chang et al., 2019)	2019	-	filter bank	Novel neural sequence to sequence architecture	60%WER reduction
(Shi & Hain, 2021)	2021	-	MFCC	TDNN	96.9%

Additional techniques were highlighted for their performance improvements. In 2018, spectrogram features with Convolutional Neural Network (CNN) and filter bank features with Encoder, Attention, and Decoder (AED) achieved a 97% WER reduction. Serialized Output Training (SOT) using filter bank features was also explored. In 2021, Compositional embedding with MFCCs resulted in a 22.93% reduction in DER, Modification E2E SA-ASR with filter bank features yielded a 19.2% WER reduction, and Hypothesis Stitcher with raw input showed a 17.8% WER reduction. By 2023, the CNN Feature Extractor with Convolutional Neural Network approach achieved an 11.94% WER reduction, as shown in table 2.

Table 2 - Previous Studies on Librispeech Dataset

Ref.	Year	Features	Classification	Results
(Kunešová, 2018)	2018	- spectrogram	Convolutional Neural Network	
(Kanda, Gaur, Wang, Meng, & Yoshioka, 2020)	2020	- filter bank	Encoder, Attention, and Decoder (AED)	97%
(Kanda, Gaur, Wang, Meng, Chen, et al., 2020)	2020	- filter bank	Serialized Output Training (SOT)	
(Li & Whitehill, 2021)	2021	- MFCC	Compositional embedding	22.93%DER
(Kanda et al., 2021)	2021	- filter bank	Modification E2E SA-ASR	19.2% WER reduction
(Chang et al., 2021)	2021	- raw input	Hypothesis Stitcher	17.8% WER reduction
(Meng et al., 2023)	2023	- CNN Feature Extractor	Convolutional Neural Network	11.94%WER reduction

The studies using the AMI dataset provided additional insights. In 2017, a combination of MFCC, Spectral Energy, Loudness, In-Band Energy, Spectral Flux, and Spectral Kurtosis with a Source mixing approach achieved an 80% improvement. In 2018, filter bank features with PIT-ASR model and CNN-BLSTMs resulted in a 10% WER reduction, while Permutation Invariant Training (PIT) showed a 25% WER reduction. In 2019, MFCC features with an Acoustic Model (AM) achieved a 30.12% WER reduction. These studies collectively showcase the significant advancements in speech processing accuracy through the application of diverse features and sophisticated classification methods, as shown in table 3.

Table 3 - Previous Studies on AMI Dataset.

Ref.	Year	Features	Classification	Results
(Andrei et al., 2017)	2017	- MFCC - Spectral Energy - Loudness - In-Band Energy - Spectral Flux - Spectral Kurtosis	Source mixing approach	80%
(Chang et al., 2018a)	2018	- filter bank	PIT-ASR model with CNN-BLSTMs	10% WER reduction
(Qian et al., 2018)	2018	- filter bank	Permutation Invariant Training (PIT)	25% WER reduction
(Kanda et al., 2019)	2019	- MFCC	Acoustic Model (AM)	30.12%WER reduction

For statistical methods review, in 2016,(Kumar & Kumaraswamy, 2016) proposed to separate the co-channel speech mixture Empirical mode decomposition (EMD) by denoising at the front end along with pitch estimation. EMD relies on analyzing the signal to its source components without the need for prior training on big data as in deep learning techniques. Results show that speaker identification correct rate increases, the feature extraction used in this method is a mel cepstral coefficients, and the data collected from 20 speakers.

In 2019,(Wang & Sun, 2019) used Mel-Frequency Cepstral Coefficients (MFCCs) to extract the acoustic features of the voices. Calculate the Euclidean distance between the input voice and the average voice of each speaker. and then compare the Euclidean distance between the input voice and the mixed voices of different speaker combinations to determine which speaker is the least likely. In the end calculate the Manhattan distance between the input voice and the MFCCs of each speaker to find the closest match.

In 2019,(Bansal et al., 2019) proposed hybrid (deep and statistical) model, by using Gaussian Mixture Model and wavelet decomposition to enhance the speech quality and High-level noise suppression. The next step is analyzing frequency-to-delay ratio, dictionary tracing, and channel compensation methods to separate the speech segments and Participant contribution extraction. Feature generation using wavelet shrinkage, I-vector scoring, acoustic scoring, log-likelihood measure, and MFCC to create a vast adaptive feature set. This represented the

statistical part. Classification of speakers using a probabilistic deep neural network trained on the extracted feature pool.

In 2021,(Chang et al., 2021) proposed new approach integrates both statistical techniques and deep learning, with a fresh perspective. Initially, traditional statistical methods like decoding and frequency-to-delay analysis are employed. They play a crucial role in enhancing the model's capabilities in speech recognition and speaker identification. Let me draw your attention to an interesting point: considering the stitching of deep neural networks in the E2E SA-ASR model can be seen as an application of deep learning itself. Deep neural networks constitute the core of this model by generating diverse hypotheses through various means, such as using different levels in feature extraction from speech signals all aimed at improving speech recognition while also delving deeper into speaker identification.

7. Conclusion

The following important issues are highlighted in this review, which concentrates on the field of crowd speech identification: Datasets: WSJ, LibriSpeech, and AMI are the three most often used datasets in crowd speech identification. the WSJ dataset highlights the advancements in speech recognition technologies for handling overlapping speech. Various feature extraction methods, notably MFCCs, filter banks, spectrograms, and magnitude spectrum, have been effectively utilized. The predominance of deep learning models such as RNNs, CNNs, and BLSTMs, along with innovative architectures like ADAN, GCN, and PIT, showcases the trend towards leveraging sophisticated algorithms. Significant improvements in Word Error Rate (WER) reduction, with some methods achieving over 60% improvement, reflect the progress in making speech recognition systems more robust in multi-speaker scenarios. The integration of traditional feature extraction methods with deep learning techniques consistently yields the best results, underscoring the importance of combining domain knowledge with advanced algorithms. Overall, the table indicates rapid advancements and effective solutions in addressing the challenges of crowded speech environments. The tables summarizing studies on the LibriMix and AMI datasets reveal notable advancements and methodologies in addressing the challenges of overlapping speech recognition. For the LibriMix dataset, diverse feature extraction methods like MFCCs, filter banks, spectrograms, and raw input are combined with advanced classification techniques, including various neural network architectures such as CNNs, RNNs, and novel models like uPIT-DL and SOT. The results demonstrate substantial improvements in performance metrics, with some methods achieving up to 97% WER reduction, indicating the effectiveness of these approaches in noisy and complex speech environments.

In the case of the AMI dataset, studies primarily focused on feature extraction methods such as MFCC, filter banks, and spectral features, integrated with models like PIT-ASR, CNN-BLSTMs, and the Acoustic Model (AM). These methods have shown significant reductions in WER, with some approaches achieving up to 30.12% improvement. The consistent use of advanced neural network models across both datasets highlights the trend towards leveraging deep learning for enhancing speech recognition accuracy. Overall, the tables reflect the ongoing progress and the importance of combining traditional feature extraction techniques with cutting-edge deep learning models to tackle the complexities of crowded speech scenarios.

Acknowledgement

Our sincere gratitude to all researchers in the field of speech recognition for their tremendous work in this area and for assisting us in finishing this review study

References

Abdulmohsin, H. A. (2022). Automatic Health Speech Prediction System Using Support Vector Machine. *Proceedings of International Conference on Computing and Communication Networks: ICCCN 2021*, 165–175.

Abdulmohsin, H. A., Al-Khateeb, B., Hasan, S. S., & Dwivedi, R. (2022). Automatic illness prediction system through speech. *Computers and Electrical Engineering*, *102*, 108224.

Abdulmohsin, H. A., Stephan, J. J., Al-Khateeb, B., & Hasan, S. S. (2022). Speech Age

Estimation Using a Ranking Convolutional Neural Network. *Proceedings of International Conference on Computing and Communication Networks: ICCCN 2021*, 123–130.

Alsalam, E. A., Razoqi, S. A., & Ahmed, E. F. (2021). Effects of using static methods with contourlet transformation on speech compression. *Iraqi Journal of Science*, *62*(8), 2784–2795. https://doi.org/10.24996/ijs.2021.62.8.31

Andrei, V., Cucu, H., & Burileanu, C. (2017). Detecting Overlapped Speech on Short Timeframes Using Deep Learning. *Interspeech*, 1198–1202.

Arons, B. (1992). A review of the cocktail party effect. *Journal of the American Voice I/O Society*, *12*(7), 35–50.

Bansal, P., Singh, V., & Beg, M. T. (2019). A multi-featured hybrid model for speaker recognition on multi-person speech. *Journal of Electrical Engineering & Technology*, *14*, 2117–2125.

Bronkhorst, A. W. (2015). The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics*, *77*(5), 1465–1487.

Chang, X., Kanda, N., Gaur, Y., Wang, X., Meng, Z., & Yoshioka, T. (2021). Hypothesis stitcher for end-to-end speaker-attributed asr on long-form multi-talker recordings. *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6763–6767.

Chang, X., Qian, Y., & Yu, D. (2018a). Adaptive permutation invariant training with auxiliary information for monaural multi-talker speech recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5974–5978.

Chang, X., Qian, Y., & Yu, D. (2018b). Monaural Multi-Talker Speech Recognition with Attention Mechanism and Gated Convolutional Networks. *INTERSPEECH*, 1586–1590.

Chang, X., Zhang, W., Qian, Y., Le Roux, J., & Watanabe, S. (2019). MIMO-Speech: End-to-end multi-channel multi-speaker speech recognition. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 237–244.

Chen, Z., Li, J., Xiao, X., Yoshioka, T., Wang, H., Wang, Z., & Gong, Y. (2017). Cracking the cocktail party problem by multi-beam deep attractor network. *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 437–444.

Fan, C., Liu, B., Tao, J., Wen, Z., Yi, J., & Bai, Y. (2018). Utterance-level permutation invariant training with discriminative learning for single channel speech separation. *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 26–30.

Gales, M., & Young, S. (2008). The application of hidden Markov models in speech recognition. *Foundations and Trends® in Signal Processing*, *1*(3), 195–304.

Huang, L., Cheng, G., Zhang, P., Yang, Y., Xu, S., & Sun, J. (2019). Utterance-level permutation invariant training with latency-controlled BLSTM for single-channel multi-talker speech separation. *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1256–1261.

Kanda, N., Chang, X., Gaur, Y., Wang, X., Meng, Z., Chen, Z., & Yoshioka, T. (2021). Investigation of end-to-end speaker-attributed ASR for continuous multi-talker recordings. *2021 IEEE Spoken Language Technology Workshop (SLT)*, 809–816.

Kanda, N., Fujita, Y., Horiguchi, S., Ikeshita, R., Nagamatsu, K., & Watanabe, S. (2019). Acoustic modeling for distant multi-talker speech recognition with single-and multi-channel branches. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6630–6634.

Kanda, N., Gaur, Y., Wang, X., Meng, Z., Chen, Z., Zhou, T., & Yoshioka, T. (2020). Joint speaker counting, speech recognition, and speaker identification for overlapped speech of any number of speakers. *ArXiv Preprint ArXiv:2006.10930*.

Kanda, N., Gaur, Y., Wang, X., Meng, Z., & Yoshioka, T. (2020). Serialized output training for end-to-end overlapped speech recognition. *ArXiv Preprint ArXiv:2003.12687*.

Kantowitz, B. H., & Sorkin, R. D. (1983). Human factors: Understanding people-system relationships. *(No Title)*.

Kumar, M. K. P., & Kumaraswamy, R. (2016). Speech separation with EMD as front-end for noise robust co-channel speaker identification. *2016 International Conference on Circuits, Controls, Communications and Computing (I4C)*, 1–4.

Kunešová, M. (2018). *Detection of overlapping speech using a convolutional neural network: first experiments*.

Li, Z., & Whitehill, J. (2021). Compositional embedding models for speaker identification and diarization with simultaneous speech from 2+ speakers. *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7163–7167.

Mao, H. H., Li, S., McAuley, J., & Cottrell, G. (2020). Speech recognition and multi-speaker diarization of long conversations. *ArXiv Preprint ArXiv:2005.08072*.

Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R., & Poria, S. (2023). A review of deep learning techniques for speech processing. *Information Fusion*, 101869.

Meng, L., Kang, J., Cui, M., Wu, H., Wu, X., & Meng, H. (2023). Unified Modeling of Multi-Talker Overlapped Speech Recognition and Diarization with a Sidecar Separator. *ArXiv Preprint ArXiv:2305.16263*.

Menne, T., Sklyar, I., Schlüter, R., & Ney, H. (2019). Analysis of deep clustering as preprocessing for automatic speech recognition of sparsely overlapping speech. *ArXiv Preprint ArXiv:1905.03500*.

Mitchell, O. M. M., Ross, C. A., & Yates, G. H. (1971). Signal processing for a cocktail party effect. *The Journal of the Acoustical Society of America*, *50*(2B), 656–660.

Mohammed, T. S., Aljebory, K. M., Rasheed, M. A. A., Al-Ani, M. S., & Sagheer, A. M. (2021). Analysis of Methods and Techniques Used for Speaker Identification, Recognition, and Verification: A Study on Quarter-Century Research Outcomes. *Iraqi Journal of Science*, *62*(9), 3255–3281. https://doi.org/10.24996/ijs.2021.62.9.38

Mohammed, Z. K., & Abdullah, N. A. Z. (2022). Survey For Arabic Part of Speech Tagging based on Machine Learning. *Iraqi Journal of Science*, *63*(6), 2676–2685. https://doi.org/10.24996/ijs.2022.63.6.33

Qian, Y., Chang, X., & Yu, D. (2018). Single-channel multi-talker speech recognition with permutation invariant training. *Speech Communication*, *104*, 1–11.

Quatieri, T. F., & Danisewicz, R. G. (1990). An approach to co-channel talker interference suppression using a sinusoidal model for speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *38*(1), 56–69.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Sato, H., Ochiai, T., Delcroix, M., Kinoshita, K., Moriya, T., & Kamo, N. (2021). Should we always separate?: Switching between enhanced and observed signals for overlapping speech recognition. *ArXiv Preprint ArXiv:2106.00949*.

Seki, H., Hori, T., Watanabe, S., Roux, J. Le, & Hershey, J. R. (2018). A purely end-to-end system for multi-speaker speech recognition. *ArXiv Preprint ArXiv:1805.05826*.

Settle, S., Le Roux, J., Hori, T., Watanabe, S., & Hershey, J. R. (2018). End-to-end multi-speaker speech recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4819–4823.

Shakat, A., Arif, K. I., Hasan, S., Dawood, Y., & Mohammed, M. A. (2021). YouTube keyword search engine using speech recognition. *Iraqi Journal of Science*, *2021*, 167–173. https://doi.org/10.24996/ijs.2021.SI.1.23

Shangguan, Y., & Yang, J. (2019). Permutation Invariant Training Based Single-Channel Multi-Talker Speech Recognition with Music Background. *2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM)*, 427–430.

Shi, Y., & Hain, T. (2021). Supervised speaker embedding de-mixing in two-speaker environment. *2021 IEEE Spoken Language Technology Workshop (SLT)*, 758–765.

Sun, H., & Ma, B. (2011). Study of overlapped speech detection for NIST SRE summed channel speaker recognition. *Twelfth Annual Conference of the International Speech Communication Association*.

Svendsen, B., & Kadry, S. (2024). *A Dataset for recognition of Norwegian Sign Language*. *2*, 2–4.

Swadi, H. M., & Ali, H. M. (2019). Mobile-based Human Emotion Recognition based on Speech and Heart rate. *Journal of Engineering*, *25*(11), 55–66.

https://doi.org/10.31026/j.eng.2019.11.05

Thakker, M., Vyas, S., Ved, P., & Shanthi Therese, S. (2018). Speaker identification in a multi-speaker environment. *Information and Communication Technology for Sustainable Development: Proceedings of ICT4SD 2016, Volume 2*, 239–244.

Wang, Y., & Sun, W. (2019). Multi-speaker recognition in cocktail party problem. *Communications, Signal Processing, and Systems: Proceedings of the 2017 International Conference on Communications, Signal Processing, and Systems*, 2116–2123.

Wu, B., Yu, M., Chen, L., Xu, Y., Weng, C., Su, D., & Yu, D. (2020). Distortionless multi-channel target speech enhancement for overlapped speech recognition. *ArXiv Preprint ArXiv:2007.01566*.

Yoshioka, T., Erdogan, H., Chen, Z., Xiao, X., & Alleva, F. (2018). Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks. *ArXiv Preprint ArXiv:1810.03655*.

Yousefi, M., & Hansen, J. H. L. (2020). Frame-based overlapping speech detection using convolutional neural networks. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6744–6748.

Zhang, W., Chang, X., & Qian, Y. (2019). Knowledge Distillation for End-to-End Monaural Multi-Talker ASR System. *INTERSPEECH*, 2633–2637.