The First Multimodal Information Based Speech Processing (Misp) Challenge Data, Tasks, Baselines And Results

Chen, Hang ; Zhou, Hengshun; Du, Jun ; Lee, Chin-Hui ; Chen, Jingdong ; Watanabe, Shinji ; Siniscalchi, Sabato Marco ; Scharenborg, Odette; Liu, Di-Yuan ; More Authors

DOI
10.1109/ICASSP43922.2022.9746683

Publication date
2022

Document Version
Final published version

Published in
Proceedings of the ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746683

THE FIRST MULTIMODAL INFORMATION BASED SPEECH PROCESSING (MISP) CHALLENGE: DATA, TASKS, BASELINES AND RESULTS

Hang Chen[1], Hengshun Zhou[1], Jun Du[1,], Chin-Hui Lee[2], Jingdong Chen[6], Shinji Watanabe[3], Sabato Marco Siniscalchi[2,4], Odette Scharenborg[7], Di-Yuan Liu[5], Bao-Cai Yin[5], Jia Pan[5], Jian-Qing Gao[5], Cong Liu[5]*

[1] University of Science and Technology of China, China [2] Georgia Institute of Technology, USA
[3] Carnegie Mellon University, USA [4] Kore University of Enna, Italy [5] iFlytek, China
[6] Northwestern Polytechnical University, China [7] Delft University of Technology, The Netherlands

ABSTRACT

In this paper we discuss the rational of the Multi-model Information based Speech Processing (MISP) Challenge, and provide a detailed description of the data recorded, the two evaluation tasks and the corresponding baselines, followed by a summary of submitted systems and evaluation results. The MISP Challenge aims at tackling speech processing tasks in different scenarios by introducing information about an additional modality (e.g., video, or text), which will hopefully lead to better environmental and speaker robustness in realistic applications. In the first MISP challenge, two benchmark datasets recorded in a real-home TV room with two reproducible open-source baseline systems have been released to promote research in audio-visual wake word spotting (AVWWS) and audio-visual speech recognition (AVSR). To our knowledge, MISP is the first open evaluation challenge to tackle real-world issues of AVWWS and AVSR in the home TV scenario.

Index Terms— MISP challenge, microphone array, audio-visual, automatic speech recognition, wake word spotting

1. INTRODUCTION

With the emergence of many speech-enabled systems, the application scenarios (e.g., home and meetings) are becoming increasingly challenging due to the factors of adverse acoustic environments (far-field audio, background noises, and reverberations) and conversational multi-speaker interactions which typically includes large portions of speech overlap. In the last decade, technology advances in speech enhancement [1, 2, 3, 4] and robust speech processing [5, 6, 7], and the availability of speech corpora recorded in various real environments [8, 9, 10, 11, 12] have caused the performances of many speech-enabled systems to improve tremendously in the above-mentioned scenarios. However, state-of-the-art speech processing techniques based on the single audio modality run into performance plateaus, e.g., the CHiME-6 [13] dinner party scenario reaches a word error rate of about 40%, which is a level of performance that falls short of the deploy ability of the application.

The cocktail party effect [14] indicates that the human auditory system can track a single target voice source in extremely noisy acoustic environment such as a cocktail party. This finding motivates us to design speech-enabled systems by drawing on the way humans perceive speech. The McGurk Effect [15] suggests a strong influence of vision on human speech perception. Other studies [16, 17, 18, 19, 20] have shown that visual cues, such as facial/lip movements, can help speech perception through supplementing speech with visual cues related to the corresponding speaker, especially in noisy environments. Inspired by those findings, speech-enabled systems utilizing both audio and visual signals have been developed.

Various audio-visual speech corpora were released to support research, e.g. TCD-TIMIT [21], LRW [22], LRS2 [23], LRS3[24]and AVSpeech [25]. Nevertheless, there is still a lack of a large-scale public audio-visual speech corpus recorded in real world scenes, especially for the Chinese language.

For the first MISP challenge, we target the home TV scenario, where several people are chatting in Chinese while watching TV and interacting with a smart speaker/TV in a living room. Carefully selected far-field/mid-field/near-field microphone arrays and cameras are used to collect both audio and video data, respectively. Time synchronization for the different microphone arrays and video cameras has been designed for conducting research on multi-modality fusion. Preliminary speech recognition results using only the audio modality show that there is much room for improvement. The challenge considers the problem of distant multi-microphone conversational audio-visual wake-up and audio-visual speech recognition in everyday home environments. The challenge features are:

- Simultaneous recordings from multiple microphone arrays and video cameras;
- Real conversation, i.e., talkers speaking in a relaxed and unscripted fashion;
- High overlaps ratios in multi-talker conversions;
- Real domestic noise backgrounds, e.g., TV, air conditioning, movement, etc.;
- 30+ real room acoustics and 250+ native Chinese, speaking Mandarin without strong accents.

The paper is structured as follows. Section 2 introduces the data collection procedure and tasks. We describe the metric, the data set, the software baseline and the corresponding submitted results for Tasks 1 and 2 in Sections 3 and 4, respectively. We conclude in Section 5. More details can be found on the challenge website[1].

2. DATASET AND TASKS

2.1. Scenario

We consider the following scenario: several people are chatting while watching TV in the living room and they can interact with a

*corresponding author

[1] mispchallenge.github.io

smart speaker/TV. An example recording scene is shown in Fig.1. In the schematic diagram, six speakers are chatting while multiple devices are used to record the audio and video in parallel.

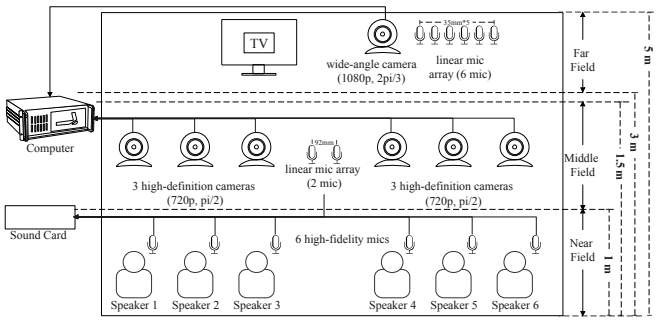

Fig. 1. Schematic overview of the recording scene.

There are some variables that can have an influence on the conversation and/or the collected audio and video that is taking place in the real living room, for example, the TV can be turned on/off, the conversation can happen during the day or night, etc. Moreover, by observing the real conversations taking place in the real living room, we found that speakers could be divided into several groups to discuss different topics. This is a common natural conversation phenomenon. Compared with the situation when all speakers are discussing the same topic, the grouping results in higher overlap ratios in the audio. We control the above variables to cover as many real scenes as possible during recording.

2.2. Audio

Three types of recording devices were used. The type of recording device was dependent on its distance to the speaker. The far recording device is a linear microphone array of 6 sample-synchronised omnidirectional microphones, which is placed 3-5m away from the speaker. The distance between adjacent microphones is 35 mm. The far linear microphone array is recorded onto a laptop computer. At a position 1-1.5 m away from the speaker, we placed a linear microphone array of 2 sample-synchronised omnidirectional microphones. The distance between adjacent microphones is 92mm. To facilitate transcription, each speaker wore a high-fidelity microphone, on the middle of chin. The audio from the middle linear microphone array and each near high-fidelity microphone was recorded via a sound board.

2.3. Video

Two types of recording devices were used to record the video. The type of recording device was dependent on its distance to the speaker. There is a wide-angle camera placed 3-5m away from the speakers. The far wide-angle camera is fixed on the bottom of the far linear microphone array. The x-axis of the camera coordinate system is parallel to the array, and the origin coincides with the midpoint of the array. All participants appear in the camera, which brings speakers position information while reducing the resolution of the lip region of interest (ROI). There is also a high-definition camera placed 1-1.5m away from each speaker. There is only the corresponding speaker in each camera, the lip ROI can be seen clearly and recorded.

2.4. Synchronization

Each device has its own clock, which may each run at their own time/speed. This may result in inconsistency in the clocks [13]. We try to minimize the potential asynchrony in time through the use of synchronization devices and manual post-processing. The clocks of all cameras are synchronized using software, named Vicando, while the clocks of the middle linear microphone array and near high-fidelity microphones are synchronized using the sound card. There are still 3 different clocks, i.e., the clock of the sound card, the clock of the far linear microphone array and the clock of videos. They are synchronized by finding the mark point manually. The mark point is a specific behavior, i.e. a speaker hits a cup with a lid. The video frame where the cup and the lid are in contact and the waveform point which is corresponding to the impact sound is aligned manually.

2.5. Transcription

The conversations have been manually transcribed and segmented at the sentence level. For each speaker, a transcription is constructed in which, for each utterance produced by that speaker, the start and end times and the word sequence are manually obtained by listening to the recording from the near high-fidelity microphone, which is worn in the middle of corresponding speaker's chin. For each other recording device, the utterance's start and end times are consistent with the near recording due to the synchronization between devices (see Section 2.4).

2.6. Tasks

The challenge features two tasks:

1 Audio-Visual Wake Word Spotting: identify a predefined word in a given evaluation utterance and video;

2 Audio-Visual Speech Recognition with Oracle Speaker Diarization: recognize a given evaluation utterance and video while ground truth diarization information is avail.

Task 1 is similar to the Alpha-mini Speech Challenge (ASC) [26] while Task 2 is similar to the "ASR only" track of the CHiME-6 challenge [13], with the key difference being that in our case the audio-only task evolved into the audio-visual task. Task 1 and Task 2 will be introduced in detail in Section 3 and in Section 4, respectively.

3. TASK 1: AUDIO-VISUAL WAKE WORD SPOTTING

3.1. Evaluation

Following the the single-modality challenge in [26], the combination of false reject rate (FRR) and false alarm rate (FAR) is adopted as the criterion, which is defined as follows.

$$Score^{\text{WWS}} = FRR + FAR = \frac{N_{\text{FR}}}{N_{\text{wake}}} + \frac{N_{\text{FA}}}{N_{\text{non-wake}}} \qquad (1)$$

where N_{wake} and $N_{\text{non-wake}}$ denote the number of samples with the wake word and without the wake word in the evaluation set, respectively. N_{FR} denotes the number of samples that include the wake word but where the WWS system erroneously did not detect it and N_{FA} is the number of samples that do not contain the wake word but where the WWS system erroneously detected it. The lower $Score^{\text{WWS}}$, the higher ranking.

3.2. Training, development, and evaluation sets

The database used for task 1 contains 124.79 hours of audio-visual data. Table 1 shows the division of the audio-visual data into a training, development, and evaluation set and indicates details regarding the number of sessions, the type of room, and the number of male/female speakers. The wake word is "Xiao T Xiao T". The data set includes 118 sessions. The number of speakers within one conversation session ranges from 1 to 6. The total number of speakers in the data set is 347. All speakers are native Chinese speaking Mandarin without strong accents. Various conversation topics were recommended during recording. Due to the final ranking only lies on the results of the far recordings, the evaluation set only contains the recordings from the far devices, but the middle and near recordings are avail in the training and development sets. Some real noise data is also provided.

Table 1. Overview of the MISP2021-AVWWS corpus. [P: for presence of wake word, N: for absence of wake word]

Dataset	Training		Dev		Eval	Total
	P	N	P	N		
Duration (h)	5.67	112.86	0.62	2.77	2.87	124.79
Session	89	89	10	10	19	118
Room	25	25	5	5	8	38
Participant	258	258	35	35	54	347
Male	81	81	11	11	31	123
Female	177	177	24	24	23	224

3.3. Baseline and Results

For Task 1, we provide three baselines: 1) audio-only and 2) video-only wake word spotting baseline systems, and 3) data simulation tool which is used to add reverberation and noise to the near mono speech.[2]

3.3.1. Data simulation

Data simulation increases the quantity of training data, which is beneficial to increase the generalization of the system. The Room Impulse Response (RIR) is generated according to the actual room size and microphone position by using an open-source toolkit, i.e. pyroomacoustic[27]. In addition, we provide a simple tool to add noise with 7 different signal-to-noise ratios (from -15dB to 15dB with a step of 5dB).

3.3.2. Audio-only model for wake word spotting

Inspired by the work in [28], we design the proposed audio-only WWS architecture in an end-to-end manner. Due to there is only one wake word in the data set, our model outputs the probability of wake-up. The optimisation objective is a binary cross-entropy loss. We use 40-dimensional filter bank (FBank) features standardized by global mean and variance as the audio features.

3.3.3. Video-only model and audio-visual fusion

The video-only WWS architecture is the same as the audio-only model except that the visual input and embedding module replace the audio input. A visual embedding is extracted from the input image frame sequence using a lipreading model [29], which is pre-trained on a word-level lip reading task and achieves 85.5% classification accuracy on the LRW dataset [22]. For every video frame, the network outputs a compact 512-dimensional feature vector. The original implementation of the lipreading model[3] was adopted to extract the visual feature. For the audio-visual fusion, the scores of the two systems are weighted and added together.

3.3.4. Results

A total of 16 teams from academia and industry participated in Task 1. Table 2 shows $Score^{\text{WWS}}$ results for Task 1. The results of the challenge baseline are quite high which is due to the challenging environments of MISP2021. It also can be observed that the best result of 0.058 is obtained and the performance of many teams outperforms better than the baseline system. Most of the proposed techniques are based on data augmentation and ensemble of networks.

Table 2. $Score^{\text{WWS}}$ of all submissions in the evaluation set on Task 1.

Team ID	$Score^{\text{WWS}}$	Team ID	$Score^{\text{WWS}}$
T01	0.058	T10	0.123
T02	0.071	T11	0.195
T03	0.091	T12	0.241
T04	0.101	T13	0.254
T05	0.108	Baseline	0.322
T06	0.109	T14	0.402
T07	0.110	T15	0.498
T08	0.113	T16	0.898
T09	0.122		

4. TASK 2: AUDIO-VISUAL SPEECH RECOGNITION WITH ORACLE SPEAKER DIARIZATION

4.1. Evaluation

We adopt Character Error Rate (CER) as the metric. It is represented with Eq. 2:

$$CER = \frac{S + D + I}{N} \times 100 \tag{2}$$

where S, D, I and N are the number of substitutions, deletions, insertions and characters in the ground truth, respectively.

The lower the CER value (with 0 being a perfect score), the better the recognition performance. Due to the multi-speaker interaction in our scenario, there are speech segments with multiple speakers talking simultaneously. For such speech overlap segments, we calculate all the S/I/D errors based on the recognition results and the ground truth for each speaker based on the oracle speaker diarization results.

[2] github.com/mispchallenge/misp2021_baseline/tree/master/task1_wws

[3] github.com/mpc001/Lipreading_using_Temporal_Convolutional_Networks

4.2. Training, development, and evaluation sets

Task2 uses the MISP2021-AVSR corpus, which contains 122.53 hours of audio-visual data. The data set includes 376 sessions. Each session consists of a discussion of about 20 minutes. The total number of speakers in the data set is 248. The data set is collected in 30 real living rooms, whose size range from $3.2 \times 2.56 \times 2.54$ to $5.2 \times 4.2 \times 2.8\ m^3$.

The recorded data were split into three subsets for training, development, and evaluation, respectively. There is no overlap in speakers and recording rooms among the data in each subset. For the challenge, all teams are ranked based on the recognition results of far-field data. So only recordings from the far devices are available in the evaluation set.

Table 3. Overview of the MISP2021-AVSR corpus for the audio-visual speech recognition task.

Dataset	Training	Dev	Eval	Total
Duration (h)	101.12	9.83	9.94	120.89
Session	304	37	32	373
Room	20	5	5	30
Participant	200	21	27	248
Male	79	9	7	95
Female	121	12	20	153

4.3. Baseline and Results

For Task 2, we provide baseline systems for speech enhancement, audio-only and audio-visual speech recognition.[4]

4.3.1. Speech enhancement

The baseline multi-channel speech enhancement front-end consists of a weighted prediction error (WPE) dereverberation [30] followed by a weighted delay-and-sum beamformer (BeamformIt [31]), which is similar to the CHiME-6 recipe [13]. Both of the WPE and the BeamformIt can be installed in the Kaldi [32] tool installation directory.

4.3.2. Audio-only speech recognition (ASR)

The conventional ASR baseline consists of the preparation of the dictionary and the language model, the audio feature extraction, the Gaussian mixture model-hidden Markov (GMM-HMM) model training, and the audio-only model training.

We use a DaCiDian[5] dictionary as the basic pronunciation dictionary. A 3-gram language model is trained by the maximum entropy modeling method implemented in the SRILM toolkit [33].

We extract 13-dimensional Mel-frequency cepstral coefficient (MFCC) features for GMM-HMM systems and 40-dimensional high resolution MFCC features for NN-HMM systems.

The GMM stages include standard triphone-based acoustic model building with various feature transformations including linear discriminant analysis, maximum likelihood linear transformation, and feature space maximum likelihood linear regression with speaker adaptive training. These models are used for generating lattices for training the chain model.

We use a factorized time delay neural network (TDNN-F) adapted from the Switchboard recipe 7q model [34]. Speed-perturbation [35] is adopted to increase the quantity of training data.

4.3.3. Audio-visual speech recognition (AVSR)

The AVSR baseline has the same preparation of the dictionary and the language model, the audio feature extraction, the GMM-HMM model training as the audio-only ASR baseline. Generated lattices and fused audio-visual embedding are used to train the audio-visual model. We use the same visual embedding extraction process as explained in Section 3.3.3. Then, high resolution MFCC feature and the visual embedding are concatenated along the channel dimension. The mismatch in the number of frames between audio and video is solved by repeating a video frame for several audio frames. Finally, we use the same model as the ASR baseline, but replace the input with the concatenated features mentioned above.

4.3.4. Results

There are 9 submissions from academia and industry participated in Task 2. Table 4 provides the CERs of all submitted systems and the baseline AVSR system for the evaluation set. The baseline AVSR system achieves a CER of 62.74%, there is indeed still a lot of room of improvement. There are 2 systems achieving a CER of less than 30%. The main performance improvement lies in novel audio-visual fusion methods and various audio data augmentation strategies.

Table 4. CERs of all submissions in the evaluation set on Task 2.

Team ID	CER (in %)	Team ID	CER (in %)
T01	25.07	T06	46.82
T02	27.17	T07	51.53
T03	34.02	T08	60.88
T04	38.87	T09	62.14
T05	42.33	Baseline	62.74

5. SUMMARY & CONCLUSIONS

The MISP2021 challenge focuses on the Audio-Visual Wake Word Spotting and the Audio-Visual Speech Recognition tasks in the Home TV scenario. A set of challenge instructions has been carefully designed to allow meaningful comparison between systems and maximize scientific outcomes. The results of this challenge shows that the visual modality could be a powerful supplement input to improve environmental robustness. In the future, we also further study audio-visual fusion methods and video data augmentation strategies.

6. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China under Grants No. 62171427. The authors were grateful to Yen-Ju Lu for providing an end-to-end AVSR baseline[6], Zhao-Xu Nian and Yu-Sheng Dai for the assistance with the code. Thanks to Ya-Jian Wang, Ya Jiang, Zhe Wang and Shu-Xian Wang for their support on the official website and operations. Thanks to AI Resources Department of iFlyTek for their support on data collection and annotation.

[4] github.com/mispchallenge/misp2021_baseline/tree/master/task2_avsr_nn_hmm

[5] github.com/aishell-foundation/DaCiDian

[6] github.com/espnet/espnet/tree/master/egs2/misp2021

7. REFERENCES

[1] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust mvdr beamforming using time-frequency masks for online/offline asr in noise," in *Proc. ICASSP 2016*, 2016, pp. 5210–5214.

[2] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. ICASSP 2016*, 2016, pp. 196–200.

[3] H. Erdogan, J. R. Hershey, S. Watanabe, et al., "Improved MVDR Beamforming Using Single-Channel Mask Prediction Networks," in *Proc. Interspeech 2016*, 2016, pp. 1981–1985.

[4] L. Chai, J. Du, Q.-F. Liu, et al., "A cross-entropy-guided measure (cegm) for assessing speech recognition performance and optimizing dnn-based speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 106–117, 2021.

[5] T. Virtanen, R. Singh, and B. Raj, *Techniques for noise robustness in automatic speech recognition*, John Wiley & Sons, 2012.

[6] J. Li, L. Deng, R. Häb-Umbach, et al., *Robust Automatic Speech Recognition: A Bridge to Practical Applications*, Elsevier Science, 2015.

[7] S. Watanabe, M. Delcroix, F. Metze, et al., *New Era for Robust Speech Recognition - Exploiting Deep Learning*, Springer, 2017.

[8] E. Vincent, J. Barker, S. Watanabe, et al., "The second 'chime' speech separation and recognition challenge: Datasets, tasks and baselines," in *Proc. ICASSP 2013*, 2013, pp. 126–130.

[9] A. Brutti, L. Cristoforetti, W. Kellermann, et al., "Woz acoustic data collection for interactive tv," *Language Resources and Evaluation*, pp. 205–219, 2010.

[10] S. Renals, T. Hain, and H. Bourlard, "Recognition and understanding of meetings the ami and amida projects," in *Proc. ASRU 2007*, 2007, pp. 238–247.

[11] A. Janin, D. Baron, J. Edwards, et al., "The icsi meeting corpus," in *Proc. ICASSP 2003*, 2003, vol. 1, pp. I–I.

[12] W. Rao, Y.-H. Fu, Y.-X. Hu, et al., "Conferencingspeech challenge: Towards far-field multi-channel speech enhancement for video conferencing," in *Proc. ASRU 2021*, 2021.

[13] S. Watanabe, M. Mandel, J. Barker, et al., "CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings," in *Proc. 6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, 2020, pp. 1–7.

[14] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, pp. 975–979, 1953.

[15] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, pp. 746–748, 1976.

[16] L. E. Bernstein and C. Benoit, "For speech perception by humans or machines, three senses are better than one," in *Proc. ICSLP 1996*, 1996, pp. 1477–1480 vol.3.

[17] A. MacLeod and Q. Summerfield, "Quantifying the contribution of vision to speech perception in noise," *British journal of audiology*, pp. 131–141, 1987.

[18] D. W. Massaro and J. A. Simpson, *Speech perception by ear and eye: A paradigm for psychological inquiry*, Psychology Press, 2014.

[19] L. D. Rosenblum, "Speech perception as a multimodal phenomenon," *Current Directions in Psychological Science*, pp. 405–409, 2008.

[20] H. Chen, J. Du, Y. Hu, et al., "Correlating subword articulation with lip shapes for embedding aware audio-visual speech enhancement," *Neural Netw.*, vol. 143, no. C, pp. 171–182, nov 2021.

[21] N. Harte and E. Gillen, "Tcd-timit: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, pp. 603–615, 2015.

[22] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. ACCV 2016*. Springer, 2016, pp. 87–103.

[23] T. Afouras, J. S. Chung, A. W. Senior, et al., "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[24] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," in *arXiv preprint arXiv:1809.00496*, 2018.

[25] A. Ephrat, I. Mosseri, O. Lang, et al., "Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation.," *ACM Trans. Graph.*, pp. 112:1–112:11, 2018.

[26] Y.-H. Fu, Z.-Y. Yao, W.-P. He, et al., "Ieee slt 2021 alpha-mini speech challenge: Open datasets, tracks, rules and baselines," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 1101–1108.

[27] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. ICASSP 2018*, 2018, pp. 351–355.

[28] L. Momeni, T. Afouras, T. Stafylakis, et al., "Seeing wake words: Audio-visual keyword spotting," in *BMVA*, 2020.

[29] B. Martinez, P.-C. Ma, S. Petridis, et al., "Lipreading using temporal convolutional networks," in *Proc. ICASSP 2020*, 2020, pp. 6319–6323.

[30] L. Drude, J. Heymann, C. Boeddeker, et al., "Nara-wpe: A python package for weighted prediction error dereverberation in numpy and tensorflow for online and offline processing," in *Speech Communication; 13th ITG-Symposium*, 2018, pp. 1–5.

[31] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 2011–2022, 2007.

[32] D. Povey, A. Ghoshal, G. Boulianne, et al., "The kaldi speech recognition toolkit," in *Proc. ASRU 2011*. IEEE Signal Processing Society, 2011.

[33] A. Stolcke, "Srilm–an extensible language modeling toolkit," *Proc. ICSLP 2002*, 2004.

[34] D. Povey, G.-F. Cheng, Y.-M. Wang, et al., "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks," in *Proc. Interspeech 2018*, 2018, pp. 3743–3747.

[35] T. Ko, V. Peddinti, D. Povey, et al., "Audio augmentation for speech recognition," in *Sixteenth annual conference of the international speech communication association*, 2015.