EXPLORING MULTI-CHANNEL FEATURES FOR SPEAKER VERIFICATION WITH JOINT VAD AND SPEECH ENHANCEMENT

Saurabh Kataria\textsuperscript{1*}, Shi-Xiong Zhang\textsuperscript{2}, Dong Yu\textsuperscript{2}

skatari1@jh.edu, auszhang@tencent.com, dyu@tencent.com

\textsuperscript{1}Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA

\textsuperscript{2}Tencent AI Lab, Bellevue, WA, USA

ABSTRACT

To improve multi-channel Speaker Verification, we propose to bring together advancements made in multi-channel speech features. For robust far-field scenarios, we combine so-called spectral, spatial, and directional features. Using a simple feature concatenation scheme, we incrementally introduce new features on top of single-channel features. Experimented features include Inter-channel Phase Difference, multi-channel sinc convolutions, directional power ratio features, and angle features. To maximally leverage supervised learning, our framework is also equipped with multi-channel speech enhancement and Voice Activity Detection. On a challenging language-mismatched test set, we obtain consistent improvements on various degradation levels. On noisy condition, we achieve a 32% relative reduction in Equal Error Rate (EER) w.r.t. single-channel baseline. We find the improvements from speaker-dependent directional features more consistent in noisy condition than clean. This is established via an analysis on several background disturbance levels. Lastly, we investigate if the learned multi-channel speaker embedding space can be made more discriminative through a contrastive loss based fine-tuning. With a simple choice of Triplet loss, we observe a further 8.3% relative reduction in EER.

Index Terms— multi-channel Speaker Verification, joint VAD learning, joint speech enhancement, contrastive learning

1. INTRODUCTION

Speech devices are increasingly getting equipped with multi-channel and multi-modal capabilities, in turn improving spatial ambiguity \cite{1} and directivity \cite{2}. Such information can be fused at various levels: signal, feature, embedding, and score \cite{3}. They are shown beneficial for Automatic Speech Recognition (ASR) \cite{4}, Speaker Recognition \cite{5}, Speech Enhancement \cite{6}, and Source Separation \cite{7}.

Several spectral, spatial, and directional features \cite{2} are proposed for such problems. In \cite{3}, authors use multi-channel sinc convolution filters and show the importance of phase for time-domain multi-channel speech enhancement. Source Separation work of \cite{9} proposed multi-channel Deep Clustering which employs Inter-channel Phase Difference (IPD) and asserts that spatial features are helpful even for arbitrary mic configurations. \cite{6} improved over IPD features through Inter-channel Convolution Difference (ICD) features by directly learning on multi-channel temporal data. In \cite{10}, authors showed the effectiveness of Angle Features (AF) \cite{11} for targeted speech extraction.

\textsuperscript{*}This work was done while Saurabh Kataria was a research intern at Tencent AI Lab, Bellevue, USA.

Multi-Channel Automatic Speaker Verification (MCASV) is an under-explored problem with no common benchmarks except CHiME-5 \cite{12}. It is mostly pursued using pre-processing techniques like dereverberation and mask-based beamforming. In \cite{13}, authors combined Weighted Prediction Error (WPE) dereverberation and Minimum Variance Distortion-less Response (MVDR) beamforming for MCASV. In \cite{14}, authors search for optimal beamformer among variants of Ideal Ratio Mask (IRM) based MVDR and Generalized Eigen-Value (GEV) beamformers. \cite{15} showed the utility of multi-channel speech enhancement for MCASV. \cite{5} found that multi-channel Time-Frequency (T-F) representations are superior to single-channel especially when 3D convolutions are used instead of 2D in a deep multi-channel Convolutional Neural Network (CNN).

We believe prior MCASV works are not comprehensive in terms of leveraging all available information. They also do not investigate if location clue of sound sources can improve directivity. \cite{9} noted that linear beamformer filtering cannot capture multi-channel non-linear information. Towards building a complete multi-channel speaker embedding system, we propose to incorporate various multi-channel features and learn a multi-channel CNN embedding network in conjunction with joint (multi-channel) enhancement and Voice Activity Detection (VAD). Specifically, we pursue text-independent wide-band informed \cite{16} (opposed to blind \cite{17}) MCASV.

Our contributions are: (1) we devise a challenging simulation based location-aware MCASV setup with joint enhancement and VAD learning for a language-mismatched (English-Mandarin) test set; (2) in our knowledge, for the first time, we propose to integrate (multi-channel) spectral, spatial, and directional speech features for robust MCASV, and report consistent improvements; (3) we show that such improvement can be further extended by fine-tuning the multi-channel system with a simple choice of contrastive loss.

2. MULTI-CHANNEL SPEAKER VERIFICATION

2.1. Proposed Multi-Task Supervised Learning Framework

To maximally exploit supervised learning, we devise a multi-task framework which jointly optimizes three multi-channel tasks: speaker embedding learning, Voice Activity Detection, and speech enhancement. In our simulation, we compute and utilize ground truth labels for the latter two tasks. Using the initial single-channel clean speech, we compute its energy VAD and spectrogram as targets for the two tasks respectively. Loss for complete framework is $L_{\text{all}} = L_{\text{BASE}} + \lambda_{\text{enh}} L_{\text{enh}}$, where $L_{\text{BASE}} = L_{\text{emb}} + \lambda_{\text{VAD}} L_{\text{VAD}}$, and $\lambda_{\text{VAD}}$, $\lambda_{\text{enh}}$ are regularization weights for VAD and enhancement task respectively. They are set to 0.1 and 0.0005. $L_{\text{enh}}$ is the Cross Entropy (CE) based speaker classification loss. $L_{\text{VAD}}$ is the frame-wise Binary Cross Entropy (BCE) loss. $L_{\text{emb}}$ is the spectrogram-domain
Mean Absolute Error (MAE) loss. $L_{\text{BASE}}$ is the loss for the BASE model which is a single-channel joint speaker embedding and VAD system without the speech enhancement module.

Fig. [1] illustrates our proposal. Note that input features for all tasks are identical but their outputs are later combined in the pipeline. Our methodology is to incrementally incorporate various multi-channel features on top of the BASE system in order to maximize speaker recognition performance. We do not investigate feature fusion schemes and simply concatenate all features along the channel dimension of our multi-channel CNNs. We hope to discover complementary effect of such features [29] since some of them may fail in adverse scenarios. For e.g., under heavy reverberations, IPDs get degraded. Hence, we are interested in a MCASV system which has good robustness and generalization properties. For this, we choose a test set which is language-mismatched (English-Mandarin) and is corrupt with reverberations, background noises, and speaker interferences. Since there is no MCASV work directly comparable to our setup, for the baseline, we use the standard single-channel 80-D Log-Mel Filter Bank (LMFB) features.

2.2. Large-Margin Contrastive Fine-Tuning

We are inspired by [13], which proposed to switch from CE loss to Triplet loss towards the end of speaker embedding training and showed it generalizes better compared to fixed CE loss training. We investigate if the learned multi-channel speaker embedding space can be made more discriminative using contrastive loss based post-processing. On the pre-trained multi-channel model, we minimize $L_{\text{Triplet}} = f_s(d(a, p) - d(a, n) + m)$, where $f_s(x) = \beta^{-1}\log(1 + \exp(\beta x))$. $L_{\text{Triplet}}$ refers to the contrastive (Triplet) loss, $f_s(\cdot)$ is the sofsoftplus function (smooth version of hinge function [19]), $\beta$ is a non-negative constant, and $d$ is the Euclidean distance function. In the Triplet loss terminology [13], $a$ is an anchor example, $p$ is a positive example, $n$ is a negative example, and $m$ is the margin. $a$ and $p$ belongs to the same class while $n$ belongs to a different class. For triplet formation, we follow the hardest mining strategy [19]. It refers to choosing hardest positive and hardest negative example for each anchor with $d$ as the criterion. For this, training batch must contain multiple examples per speaker. We follow PK sampling by choosing K examples per P unique speakers. We are particularly interested in large margins in this formulation.

3. MULTI-CHANNEL FEATURES

Inter-Channel Phase Difference: IPD is a common spatial feature which measures the phase difference between complex Short-Time Fourier Transform (STFT) of signals at two different microphones.

$$\text{IPD}_{ij} = \angle \left( \frac{Y_{ij}}{Y_{ij}} \right), i, j = 1 \ldots M, i \neq j. \tag{1}$$

Here, $M$ is the number of microphones in the array, $i, j$ are microphone indices, $(t, f)$ is the current T-F bin, and $Y$ is the STFT. Note that this gives us $M(M-1)$ pairs but we later choose only a pre-defined small subset. Cosine and sine of IPD are extensively used in prior works [6][7][22][29] and they are referred to as cosIPD and sinIPD.

Directional Power Ratio (DPR): We adopt directional features from [29] based on the output power of multi-look fixed beamformers and direction beam of target speaker $\theta_p$. For a set of direction grid of beams $\{\theta_1, \ldots, \theta_P\}$ and the corresponding filters $w_f(\theta_p)$, we can compute the directional energy $\text{DPR}_{ij} = \frac{\sum_{p=1}^{P} ||w_f(\theta_p)Y_{ij}||_2^2}{\sum_{p=1}^{P} ||w_f(\theta_p)Y_{ij}||_2^2}$.

Angle Features: We adopt the Angle Features (AF) from [1]. The AF is formed according to the angle $\theta$ of a target speaker and measures the cosine distance between the steering vector and IPD:

$$A_{ij}(\theta) = \sum_{(i,j)} \langle e^{\text{IPD}_{ij}}, t, f \rangle,$$

$$\text{TPD}_{ij}(t, f) = 2\Delta_{ij} \cos(\theta(t)) f / (c \cdot f_s) \tag{3}$$

Vector $e^{\cdot}$ is the mixed multi-channel complex spectrum. The DPR features quantify how well a T-F bin is represented by a signal from a direction beam $\theta_p$.

where $Y$ is the mixed multi-channel complex spectrum. The DPR features quantify how well a T-F bin is represented by a signal from a direction beam $\theta_p$.

4. EXPERIMENTS

We use a proprietary 15-channel non-uniform linear microphone array setup with spacing 7-6-5-4-3-2-1-1-2-3-4-5-6-7 (in cm). The real recorded multi-channel audio visual data will be open released soon [23]. For this paper, all data (training, validation, evaluation) is originally single-channel 16KHz and is simulated to 15-channel on-the-fly. The details of the simulation can be found in the Alg. 1 of [1]. To construct training data, we first combine a portion of VoxCeleb [25] (450 hrs, 7285 speakers) and small amount of internal Mandarin data (50 hrs, 2265 speakers). This gives us approximately 500 hrs and 9550 speakers which is then split in 9:1 for training and validation data. Initial evaluation data consists of clean internal Mandarin corpus (36 hrs, 364 speakers). For data augmentation, we use a large internal noise and Room Impulse Response (RIR) corpora which contains 700 noise files and 3000 RIRs. The RIRs are multi-channel signals simulated in various virtual room configurations. They are also similarly split to create disjoint training, validation and evaluation data. We constrain all sound sources (target speaker, interference speaker, noise sources) to azimuth angle range of $[0^\circ, 180^\circ]$. For DPR features, we choose spatial resolution of 10°and hence $P = 10$ in Equation [2]. Ignoring spatial ambiguity issue [26], we keep the angle between target and interference speaker unconstrained. Position of all sources are static. Only one interference speaker is allowed with probability $p_{\text{uw}}$. For training and
In Table 1 we evaluate three versions of our multi-channel test set: clean (no data augmentation), noisy, and noisy with SIR fixed to 0. The last version explores a slightly more challenging scenario. We observe that 257-D Log Power Spectrum (LPS) features are better than 80-D LMFBS perhaps due to higher dimensionality. In clean condition, MultChanSinc features are better than single-channel features, but perform worse in noisy condition. This suggests that these features might not be robust to noise or they are sensitive to its hyper-parameters. LPS+cosIPD are conventional choice of features and are superior to MultChanSinc in our case. By experimenting with three choices of microphone pairs, we observe that this choice is important and perhaps can be even learned. Performance-wise ranking of such pairs is v2 > v0 > v1. They also achieve best clean condition performance. Adding beamformer features (DPR) hurt in clean condition while maintaining performance in noisy condition. This is possibly because these features fail to complement other features in our simple feature concatenation scheme. We make similar observation for angle features (AF). When both type of directional features are combined, complementary behavior is observed and we get best results on noisy condition while performance on clean condition is also better than the baseline.

Incorporating sinIPD features vastly degrade performance. This ill-effect is perhaps also due to concatenating vastly different types of features on channel dimension and not learning to fuse them. However, adding interference location information gives some expected improvement. This demonstrates a cancelling ability in our directional features. Including absolute phase information of the first microphone phase0 hurts, which shows that relative phase information (through IPDs) is sufficient. Note that, till now, enhancement is not incorporated. Using it further improves performance. However, excluding VAD from this system degrades minDCF while improving EER. This suggests that there is an interplay between enhancement and VAD module and therefore a further exploration is required.

An critical observation is that progressively incorporating more information (our methodology) leads to consistent improvement in validation accuracy (better speaker re-identification) but generalization (measured by speaker verification task) patterns vary significantly. Finally, we note that the combination LPS + cosIPDv0 + DPR(1) + AF(1) has drastically bridged the gap between performance on noisy condition (16.9%) and clean baseline (13.9%).

5.2. Multi-channel system evaluation on various SNR, SIR pairs
In Table 2 we evaluate few key feature combinations on specific (SNR,SIR) pairs. We note that the trend of consistent improvement with addition of features hold true for challenging conditions (low
SNR, low SIR) only. On cleaner conditions, spatial features help but directional features degrade performance. This explains some unexpected observations in Table 1 since we report averaged performance there. This strongly suggests the sub-optimality of our feature concatenation scheme for directional features. We also observe that the performance of multi-channel features is always better than single-channel and only comes close (16.6% vs 16.7%) in high SNR, high SIR condition. Combining spectral, spatial, and directional features in low SNR=-2 dB gives almost identical performance for both SIR=0 and SIR=6. This demonstrates the strong ability of our system to handle interference speaker.

### Table 2: EER (in %) performance for key feature combinations on various pairs of SNRs and SIRs (in dB)

| SIR=0 | BASE features ↓ \ SNR → | CLEAN | NOISY (all conds) | NOISY (SIR=6) |
|-------|--------------------------|-------|-------------------|---------------|
|       |                          | EER (%) | minDCF | EER (%) | minDCF | EER (%) | minDCF |
| FRANK | 32.0                     | 24.6    | 20.6   |          |        |          |        |
|       | LPS                      | 27.4    | 26.1   | 19.9     | 16.1   |          |        |
|       | LPS + cosIPDv0           | 26.1    | 19.9   | 16.1     |        |          |        |
|       | LPS + cosIPDv0 + AF(1)   | 23.4    | 19.2   | 16.9     |        |          |        |
|       | LPS + cosIPDv0 + DPR(2) + AF(2) + sinIPDv0 | 23.5 | 18.8 | 17.0 | | |
|       | LPS + cosIPDv0 + AF(2) + sinIPDv0 + enh - VAD | 22.1 | 18.7 | 17.0 | | |

| SIR=6 | BASE features ↓ \ SNR → | CLEAN | NOISY (all conds) | NOISY (SIR=6) |
|-------|--------------------------|-------|-------------------|---------------|
|       |                          | EER (%) | minDCF | EER (%) | minDCF | EER (%) | minDCF |
| FRANK | 30.8                     | 22.8   | 18.1   |          |        |          |        |
|       | LPS                      | 26.2    | 20.1   | 16.7     |        |          |        |
|       | LPS + cosIPDv0           | 24.9    | 18.4   | 14.5     |        |          |        |
|       | LPS + cosIPDv0 + AF(1)   | 23.1    | 18.6   | 16.3     |        |          |        |
|       | LPS + cosIPDv0 + DPR(2) + AF(2) + sinIPDv0 | 23.0 | 18.4 | 16.3 | | |
|       | LPS + cosIPDv0 + AF(2) + sinIPDv0 + enh - VAD | 22.1 | 18.4 | 16.6 | | |

5.3. Contrastive fine-tuning of trained multi-channel system

We investigate if our best multi-channel system of Table 1 (LPS + cosIPDv0 + DPR(1) + AF(1)) can be fine-tuned to improve verification performance by a simple choice of contrastive loss: Triplet Loss. By experimenting with various values of margin $m$, we obtain large improvements in clean as well as noisy condition. This suggests that softmax training leaves some scope of improvement in embedding space. $m = 2$ gives the best results contrary to small values like 0.1 and 0.2 used in previous works [18, 27]. For noisy test set, compared to single-channel baseline of Table 1 we reduce EER from 24.9% to 15.5% here. Hence, we demonstrate that MCASV performance can be improved via better design of input feature space as well as target embedding space.

### Table 3: Fine-tuning the multi-channel system (LPS + cosIPDv0 + DPR(1) + AF(1)) with different Triplet margin $m$

|          | CLEAN | NOISY (all conds) |
|----------|-------|-------------------|
|          | EER (%) | minDCF | EER (%) | minDCF |
| no fine-tuning | 12.0 | 2.06 | 16.9 | 0.818 |
| $m = 0.3$ | 11.0 | 1.07 | 16.8 | 0.813 |
| $m = 1$   | 10.2 | 0.69 | 16.4 | 0.834 |
| $m = 2$   | 10.5 | 0.68 | 15.5 | 0.780 |
| $m = 3$   | 11.3 | 0.69 | 16.4 | 0.827 |

6. CONCLUSION

To advance robust location-aware multi-channel Speaker Verification, we tackle the problem of designing better input feature as well as embedding space. For the former task, we explored various combinations of spectral, spatial, and directional features to find that single-channel baseline can be vastly improved under all testing conditions with combinations of spectral and spatial features. We find the benefit of adding directional features more prominent in noisy conditions while an analysis showed that further exploration is required to improve our simple feature concatenation scheme. Overall, we observe 32% relative reduction in Equal Error Rate (EER) on noisy condition. We also show that the discriminativeness of speaker embedding space can be significantly improved via a contrastive loss based fine-tuning of our multi-channel system. In future, we can investigate (1) feature fusion learning schemes; (2) robustness to sound source(s) location information (modality robustness problem (1)); and (3) explore more multi-channel features like Inter-channel Convolution Differences (ICD) [6].
7. REFERENCES

[1] Rongzhi Gu, Shi-Xiong Zhang, Yong Xu, Lianwu Chen, Yuxian Zou, and Dong Yu, “Multi-modal multi-channel target speech separation,” IEEE Journal of Selected Topics in Signal Processing, 2020.

[2] Zhuo Chen, Xiong Xiao, Takuya Yoshioka, Hakan Erdogan, Jinyu Li, and Yifan Gong, “Multi-channel overlapped speech recognition with location guided speech extraction network,” in 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018, pp. 558–565.

[3] LADISLAV MOŠNER, “Far-field speaker verification incorporating multichannel processing,” BRNO UNIVERSITY OF TECHNOLOGY thesis, 2020.

[4] Tara N Sainath, Ron J Weiss, Kevin W Wilson, Bo Li, Arun Narayanan, Ehsan Variani, Michel Bacchiani, Izhak Shafran, Andrew Senior, Kea Chin, et al., “Multichannel signal processing with deep neural networks for automatic speech recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 5, pp. 965–979, 2017.

[5] Danwei Cai, Xiaoyi Qin, and Ming Li, “Multi-channel training for end-to-end speaker recognition under reverberant and noisy environment,” in INTERSPEECH, 2019, pp. 4365–4369.

[6] Rongzhi Gu, Shi-Xiong Zhang, Lianwu Chen, Yong Xu, Meng Yu, Dan Su, Yuxian Zou, and Dong Yu, “Enhancing end-to-end multi-channel speech separation via spatial feature learning,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 7319–7323.

[7] Zhong-Qiu Wang and DeLiang Wang, “Combining spectral and spatial features for deep learning based blind speaker separation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 2, pp. 457–468, 2018.

[8] Chang-Le Liu, Sze-Wei Fu, You-Jin Li, Jen-Wei Huang, Hsin-Min Wang, and Yu Tsao, “Multichannel speech enhancement by raw waveform-mapping using fully convolutional networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 1888–1900, 2020.

[9] Zhong-Qiu Wang, Jonathan Le Roux, and John R Hershey, “Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent deep speaker separation,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 1–5.

[10] Aswin Shanmugam Subramanian, Chao Weng, Meng Yu, Shi-Xiong Zhang, Yong Xu, Shinji Watanabe, and Dong Yu, “Far-field location guided target speech extraction using end-to-end speech recognition objectives,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 7299–7303.

[11] Zhong-Qiu Wang and DeLiang Wang, “On spatial features for supervised speech separation and its application to beamforming and robust asr,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 5709–5713.

[12] Daniel Garcia-Romero, David Snyder, Shinji Watanabe, Gregory Sell, Alan McCree, Daniel Povey, and Sanjeev Khudanpur, “Speaker recognition benchmark using the chime-5 corpus,” in INTERSPEECH, 2019, pp. 1506–1510.

[13] Joon-Young Yang and Joon-Hyuk Chang, “Joint optimization of neural acoustic beamforming and dereverberation with x-vectors for robust speaker verification.” in INTERSPEECH, 2019, pp. 4075–4079.

[14] Hassan Tahirian, Z-Q Wang, and DeLiang Wang, “Deep learning based multi-channel speaker recognition in noisy and reverberant environments,” in Interspeech, 2019.

[15] Hassan Tahirian, Zhong-Qiu Wang, Jorge Chang, and DeLiang Wang, “Robust speaker recognition based on single-channel and multi-channel speech enhancement,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 1293–1302, 2020.

[16] Marc Delcroix, Katerina Zmolkova, Keisuke Kinoshita, Atsunori Ogawa, and Tomohiro Nakatani, “Single channel target speaker extraction and recognition with speaker beam,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 5554–5558.

[17] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen, “Permutation invariant training of deep models for speaker-independent multi-talker speech separation,” in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 241–245.

[18] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu, “Deep speaker: an end-to-end neural speaker embedding system,” arXiv preprint arXiv:1705.02304, vol. 650, 2017.

[19] Alexander Hermans, Lucas Beyer, and Bastian Leibe, “In defense of the triplet loss for person re-identification,” arXiv preprint arXiv:1703.07737, 2017.

[20] Rongzhi Gu, Lianwu Chen, Shi-Xiong Zhang, Jincheng Zheng, Yong Xu, Meng Yu, Dan Su, Yuxian Zou, and Dong Yu, “Neural spatial filter: Target speaker speech separation assisted with directional information,” in Proc. Interspeech, 2019.

[21] Davis E. King, “Dlib-ml: A machine learning toolkit,” Journal of Machine Learning Research, vol. 10, pp. 1755–1758, 2009.

[22] Aswin Shanmugam Subramanian et. al., “Directional ASR: A new paradigm for E2E multi-speaker speech recognition with source localization,” in ICASSP. IEEE, submitted.

[23] Shi-Xiong Zhang, Yong Xu, Meng Yu, Lianwu Chen, Jianwei Yu, Mingjie Jin, Dan Su, and Dong Yu, “M*: Multi-Modal Multi-channel dataset for cocktail party problems,” to be released, 2020.

[24] Mirco Ravanelli and Yoshua Bengio, “Speaker recognition from raw waveform with sincnet,” in 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018, pp. 1021–1028.

[25] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman, “Voxceleb: Large-scale speaker verification in the wild,” Computer Speech & Language, vol. 60, pp. 101027, 2020.

[26] Youngmoon Jung, Yejun Choi, and Hoin Kim, “Self-adaptive soft voice activity detection using deep neural networks for robust speaker verification,” in 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2019, pp. 365–372.

[27] Hervé Bredin, “Tri스트론트: triplet loss for speaker turn embedding,” in 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2017, pp. 5430–5434.
This figure "multchan.png" is available in "png" format from:

http://arxiv.org/ps/2010.12692v1