Identify Speakers in Cocktail Parties with End-to-End Attention

Junzhe Zhu, Mark Hasegawa-Johnson, and Leda Sarı

ECE Department, University of Illinois, USA

{junzhez2,jhasegaw,lsari2}@illinois.edu

Abstract

In scenarios where multiple speakers talk at the same time, it is important to be able to identify the talkers accurately. This paper presents an end-to-end system that integrates speech source extraction and speaker identification, and proposes a new way to jointly optimize these two parts by max-pooling the speaker predictions along the channel dimension. Residual attention permits us to learn spectrogram masks that are optimized for the purpose of speaker identification, while residual forward connections permit dilated convolution with a sufficiently large context window to guarantee correct streaming across syllable boundaries. End-to-end training results in a system that recognizes one speaker in a two-speaker broadcast speech mixture with 99.9% accuracy and both speakers with 93.9% accuracy, and that recognizes all speakers in three-speaker scenarios with 81.2% accuracy.[1]

Index Terms: speaker recognition, source separation, cocktail party effect

1. Introduction

The "Cocktail Party" is the problem of recognizing the identity of multiple speakers who are talking at the same time; in order to distinguish it from other types of simultaneous-speech problems, this problem is also called co-channel speaker identification [1]. Different methods have been proposed for scenarios where there are multiple microphone channels, or when there are visual inputs [2], but identification of multiple speakers from a monaural signal is challenging, because the same band now contains sounds from different sources [3]. Full-band speaker identification technologies, like Gaussian mixture models (GMM) [4] and i-vectors [5], fail. Early approaches to co-channel speaker identification simply tried to identify temporal segments sufficiently dominated by one speaker, to permit single-speaker identification [6], but since single-channel source separation became possible [7], its methods were applied to the problem of co-channel speaker ID [8]. Automatic speaker identification systems based on GMMs were applied using the fusion of speaker-dependent models [9], using the inter-GMM Kullback-Leibler divergence [10], using vector quantization [11], and by integrating the GMMs in a graphical model [12].

In a wide variety of applications recently, it has been demonstrated that end-to-end training of a deep neural network can result in lower error rates than a system built up from separately optimized components [13, 14]. End-to-end co-channel speaker identification is also possible: the speaker identification loss can be back-propagated to train the weights of the source separation algorithm. Multi-speaker recognition is, obviously, a type of multi-label classification problem [15], and it is possible to train a co-channel speaker ID system using multi-label classification methods [16]. Improved accuracy is possible, however, if one can force the network to output the correct number of speakers, rather than allowing the network to choose the number of speakers. Doing so requires solving the streaming problem (the problem of correctly assigning each decoded feature vector to a feature stream) and the permutation problem (the problem of computing gradients based on the correct assignment of feature streams to ground truth speaker identities). The streaming problem can be solved using stream-assignment methods from computational auditory scene analysis [17], or using dilated convolution [18]. The permutation problem can be solved using a dominant-speaker model [19], but is usually solved using some variant of permutation-invariant training (PIT) [20, 17], in which feature streams are assigned to ground truth reference signals in whatever order minimizes the reconstruction error.

This paper develops a new algorithm for co-channel speaker identification, based on methods of residual attention, dilated convolution with residual connections, Siamese networks for speaker identification, and a novel max-pooling error metric. Residual attention computes a spectrogram transformation and spectrogram mask that are simultaneously optimized [21]. Residual connections permit us to implement a much deeper dilated convolution than the one proposed in [18] (18 layers compared to their 3 layers), thereby permitting a more robust streaming of the segregated features. Siamese networks are used to pre-train the speaker recognition network [22, 23]. The entire end-to-end network is then trained for optimum co-channel speaker identification using a max-pooling error criterion that ignores permutation, in a manner similar to PIT [20], but without PIT's computational expense.

2. Source extractor model

The first stage of our pipeline is an extractor that uses residual attention and dilated convolution to extract and stream individual source spectrograms from the mixed speech spectrogram.

The scaling of speech spectrograms is problematic. The magnitude STFT requires different network sensitivity at different frequencies: amplitudes of low-frequency pitch lines are dozens of times higher than those of high-frequency pitch lines. The decibel scale is unsatisfactory: logarithms of small numbers are large negative numbers. Standard spectrogram visualizations (e.g. [24]) truncate the spectrogram 60dB below its maximum, but the statistics of a truncated input are difficult for a neural network to learn. A useful balance is obtained using $X = \log(1 + S)$ as the network input, where $S = |STFT|$ is the magnitude of the short-time Fourier transform for mixed and separated speech signals.

The task of separating a mixed spectrogram boils down to assigning pitch lines, frication frames, and aspiration frames correctly to each source, then refining the time-frequency bins that are under interference. Therefore, we hypothesize that in order to separate the sources as closely as possible, the neural

---

[1]Code: https://github.com/JunzheJosephZhu/Identify-Speakers-in-Cocktail-Parties-with-E2E-Attention.git
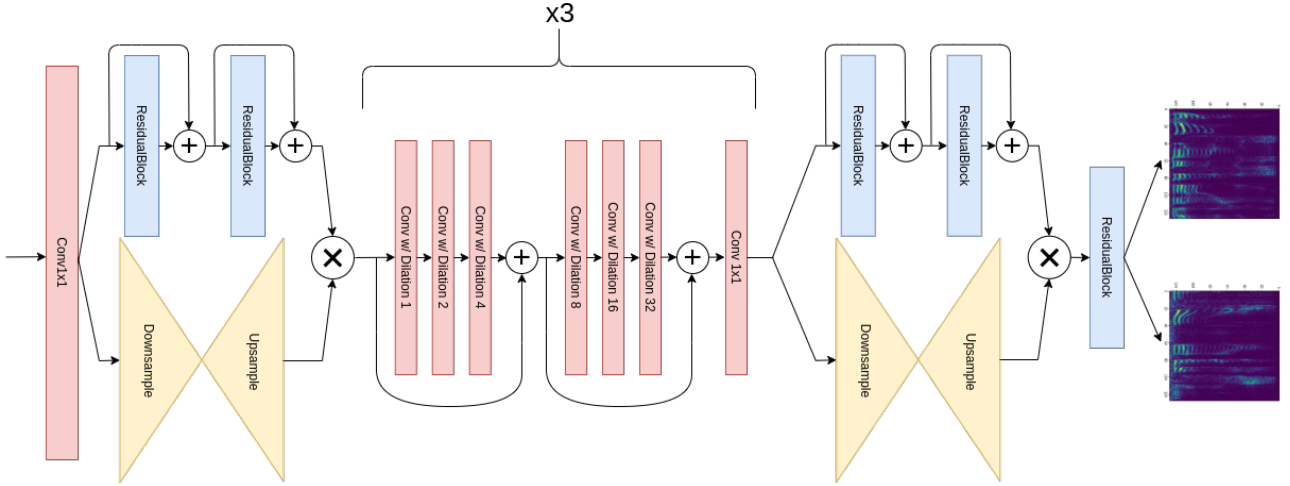
Figure 1: *Extractor module for 2-speaker model. The residual blocks are the 'trunk' branch, while the hourglass architecture is the 'mask' branch. All convolutions are 2D.*

network needs to first have some higher-level knowledge about each source, in order to correctly identify features from different sources. For this purpose we employ residual attention [25] as the main backbone of the model; its hourglass architecture collects contextual information and generates masks that are able to suppress fine-grained regions of the spectrogram.

The extractor model begins and ends with residual attention networks (the dual-path sections at left and right sides of Fig. 1). Residual attention computes two different types of nonlinear tensor transformations: a trunk branch, $T(X)$, and a set of attention masks, $M(X)$. The trunk branch ($T(X)$) is a sequence of convolutional layers with residual connections, shown in the top path of Figure 1. The masks ($M(X)$) are computed in a series of downsample-then-upsample operations, arranged in a U-Net [26] architecture, shown in the bottom path of Figure 1. The U-Net architecture learns global structure (e.g., harmonic structure, and the alternation of voiced and unvoiced speech frames) with which to estimate $M(X)$. The residual attention layers compute $(1+M(X)) \odot T(X)$, an initial estimate of the separated speech sources, where the plus term represents a residual connection for better gradient propagation. Several residual attention architectures were tested; the best-performing system uses residual attention blocks with 128 channels and 8x downsample-upsample ratio in the mask branch.

The residual attention block ensures consistency of the harmonic structure in each continuously-voiced interval, but does not ensure consistency of speaker identity between voiced regions. A convolution kernel composed mostly of zeros, with nonzero values only once every M samples, is called a dilated convolution [27]. It has been demonstrated that dilated convolutional layers, with dilation factors that increase exponentially from layer to layer, can accumulate long-range context in order to improve image segmentation [28] and music source separation [29], as well as speech denoising [30]. Previous work demonstrated the benefit of short-range dilated convolution for co-channel speaker identification [18], but was unable to leverage dilated convolution effectively for the streaming of consecutive syllables, because vanishing gradients limited the number of dilated convolutional layers that could be effectively trained. Better gradient propagation, and a much deeper stack of dilated convolutions, are possible by adding a residual connection every

3 convolutional layers in the dilated convolutional network; besides eliminating the vanishing gradient problem, residual connections help the network to maintain the focus on local time-frequency magnitude information while aggregating context.
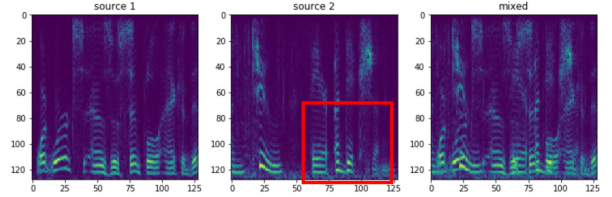

Figure 2: *Example spectrograms of target source 1, target source 2, and the mixed signals.*
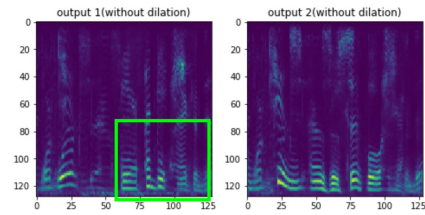

Figure 3: *Example spectrograms: prediction without dilated convolution. It is obvious that left & right halves of both output channels correspond to different sources.*

Figure 3 shows the result of extraction without using dilation with Figure 2 as input. It can be seen that the left and right parts of the extracted spectrogram come from mismatched sources. Figure 4 shows the result of extraction with 5 blocks of dilated convolution, and demonstrates that the mistake in Figure 3 has been corrected, because of the increased receptive field provided by the dilated convolutions. Several architectures were tested; the best-performing system uses 3 dilated convolution blocks with $32 \times$ #talker channels, where #talker is the number of talkers in each mixture. each with 6 dilated convolution layers, with the nth layer having dilation of 2^{n-1} pixels.
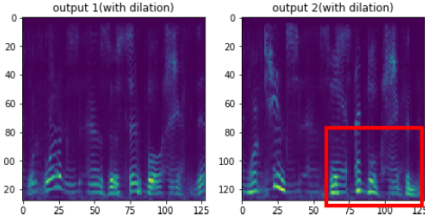
Figure 4: *Example spectrograms: prediction with 5 blocks of dilated convolution, from the ablation study.*

Residual attention and dilated convolution are pre-trained, prior to training the speaker identification block, using permutation-invariant [20] mean squared error (MSE) of the reconstructed source spectrograms,

$$MSE(\hat{S}, S) = \min_{\substack{\sigma \in \text{permutations} \\ (1:n_{\text{sources}})}} \sum_{i=1}^{n_{\text{sources}}} \|S_i - \hat{S}_{\sigma(i)}\|_2^2 \quad (1)$$

where $\hat{S}_c$ is the network output in channel c, and S_i is the ground truth spectrogram of the ith source.

3. Siamese speaker classification models

Each output channel of the extractor model corresponds to the estimated spectrogram for a single source. Each channel is classified using a Siamese (shared weights) speaker detection model [22], in order to generate a prediction vector. The output of the Siamese networks is a matrix of speaker identification probabilities, $\hat{y}_{i,c} = \text{Pr}\{\text{channel } c \text{ is speaker } i\}$, with weights shared across the channel index, c. Two different architectures were tested for the Siamese networks: syllable-level LSTM, and ResNet34.

The syllable-level LSTM is computed by chopping the extracted spectrogram into chunks of 312.5ms, which is about the length of a syllable, with a shift of 15.6ms between each chunk (Fig. 5). For each chunk, a 3-layer bi-LSTM with a hidden layer size of 200 computes features, which are concatenated into a single vector per chunk. Each chunk is classified using a 3-layer fully connected network (hidden layer size of 800, output size equal to the number of speakers), whose softmax outputs are averaged over the whole signal. By limiting window size to 312.5ms and averaging over the signal, we reduce the chance of extraction error affecting the prediction result by preventing it from changing the hidden state of the LSTM over a long range.
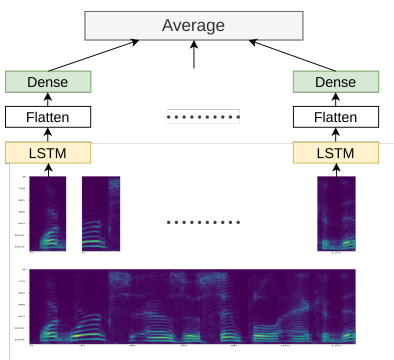

Figure 5: *Syllable-level LSTM*

The ResNet34 classifier is a direct implementation of Resnet34 [31], with softmax activation at outputs. We found that for the global pooling layer, average pooling works significantly better than max pooling, probably due to the suppression of misassigned phonemes in the averaging operation.

In order to make the speaker identification network robust to output noise in the extracted spectrograms, the network is trained directly using extracted spectrograms as inputs. However, if we directly use the permutation decided by PIT [20] as input-target pairs, the accuracy of speaker classification is upper-bounded by the accuracy with which PIT is able to match extractor channels to reference sources. This is problematic especially in cases with more than 2 speakers, because PIT uses a minimum-MSE criterion to match output-reference pairs, but MSE does not directly measure speaker identification accuracy. Incorrect matching caused by PIT can result in the speaker classification network being trained on the wrong input-target pair, which will harm the accuracy of the system.

To address this issue, we propose channel-wise max pooling as a new way of aggregating output for such systems. To produce the final prediction vector for a mixed spectrogram, for each prediction index, we select its value from the Siamese network that yields the highest value. This is equivalent to stacking the Siamese network outputs in the channel dimension and performing channel-dimension max pooling. We then minimize the categorical cross entropy (CCE),

$$CCE(\hat{y}, y) = \sum_{i=1}^{n_{\text{speakers}}} -y_i \log\left(\max_{c=1}^{n_{\text{channels}}} \hat{y}_{i,c}\right) \quad (2)$$

between the predicted vector and ground truth, where $y_i = 1$ when speaker i is among the target speakers, and n_{speakers} is the number of speakers known to the system. By letting the neural network vote for the match between ground truth speakers and predicted speakers, using the max-pooling framework shown in Eq. (2), we bypass the upper limit in the accuracy of PIT matching.

4. Joint training of extractor and classifier with end-to-end back-propagation
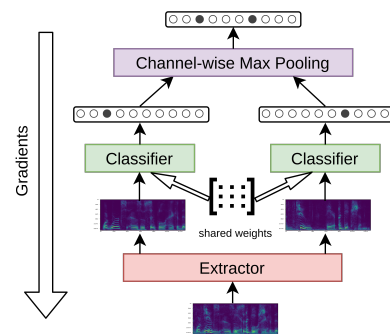

Figure 6: *Joint training of system*

Training the extractor model to minimize MSE is sensible if the training data contain clean, noise-free recordings of each target speaker. Target speakers recorded from Broadcast News data, however, may be corrupted by music or other background noise signals, which are irrelevant to the task of speaker identification; an extractor trained in order to minimize MSE in the

reconstruction of such sources will learn to waste parameters on signal components that are hard to estimate, and irrelevant to the task of speaker identification. To address this issue, we use joint training as shown in Fig. 6 to further maximize the accuracy by pushing the network to place emphasis on features that help speaker classification.

Since speaker classification loss is differentiable w.r.t. the separated spectrograms, its gradients can be back-propagated to the extractor model. The extractor model is then trained using a two-part loss function, including CCE of the classifier, and scaled MSE of the extracted spectrograms:

$$\mathcal{L} = \alpha \mathrm{MSE}(\hat{S}, S) + \mathrm{CCE}(\hat{y}, y) \qquad (3)$$

We set $\alpha = 20$ in 2 speaker case, and $\alpha = 300$ in 3 speaker case, to level $\alpha \mathrm{MSE}(\hat{S}, S)$ and $\mathrm{CCE}(\hat{y}, y)$ to similar scales.

5. Experimental methods and results

Training, development, and test data were extracted from Hub-4 broadcast news recordings [32]. We select top 20 speakers with the longest speech time, and for each person, extract all the non-overlapping 2s segments. All segments on average have 6.2% silence time at -30dB threshold. We split those segments into 3 groups of 38424 training segments, 4803 validation segments, and 4886 test segments. During training/testing time, for each source mixture, we deterministically select the first source by iterating the dataset, select the other sources randomly from the remaining speakers, and add the signals to produce a mixture. 128-dimensional magnitude spectrum of each mixture audio is computed with window size of 32ms and hop size of 16ms.

Two versions of the proposed end-to-end system were tested. Both versions use the same extractor model (Sec. 2), but different classifiers (Sec. 3): one of the classifiers is a syllable-level LSTM, one is ResNet34. The extractor is pre-trained using MSE, and the classifier using CCE, then the extractor and classifier are pipelined, and trained using end-to-end back-propagation of the joint loss (Eq. (3)).

Two baselines were implemented. First, among previously published co-channel speaker identification systems, the one most comparable to ours is the system of Wang2018 [18]: that system uses three layers of dilated convolution, but no residual connections or residual attention, and directly estimates speakers from mixtures without reconstructing sources. Experimental tests reported in [18] demonstrated its success in separating laboratory recordings with predictable lexical content, but it has not been tested on Broadcast News recordings. Second, we implemented a 2-stage modular baseline that we built ourselves with state-of-the-art networks. Conv-Tasnet [33] is used as extractor while SincNet is used as speaker classifier. Conv-Tasnet is first trained on the same training mixtures using PIT loss. Extracted signals are then paired with ground truth input sources using PIT in order to match each extracted signal with a target speaker, and SincNet [34] is trained using the matched pairs. Training uses non-overlapping 0.2s chunks that cover the whole 2s signal as training inputs: SincNet is applied to each chunk, and its softmax outputs are averaged over time. Testing uses 0.2s chunks with 10ms shift, and time averaging of SincNet softmax outputs. Joint training is not used.

We also did an ablation study with the ResNet34 version of our pipeline. We replaced our residual attention blocks with dilated convolutional blocks, with the same number of channels, to see how much residual attention increases the accuracy.

Table 1 lists experimental results. The columns headed by M/N list the percent of test data where at least M speakers out

Table 1: *Results; column M/N lists % of test data where at least M talkers out of N-talker mixture are correctly identified.*

(# spkr corr)/(# in mixture)			(s)	(10^9)	(10^6)
Algorithm	1/2	2/2	3/3	time	FLOPS	#params
Proposed, LSTM	99.9	93.6	77.7	41	41.1	12.4
Proposed, ResNet	99.9	93.9	81.2	47	158.6	20.1
TasNet + SincNet	99.7	91.0	74.1	52	23.8	23.0
Wang2018	95.2	36.3	12.5	2	0.15	1.36
Ablation	99.8	92.1	-	50	162.1	17.4

of an N speaker mixture are correctly identified. The columns headed "time," "FLOPS," and "#params" list the computation time, number of floating point operations, and number of parameters instantiated for the recognition of 4886 two-speaker mixtures with a batch size of 32 on $4\times$GTX 1080 Ti. The Wang2018 baseline system has far fewer parameters than other systems, and is unable to recognize both speakers in a mixture of Broadcast News speech. For the 2-speaker case, the modular baseline system, TasNet+SincNet, performs better, with correct recognition of both speakers in 91.0% of test data. The ablation study performs slightly better (92.1%) than TasNet+SincNet, and does not approach the accuracies of the complete system (93.6% and 93.9% with syllable-level LSTM and ResNet34, respectively), suggesting that dilated convolution alone is insufficient for this task: the residual attention modules are critical for achieving high accuracy in this real-world scenario. For the 3-speaker case, the proposed systems(77.7% with syllable-level LSTM, 81.2% with ResNet34) also outperformed the modular baseline(74.1%) We also found that joint training brings the reconstruction MSE loss down, from 0.19 to 0.16 in the 2-speaker case, and from 0.32 to 0.27 in the 3-speaker case, possibly because extractor learns speaker-dependent features from classification loss gradients.

6. Conclusions

This paper presents a co-channel speaker identification system using residual attention, deep dilated convolution, and end-to-end optimization. The proposed system outperforms pipelines built with state-of-the-art neural networks: jointly optimizing the model, and using residual attention for the extractor, lead to better performance than cascading state-of-the-art source separation and speaker recognition systems. All three systems achieve near-perfect accuracy in identifying at least one speaker out of two, therefore, it can be inferred that most errors in multi-speaker classification come from imperfect source reconstruction. The proposed system avoids source reconstruction errors using a spectrogram masking method optimized by the back-propagation of speaker identification errors, and by using an 18-layer dilated convolution with residual connections in order to enforce correct streaming of the component speech signals. Compared to the TasNet+SincNet baseline, the proposed model has fewer channels and smaller kernels, therefore a smaller number of parameters and faster runtime possibly due to better parallelism.

7. Acknowledgements

8. References

[1] D. Morgan, E. George, L. Lee, and S. Kay, "Co-channel speaker separation," in *Proc. ICASSP*, 1995, pp. 828–831.

[2] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party," *ACM Transactions on Graphics*, vol. 37, no. 4, p. 1–11, Aug 2018. [Online]. Available: http://dx.doi.org/10.1145/3197517.3201357

[3] J. B. Allen, "How do humans process and recognize speech?" *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 567–577, Oct 1994.

[4] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 1, 1995.

[5] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[6] K. R. Krishnamachari, R. E. Yantorno, D. S. Benincasa, and S. J. Wenndt, "Spectral autocorrelation ratio as a usability measure of speech segments under co-channel conditions," in *IEEE Symposium on Intelligent Signal Processing and Communication systems*, 2000.

[7] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, no. 4, pp. 297–336, 1994.

[8] Y. Shao and D. Wang, "Co-channel speaker identification using usable speech extraction based on multi-pitch tracking," in *Proc. ICASSP*, vol. II, 2003, pp. 205–208.

[9] P. Mowlaee, R. Saeidi, Z. H. Tan, M. G. Christensen, P. Fränti, and S. H. Jensen, "Joint single-channel speech separation and speaker identification," in *Proc. ICASSP*, 2010, p. 4430–4433.

[10] R. Saeidi, P. Mowlaee, T. Kinnunen, Z.-H. Tan, M. G. Christensen, S. H. Jensen, and P. Fränti, "Signal-to-signal ratio independent speaker identification for co-channel speech signals," in *International Conference on Pattern Recognition (ICPR)*, 2010, pp. 4565–4568.

[11] P. Li, Y. Guan, S. Wang, B. Xu, and W. Liu, "Monaural speech separation based on maxvq and casa for robust speech recognition," *Computer Speech and Language*, vol. 24, no. 1, p. 30–44, 2010.

[12] J. Hershey, S. Rennie, P. Olsen, and T. Kristjansson, "Superhuman multi-talker speech recognition: A graphical modeling approach," *Computer Speech and Language*, vol. 24, no. 1, p. 45–66, 2010.

[13] A. S. Subramanian, C. Weng, M. Yu, S.-X. Zhang, Y. Xu, S. Watanabe, and D. Yu, "Far-field location guided target speech extraction using end-to-end speech recognition objectives," in *Proc. ICASSP*, 2020, pp. 7299–7303.

[14] L. Sarı, S. Thomas, and M. Hasegawa-Johnson, "Training spoken language understanding systems with non-parallel speech and text," in *Proc. ICASSP*, 2020, pp. 8109–8113.

[15] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Journal of Data Warehousing and Mining*, vol. 3, no. 3, pp. 1–17, 2007.

[16] M. Khademian and M. Homayounpour, "Monaural multi-talker speech recognition using factorial speech processing models," *Speech Communication*, vol. 98, pp. 1–16, 2018.

[17] Y. Liu and D. Wang, "A casa approach to deep learning based speaker-independent co-channel speech separation," in *Proc. ICASSP*, 2018, pp. 5399–5403.

[18] S. Wang, Y. Qian, and K. Yu, "Focal kl-divergence based dilated convolutional neural networks for co-channel speaker identification," in *Proc. ICASSP*, 2018.

[19] J. Janský, J. Málek, J. Čmejla, T. Kounovský, Z. Koldovský, and J. Žd'ánský, "Adaptibe blind audio source extraction supervised by dominant speaker identification using x-vectors," https://arxiv.org/pdf/1910.11824.pdf, 2019.

[20] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. ICASSP*, 2017, pp. 241–245.

[21] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech and Language*, vol. 23, no. 12, pp. 2136–2147, 2015.

[22] K. Chen and A. Salman, "Learning speaker-specific characteristicswith a deep neural architecture," *IEEE Trans. on Neural Networks*, vol. 22, no. 11, pp. 1744–1756, 2011.

[23] L. Sarı, S. Thomas, M. Hasegawa-Johnson, and M. Picheny, "Pretraining of speaker embeddings for low-latency speaker change detection in broadcast news," in *Proc. ICASSP*, 2019, pp. 6286–6290.

[24] R. K. Potter, G. A. Kopp, and H. C. Green, *Visible Speech*, 2nd ed. New York: Dover Publications, 1966.

[25] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *IEEE Conf. Computer Vision and Pattern Recogn. (CVPR)*, 2017, pp. 3156–3164.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

[27] P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[28] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets and fully connected CRFs," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[29] J.-Y. Liu and Y.-H. Yang, "Dilated convolution with dilated gru for music source separation," in *IJCAI International Joint Conference on Artificial Intelligence*, 2019, pp. 4718–4724.

[30] K. Tan, J. Chen, and D. Wang, "Gated residual networks with dilated convolutions for supervised speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 21–25.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Computer Vision and Pattern Recogn. (CVPR)*, 2015, pp. 770–778.

[32] D. Graff, "An overview of broadcast news corpora," *Speech Communication*, vol. 37, no. 1-2, pp. 15–26, 2002.

[33] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech and Language*, vol. 27, no. 8, pp. 1256–1266, 2019.

[34] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform using sincnet," in *IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021–1028.