

ANALYSIS OF IMPACT OF EMOTIONS ON TARGET SPEECH EXTRACTION AND SPEECH SEPARATION

Ján Švec¹, Kateřina Žmolíková¹, Martin Kocour¹, Marc Delcroix², Tsubasa Ochiai²,
Ladislav Mošner¹, Jan “Honza” Černocký¹

¹Brno University of Technology, Faculty of IT, IT4I Centre of Excellence
²NTT Corporation, Japan

ABSTRACT

Recently, the performance of blind speech separation (BSS) and target speech extraction (TSE) has greatly progressed. Most works, however, focus on relatively well-controlled conditions using, e.g., read speech. The performance may degrade in more realistic situations. One of the factors causing such degradation may be intrinsic speaker variability, such as emotions, occurring commonly in realistic speech. In this paper, we investigate the influence of emotions on TSE and BSS. We create a new test dataset of emotional mixtures for the evaluation of TSE and BSS. This dataset combines LibriSpeech and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Through controlled experiments, we can analyze the impact of different emotions on the performance of BSS and TSE. We observe that BSS is relatively robust to emotions, while TSE, which requires identifying and extracting the speech of a target speaker, is much more sensitive to emotions. On comparative speaker verification experiments we show that identifying the target speaker may be particularly challenging when dealing with emotional speech. Using our findings, we outline potential future directions that could improve the robustness of BSS and TSE systems toward emotional speech.

Index Terms— target speech extraction, SpeakerBeam, speech separation, Conv-TasNet, emotion

1. INTRODUCTION

Speech processing applications often suffer from reduced performance in real-world environments because of the presence of interfering speakers. Recent research tackles this problem by pre-processing the multi-talker signal to isolate the speech of individual speakers using deep learning approaches. There are two dominant approaches: blind speech separation (BSS) [1, 2, 3, 4, 5] and target speech extraction (TSE) [6, 7]. The task of BSS is to estimate all sources in a mixture; this is needed in some applications such as automatic meeting transcription [8, 9]. In contrast, TSE aims to extract the speech signal of a target speaker only, while removing all other interferences. It can be a practical alternative to BSS for applications such as smart speakers [7]. The target speaker is determined by providing, e.g., an enrollment recording of the voice of the target speaker. In contrast with BSS, TSE needs to perform both separation and identification of the target speaker.

BSS and TSE have significantly progressed with the advent of deep learning and can achieve excellent separation and extraction performance in well-controlled conditions. However, the performance sometimes degrades when tackling more realistic conditions. Understanding the cause of the performance degradation is crucial to further progress research on BSS and TSE.

There are different factors that may influence the performance of TSE and BSS. Noise and reverberation have been shown to affect the separation performance significantly [10, 11]. Some works also explored the effect of voice characteristics and showed the difficulty of performing TSE or BSS on mixtures of speakers with similar voices [12, 13]. The language [14] or environment [15] mismatch has also been shown to be detrimental to the BSS performance. These prior works have focused on the impact of external conditions (noise or reverberation) or global speaker characteristics (such as voice characteristics and languages). However, the voice of a speaker changes also due to, e.g., health condition, type of speech (read, presentation, natural conversations), emotions etc. The impact of such intrinsic speaker variability has been less explored. In this paper, we focus on the impact of emotion on BSS and TSE. Understanding the effect of emotions on BSS and TSE tasks can have crucial implication on the design of applications as we would expect that voice-user interfaces to understand us even when we, for example, become angry at them.

Several studies have shown that the emotional state of the speaker influences the articulation [16] and the prosody [17]. These changes in the speech signal lead to decreased performance of various speech technologies, including speech recognition [18] or speaker verification [19]. These findings suggest that emotions might influence also the performance of TSE and BSS. There are several ways how the emotions can affect these systems. First, the separation or extraction performance may degrade because BSS or TSE models may not represent well the characteristics of emotional speech if emotional data was not well represented in the training data. On the other hand, it may be easier to separate or distinguish speakers in the mixture if they speak with different emotions. Finally, for TSE, the target speech and the enrollment could have different emotions, which may make it hard to identify the target speaker in the mixture.

In our work, we analyze the effect of emotions on TSE and BSS systems. For achieving this, we create a dataset of emotional speech mixtures, based on The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [20]. We call this dataset RAVDESS2Mix. The dataset is prepared in a way, that enables us to isolate the impact of emotions from other factors such as semantic content. With this dataset, we reveal that BSS is not affected much by emotions. In contrast, TSE performance degrades severely when the enrollment and target utterance in the mixture have mismatched emotions. We thus hypothesize that the challenging task is the identification of the target speaker using mismatched enrollment, and we support this hypothesis with speaker identification experiments.

978-1-6654-6867-1/22/\$31.00 ©2022 IEEE

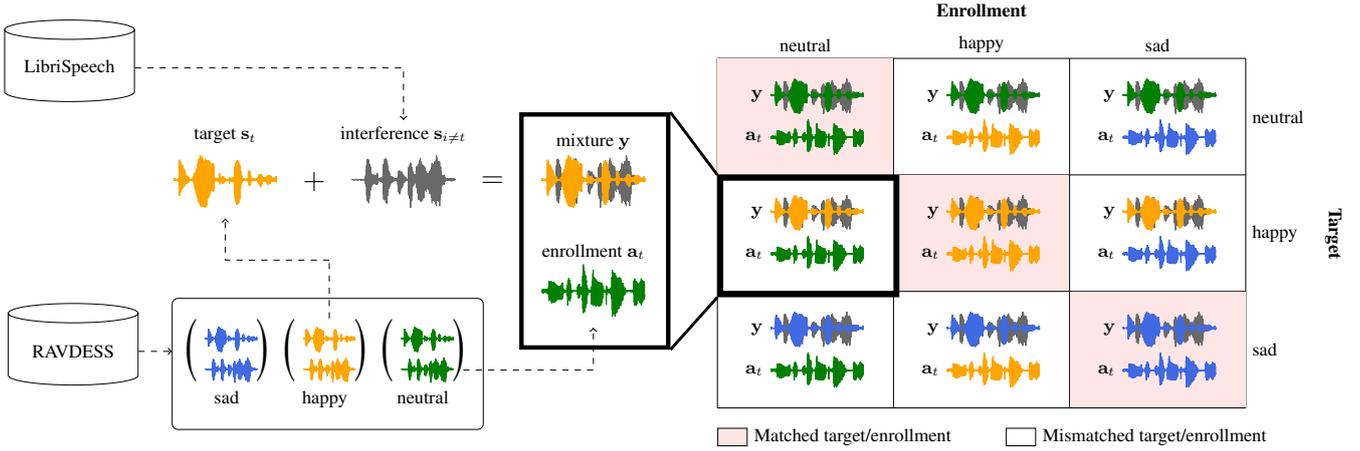


Fig. 1: RAVDESS2Mix: illustration of the dataset creation for 3 emotions, showing the detailed generation of one sample of neutral-happy set. In actual RAVDESS2Mix, we use 8 different emotions, leading to $8 \times 8 = 64$ combinations of (mix, enroll).

2. BSS AND TSE TASKS DESCRIPTION

In both TSE and BSS tasks, the input of the processing is a mixture defined as

$$\mathbf{y} = \sum_{i=1}^S \mathbf{s}_i + \mathbf{n}, \quad (1)$$

where $\mathbf{y} \in \mathbb{R}^T$, T is the duration of the signals in samples, \mathbf{s}_i is the time-domain speech signal of the i -th speaker in the mixture, S is the number of speakers and \mathbf{n} represents additional noise.

A BSS model takes the mixture \mathbf{y} as input and aims to obtain the signals of all S speakers as the output,

$$\{\hat{\mathbf{s}}_1, \dots, \hat{\mathbf{s}}_S\} = f^{(sep)}(\mathbf{y}), \quad (2)$$

where $\hat{\mathbf{s}}_i$ is the estimated signal of the i -th speaker. In this work, we implement the BSS model $f^{(sep)}$ with a fully-convolutional time-domain audio separation network (Conv-TasNet) [4], that has been widely used for speech separation.

TSE also accepts the mixture \mathbf{y} as the input but, in contrast with BSS, also expects additional information in the form of an enrollment utterance of the target speaker. We denote this utterance for target speaker t as $\mathbf{a}_t \in \mathbb{R}^{T'}$ where T' might be different from T . TSE aims to estimate the speech of speaker t at the output. This is done in two steps: first, a speaker embedding \mathbf{e}_t is extracted from the enrollment utterance, and second, the embedding is used to inform the extraction process about the target speaker:

$$\mathbf{e}_t = f^{(emb)}(\mathbf{a}_t) \quad (3)$$

$$\hat{\mathbf{s}}_t = f^{(tse)}(\mathbf{y}, \mathbf{e}_t), \quad (4)$$

where $f^{(emb)}$ and $f^{(tse)}$ represent the speaker embedding computation and speech extraction network, respectively. The speaker embedding is then used to inform the extraction network, $f^{(tse)}$, about the target speaker. The TSE formulation of the problem avoids some issues of speech separation. For instance, the number of speakers in the mixture does not need to be known. Also, the speaker information clearly determines the speaker at the output, therefore the global speaker ambiguity is avoided [6]. However, the TSE processing

$f^{(tse)}$ needs to perform the identification of the target speaker, in addition to the separation itself, which can lead to additional errors.

We follow time-domain SpeakerBeam [13] to implement $f^{(emb)}$ and $f^{(tse)}$. Time-domain SpeakerBeam is based on the Conv-Tasnet architecture, extending it to accept the speaker embedding. This is done by performing an element-wise multiplication of the embedding \mathbf{e}_t with the series of input signal hidden representations in the extraction network, $f^{(tse)}$. With SpeakerBeam, $f^{(emb)}$ is implemented as a neural network, which is jointly trained with the extraction network.

3. PROPOSED FRAMEWORK TO ANALYZE THE IMPACT OF EMOTIONS

3.1. Controlled dataset design

It may be hard to disentangle the effect of emotions from other factors that may affect separation or extraction performance such as the interference signal, the semantic content of the target or enrollment utterance using existing corpora. Here, we design a controlled experimental framework to analyze specifically the effect of the emotional speech on TSE and BSS tasks. This is done by creating mixtures of an emotional target speaker and a relatively neutral interfering speaker. We create mixtures with identical parameters (speaker identities, interference signal, mixing SNR, semantic content etc), and vary only the emotions of the target speaker. This will enable to compare BSS and TSE performance across emotions.

For TSE, we also use enrollment with various emotions. This allows us to analyze the effect of the emotional variability, both in cases when the enrollment and target have *matched* and *mismatched* emotions.

We created only emotional mixtures for evaluation and intentionally used Librimix [21] for training.

3.2. Dataset creation

We create the RAVDESS2Mix dataset of two-speaker mixtures, such that it is possible to isolate the effect of the emotion on the BSS and TSE tasks. We designed this dataset so that it relies on openly

available resources to allow reproducibility ¹.

For the emotional utterance in the dataset, we use RAVDESS [20]. The database contains 24 professional actors (12 female, 12 male), expressing acoustically and visually two semantic statements. Utterances are spoken with eight different emotions (neutral, calm, happy, sad, angry, fearful, surprise, disgust) with a North American accent. Each emotion is produced at two levels of emotional intensity (normal, strong). Thus, by utilizing RAVDESS, we can create mixtures that differ only in the emotions of the target speaker. The acoustic part of the corpus contains also songs, but we use only speech. For the interferences, we use utterances taken from the test set of LibriSpeech [22]².

Figure 1 shows the mixture creation process of RAVDESS2Mix. We created 384 mixtures for each emotion (each cell in the grid in Figure 1). The RAVDESS dataset has 24 speakers, two semantic statements per speaker, two repetitions per statement and two levels of emotional intensity which results in $24 \times 2 \times 2 \times 2 = 192$ unique utterances. Mixing these with different interfering utterances from LibriSpeech resulted into 384 mixtures.

For TSE, we use as enrollment the utterances with different semantic statement than that in the mixture. The enrollment utterance also has one of eight different emotions. There are thus eight variants of the mixtures that differ only by the emotion of the enrollment. This creates 64 (8×8) variants of the pair (mix, enroll).

4. EXPERIMENTS

4.1. Configuration of TSE and BSS

The TSE and BSS models were trained using the Asteroid [23] toolkit. For BSS, we used Conv-TasNet with the following hyper-parameters: 8 blocks, 4 repetitions, and block dimensionality (128, 512, 128), where the output is produced by skip connections. For TSE, we used the implementation of SpeakerBeam provided in our GitHub repository³ which is based on Conv-TasNet architecture. We use the same hyper-parameters as in BSS. The auxiliary network consists of eight connected 1-D Convolutional blocks [4], corresponding to one repetition in Conv-Tasnet architecture. The auxiliary network has its own encoder. The speaker information is used after the first repetition of the extraction network. For training, we dynamically created mixtures from LibriSpeech utterances with noise from WHAM [10], closely following the mixing process used in LibriMix dataset [21]. For both TSE and BSS, we set the batch size to 20, number of epochs to 200, and used the Adam optimizer [24] with an initial learning rate of $1e-3$, which is halved after five consecutive epochs with no reduction in validation loss. We worked with 8 kHz sample rate to reduce computational cost.

4.2. Results

First, we compare both TSE and BSS results on RAVDESS2Mix with those on the test-set of Libri2mix without noise [21], to get an initial idea of how challenging the emotional dataset is. Table 1 shows the signal to distortion ratio improvement (SDRi) measured according to [25] on both tasks. For RAVDESS2Mix, the results in the ‘‘Avg.’’ column indicate the SDRi of BSS averaged on both

Table 1: Overall performance of the BSS and TSE systems in terms of SDR improvements [dB] over the mixture SDR (first row).

	Libri2Mix	RAVDESS2Mix		
		Avg. over all sources	target/enroll matched	target/enroll mismatched
mix	-0.2	0.2	-3.2	
BSS	15.1	15.5	14.4	
TSE	14.8	-	14.1	12.4

the target emotional and interference speech signals. We see that the averaged SDRi of BSS on Libri2Mix and RAVDESS2Mix are similar, which confirms the validity of our experimental setup, i.e., training on LibriMix and testing on RAVDESS2Mix.

The other columns show the SDRi of the estimated target speech. For TSE, we split the results into the cases when the emotion in the enrollment is matched or mismatched with the emotion in the target utterance. The results of TSE for RAVDESS2Mix are in the same range as Libri2Mix when the target and enrollment are from matched emotions, and also similar to the SDRi obtained with BSS. However, we observe close to 2 dB degradation in the mismatched case. These results clearly show that emotion has a significant impact on TSE.

To get more insights, we present the results for individual combinations of emotions in Table 2. The diagonal in the TSE part of the table (highlighted with a pink background) shows the results for the matched cases. We observe that, in the matched cases, emotion slightly hurts the performance compared to the *neutral-neutral* case. A similar trend is present in BSS results, where changing emotion degrades the performance by up to 1.2 dB. In the mismatched case, for TSE, the impact is much stronger. This is evident, especially in cases, where the target utterance is neutral or calm, while the enrollment contains strong emotion (e.g. calm-fearful resulting in 6.71 dB SDRi). Interestingly, we observe that the results are not symmetrical between enrollment and target emotions. For example, the SDRi fluctuates less when the enrollment is neutral or calm and the target emotion in the mixture changes than when the enrollment emotion changes and the target is neutral or calm. This suggests that in practice, it would be better to collect enrollment utterances without strong emotions or maybe to increase variation of emotions in enrollment during training.

We can conclude from this experiment that emotions affect only moderately BSS performance for all eight emotions we investigated. In contrast, some emotions greatly affect TSE performance especially when there is a mismatch between the enrollment and the target speech in the mixture.

4.3. Analysis of speaker identification performance

As mentioned in Section 2, the TSE task has to solve two problems: first, identifying the target speaker in the mixture, and second, the separation itself leading to the extraction of the target signal. From the comparison of the matched and mismatched TSE cases, together with the BSS performance, we hypothesize that the issue with the emotional speech comes from the speaker identification part of the task. To support this hypothesis and to analyze this behavior further, we performed speaker verification experiments on the emotional speech.

For the speaker verification experiments, we created trials of all

¹Data generating code is available at <https://github.com/BUTSpeechFIT/RAVDESS2Mix>.

²Note that we use LibriMix for training, which prevents the systems from learning to discriminate the target and interference simply because the emotional speech and the interference are taken from different datasets.

³<https://github.com/BUTSpeechFIT/speakerbeam>

Table 2: SDRi [dB] for the estimated target source with BSS and TSE systems.

Target emotion	BSS	TSE (Enrollment emotion)							
		<i>neutral</i>	<i>calm</i>	<i>happy</i>	<i>sad</i>	<i>angry</i>	<i>fearful</i>	<i>disgust</i>	<i>surprised</i>
<i>neutral</i>	15.0	15.1	14.8	12.0	13.3	9.5	9.8	13.2	12.8
<i>calm</i>	14.5	13.6	14.1	9.2	10.8	6.8	6.7	10.9	10.6
<i>happy</i>	14.5	13.0	12.5	14.5	13.4	13.8	13.1	13.8	13.6
<i>sad</i>	14.3	13.2	12.7	12.7	13.7	11.3	11.9	13.0	12.8
<i>angry</i>	14.6	12.3	11.3	14.3	12.7	14.5	13.8	13.8	13.4
<i>fearful</i>	14.5	12.1	12.0	13.9	13.3	13.6	14.2	13.4	13.8
<i>disgust</i>	14.0	12.4	12.0	13.0	12.5	12.0	11.8	13.6	12.7
<i>surprised</i>	13.9	13.0	12.7	13.4	12.3	12.5	12.6	13.1	13.7

possible pairs of utterances in the RAVDESS dataset. We extracted embeddings \mathbf{e}_t from these utterances using the auxiliary network of the TSE system $f^{(emb)}$. The embeddings are then compared using cosine distance⁴. The left part of Table 3 shows the equal error rate (EER) results of this experiment for different pairs of emotions in the trials. Due to the limited space, we selected four representative emotions. The performance of the speaker verification decreases steeply with the presence of emotions different from neutral. This indicates that the embeddings extracted by the TSE system are strongly sensitive to emotions, which in many cases obfuscates the speaker information. This explains the degraded performance of TSE when the enrollment contains emotion, i.e., if the speaker information in the embedding is obfuscated, the TSE network might extract the incorrect speaker.

Some variants of TSE frameworks use external speaker embeddings trained on large corpora with the task of speaker classification [7, 26]. Potentially, such embeddings might be more robust and deal better with intra-speaker variability, such as emotions. To get an idea of whether using such embeddings could solve the problem with the emotional speech, we repeated the above speaker verification experiment with x-vectors. We trained a standard extraction network based on the ResNet backbone [27]. The network is trained on the VoxCeleb2-dev [28] with noise and reverberation augmentations. The results are shown in the right part of Table 3. The absolute speaker verification performance with x-vectors⁵ is about twice better than with the embedding of the TSE system — this is expected, as the embeddings of the TSE system were not explicitly trained for speaker verification and might not be well suited for cosine similarity metric. More interestingly, the performance of x-vectors is also significantly influenced by emotions. The degradation is smaller in the cases when the emotions in the trial are matched. In the mismatched case, the relative difference to the neutral-neutral case is close to the embeddings of the TSE system (about 4-6 times worse). This suggests that using x-vectors for TSE has the potential to slightly improve the performance on the emotional data, but most probably would not solve the problem completely.

Our experiments revealed that there is a need to increase the robustness of TSE systems to emotional speech. One option could be to learn speaker embeddings more robust to emotions by, e.g., augmenting the training data with real or synthetic emotional speech. An alternative option could be to exploit other clues than enrollment to identify the target speaker such as video [29, 30, 31], which may be less sensitive to emotions. We could exploit the visual resource

⁴Note that with Euclidean distance, we obtained the same trends.

⁵In our notation, x-vectors refer to speaker embeddings extracted by ResNet.

Table 3: Speaker verification with TSE and x-vector embeddings. Results are in terms of EER [%].

	TSE embeddings				x-vector embeddings			
	<i>neutr.</i>	<i>calm</i>	<i>angry</i>	<i>surpr.</i>	<i>neutr.</i>	<i>calm</i>	<i>angry</i>	<i>surpr.</i>
<i>neutral</i>	4.9	14.2	31.4	24.1	2.4	5.9	11.1	9.8
<i>calm</i>	14.2	9.6	30.3	25.2	5.9	7.3	12.9	11.4
<i>angry</i>	31.4	30.3	30.4	31.6	11.1	12.9	9.0	12.2
<i>surprised</i>	24.1	25.2	31.6	19.3	9.8	11.4	12.2	7.3

of RAVDESS dataset to investigate if audio-visual features could improve the robustness of TSE to emotions.

5. CONCLUSION

In this work, we investigated the influence of emotions on TSE and BSS. For this purpose, we created a new emotional speech mixture evaluation dataset called RAVDESS2Mix, which enabled us to perform controlled experiments to analyze the impact of eight emotions on the performance of BSS and TSE. Our experiments revealed that emotions had a relatively small impact on BSS but a larger impact on TSE, especially in the case of mismatched emotions between the enrollment and the target speech in the mixture. We hypothesized that this performance degradation is mainly due to the difficulty in identifying the target speaker, which we confirmed with speaker verification experiments. We hope that the results of this study will foster more research on solutions to increase the robustness of TSE systems to internal speaker variability such as emotions.

6. ACKNOWLEDGEMENT

The work was partly supported by European Union’s Horizon 2020 project No. 870930 — WELCOME, and by Czech Ministry of Education, Youth and Sports from project no. LTAIN19087 “Multilinguality in speech technologies”. Computing on IT4I supercomputer was supported by the Czech Ministry of Education, Youth and Sports from the Large Infrastructures for Research, Experimental Development and Innovations project “e-Infrastructure CZ — LM2018140”.

7. REFERENCES

- [1] D. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” *IEEE/ACM Transactions on*

- Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.
- [2] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in *Proc. ICASSP*. IEEE, 2016, pp. 31–35.
 - [3] D. Yu, M. Kolbaek, 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.
 - [4] Y. Luo and N. Mesgarani, “Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug 2019.
 - [5] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, “Attention is all you need in speech separation,” in *Proc. ICASSP*, 2021, pp. 21–25.
 - [6] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černocký, “Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.
 - [7] Q. Wang, I. L. Moreno, M. Saglam, K. Wilson, A. Chiao, R. Liu, Y. He, W. Li, J. Pelecanos, M. Nika, et al., “Voicefilter-lite: Streaming targeted voice separation for on-device speech recognition,” *arXiv preprint arXiv:2009.04323*, 2020.
 - [8] S. Araki, H. Sawada, and S. Makino, “Blind speech separation in a meeting situation with maximum snr beamformers,” in *Proc. ICASSP*. IEEE, 2007, vol. 1, pp. I–41.
 - [9] D. Raj, P. Denisov, Z. Chen, H. Erdogan, Z. Huang, M. He, S. Watanabe, J. Du, T. Yoshioka, Y. Luo, et al., “Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 897–904.
 - [10] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, “Wham!: Extending speech separation to noisy environments,” in *Proc. Interspeech*, Sept. 2019.
 - [11] M. Maciejewski, G. Wichern, E. McQuinn, and J. L. Roux, “Whamr!: Noisy and reverberant single-channel speech separation,” 2019.
 - [12] D. Ditter and T. Gerkmann, “Influence of speaker-specific parameters on speech separation systems,” in *Proc. Interspeech*, 2019, pp. 4584–4588.
 - [13] M. Delcroix, T. Ochiai, K. Zmolikova, K. Kinoshita, N. Tawara, T. Nakatani, and S. Araki, “Improving speaker discrimination of target speech extraction with time-domain speakerbeam,” 2020.
 - [14] M. Borsdorf, C. Xu, H. Li, and T. Schultz, “GlobalPhone Mix-To-Separate Out of 2: A Multilingual 2000 Speakers Mixtures Database for Speech Separation,” in *Proc. Interspeech*, 2021, pp. 3905–3909.
 - [15] M. Maciejewski, G. Sell, Y. Fujita, L. P. Garcia-Perera, S. Watanabe, and S. Khudanpur, “Analysis of robustness of deep single-channel speech separation using corpora constructed from multiple domains,” in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 165–169.
 - [16] S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan, “An articulatory study of emotional speech production,” in *Proc. Interspeech*, 2005, pp. 497–500.
 - [17] K. R. Scherer, “Vocal communication of emotion: A review of research paradigms,” *Speech communication*, vol. 40, no. 1-2, pp. 227–256, 2003.
 - [18] R. Munot and A. Nenkova, “Emotion impacts speech recognition performance,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, Minneapolis, Minnesota, June 2019, pp. 16–21, Association for Computational Linguistics.
 - [19] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, “X-vectors meet emotions: A study on dependencies between emotion and speaker recognition,” in *Proc. ICASSP*, 2020, pp. 7169–7173.
 - [20] S. R. Livingstone and F. A. Russo, “The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english,” *PLOS ONE*, vol. 13, no. 5, pp. 1–35, 05 2018.
 - [21] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, “Librimix: An open-source dataset for generalizable speech separation,” 2020.
 - [22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *Proc. ICASSP*, 2015, pp. 5206–5210.
 - [23] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, et al., “Asteroid: the pytorch-based audio source separation toolkit for researchers,” 2020.
 - [24] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
 - [25] E. Vincent, R. Gribonval, and C. Fevotte, “Performance measurement in blind audio source separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.
 - [26] W. Li, P. Zhang, and Y. Yan, “Tenet: target speaker extraction network with accumulated speaker embedding for automatic speech recognition,” *Electronics Letters*, vol. 55, pp. 816–819(3), July 2019.
 - [27] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
 - [28] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, “Voxceleb: Large-scale speaker verification in the wild,” *Computer Speech and Language*, 2019.
 - [29] 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: A speaker-independent audio-visual model for speech separation,” *ACM Trans. Graph.*, vol. 37, no. 4, jul 2018.
 - [30] T. Afouras, J. S. Chung, and A. Zisserman, “The Conversation: Deep Audio-Visual Speech Enhancement,” in *Interspeech*, 2018, pp. 3244–3248.
 - [31] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani, “Multimodal SpeakerBeam: Single channel target speech extraction with audio-visual speaker clues,” in *Proc. Interspeech*, 2019, pp. 2718–2722.