EXPERTS VERSUS ALL-ROUNDERS: TARGET LANGUAGE EXTRACTION FOR MULTIPLE TARGET LANGUAGES

Marvin Borsdorf [1], *Kevin Scheck*[2], *Haizhou Li*[3,1], *Tanja Schultz*[2]

[1]Machine Listening Lab (MLL), University of Bremen, Germany
[2]Cognitive Systems Lab (CSL), University of Bremen, Germany
[3]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
marvin.borsdorf@uni-bremen.de, scheck@uni-bremen.de

ABSTRACT

Target language extraction (TLE) is a novel task in the field of selective auditory attention, which seeks to extract all speech signals that are spoken in a target language from other sources in a multilingual cocktail party. In our prior studies, a TLE model was trained to extract a predefined, single target language, referred to as Single-TLE. In this paper, we extend the Single-TLE framework to Multi-TLE. Multi-TLE models can also extract all speech signals of one specific target language, but they are optimized on a set of multiple target languages during training. As such, they learn the characteristics of several target languages and can replace multiple Single-TLE models without retraining. We perform experiments on the GlobalPhoneMCP database and incorporate a dynamic language mixing scheme for training. The Multi-TLE model does not only outperform Single-TLE models, but when given a language ID as additional input, it is also able to extract the speech of a specific target language from a mixture which contains multiple learned target languages.

Index Terms— Target language extraction, selective auditory attention, multilingual, GlobalPhone, cocktail party problem

1. INTRODUCTION

In recent years, researchers have studied on how to equip machines with the human ability of selective auditory attention. This ability allows human listeners to easily disentangle overlapping sound sources within an arbitrary soundscape, which is also referred to as the "cocktail party problem" [1]. The research in this field can be roughly divided into two branches, namely blind source separation [2, 3, 4, 5, 6, 7] and speaker extraction [8, 9, 10, 11, 12]. While the former seeks to "blindly" separate all sources in a given audio mixture, the latter utilizes an additional given reference signal to find and extract only the desired target source. Both methods show some limitations, which we will briefly describe next.

Blind source separation seeks to separate multiple sources in parallel. It therefore usually requires the number of sources to be known in advance. However, such information is not always available in practice. Techniques to address this include iterative and recursive methods [13, 14], assumption of a maximum number of sources for the utterance level [4, 15, 16] or the frame level [17], or devising a fixed sequential output stream [18]. Despite much progress, these techniques face the combinatorial and computational challenges as the number of sources increases, as described, e.g., by Kanda et al. [18] and Dovrat et al. [19]. Speaker extraction considers source separation from a different point of view. This technique does not separate multiple speaker streams simultaneously, but rather extracts the desired target speaker's voice only, based on a given reference signal of the target speaker.

Recently, we introduced a novel task called "target language extraction" (TLE) [20] which seeks to extract the voices of all speakers talking in a specific target language from an audio mixture. TLE is similar to target speaker extraction, except that we are interested in a particular language instead of a speaker. In this work, we refer such models as "single-target language extraction" (Single-TLE) models. Unlike blind source separation and speaker extraction, the Single-TLE model requires neither information about the number of speakers nor any speaker characteristics in advance. It works as an "expert", i.e., it is tailored to one fixed target language and could be plugged into an arbitrary multilingual cocktail party scene to extract all speech signals that are spoken in the target language in one step, independent of the number of speakers. While this concept is well suited for applications in which the target language is consistent, one Single-TLE model is required for each target language. This limits the scope of application for multilingual settings.

In this work, we extend the Single-TLE framework to extract multiple target languages with the same model without retraining. The model is trained on mixtures in which the target language is sampled from a set of target languages. Such "all-rounder" models are referred to as "multi-target language extraction" (Multi-TLE) models and can replace multiple Single-TLE models. We show that training a Multi-TLE model with different target languages improves the performance compared to its Single-TLE counterparts in most cases. Finally, we introduce a conditioning mechanism to specify which learned target language to extract from a mixture that contains multiple target languages, as shown in Figure 1.

The rest of the paper is organized as follows. In Section 2, we describe the construction of the different types of TLE models. In Section 3, we explain the experiments, followed by Section 4, in which we discuss the results. In Section 5, we conclude the study and highlight some future steps. All information and scripts will be made publicly available[1].

2. EXPERTS AND ALL-ROUNDERS FOR TLE

2.1. From one to multiple target languages

The construction of TLE models has been introduced as supervised learning scheme in which tuples of input and ground truth mixtures

This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (University Allowance, EXC 2077, University of Bremen)

[1]https://github.com/mborsdorf/MultiTargetLanguageExtraction

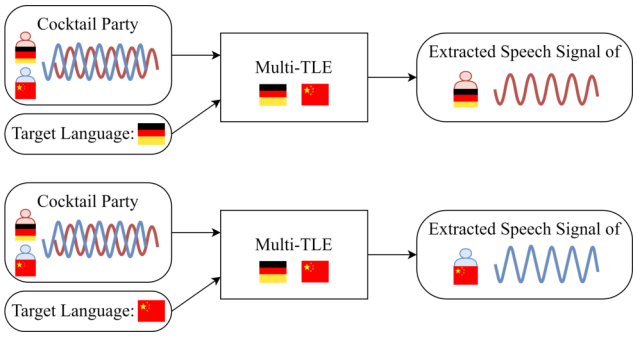


Fig. 1: Illustration of a Multi-TLE model which is trained to extract both German and Chinese Shanghai. At run-time, the Multi-TLE model either extracts German (top) or Chinese Shanghai (bottom) speech from a two-speaker mixture by specifying the target language ID as part of the input.

are given [20]. While the input data contains speech signals of two languages, the ground truth data contains the speech signals of the target language only. The models are trained to extract the target language signals from the input mixture to match the ground truth. As such, the models can learn and embed the target language specific characteristics directly from the data. During training, the target language remains the same, while the interfering language varies.

Studies on multilingual speech separation show that deep neural network based models have the ability to learn and embed characteristics of multiple languages [4, 21, 22]. This is mainly accomplished by extending the monolingual training data with multilingual data while the supervised training scheme remains the same. We believe that this method could be transferred to TLE so as to extend Single-TLE to Multi-TLE. Such a Multi-TLE model could be applied to different multilingual speaker mixtures and extract the speech signals spoken in one of the training target languages. However, it is unable to single out one particular target language when the input audio mixture contains multiple learned target languages at run-time, because the Multi-TLE model cannot determine which of the target languages to extract. We refer to these models as "multi-target language extraction fixed" (Multi-TLE-Fixed) models.

To extract a particular target language when multiple learned target languages are present in the mixture, we propose "multi-target language extraction switch" (Multi-TLE-Switch) models. In addition to the input mixture, they take the language ID of the to-be-extracted target language as input, as shown in Figure 1. This is inspired by the success of "embeddings" that are used, e.g., in speaker extraction or universal sound selection [10] to guide the separation models. Depending on the specific application, there are different ways to design such a conditioning, as, e.g., discussed by Delcroix et al. [23]. In this work, we do not use language embeddings for this conditioning, but use an one-hot encoding instead. This is because the number of possible target languages is limited compared to the number of possible speakers.

2.2. Dynamic language mixing

Similar to the dynamic mixing training scheme, introduced by Zeghidour et al. [7], we incorporate a procedure to dynamically create language-specific speech mixtures on-the-fly during training. We define a fixed set of target languages and a fixed set of interfering languages for training a TLE model. In our experiments, we only cover two-speaker mixtures in which each speaker talks in a different language. However, the concept of dynamic language mixing can be easily extended to more speakers and languages.

To create an input mixture of two utterances, we sample one utterance of the target languages set and one utterance of the interfering languages set. The target languages set comprises a single language when training Single-TLE models and multiple languages when training Multi-TLE models. The interfering language utterance depends on the to-be-trained model and can be sampled either solely from the predefined set of interfering languages or from the union of interfering languages and target languages, excluding the language of the previously sampled target language utterance. The target language utterance, which will be mixed with the interfering language utterance, represents the ground truth at the same time.

Before the training, we use the scripts introduced by Isik et al. [24] in order to calculate and store the active speech level for each utterance as well as to resample the audio data to 8 kHz. To mix the utterances on-the-fly during training, we apply the same procedure as described in Isik et al. [24]. The sampled audio data are first individually scaled according to their respective active speech level. Since we create "min" mixtures, the longer utterance is truncated to the length of the shorter one. Subsequently, we uniformly sample a signal-to-noise ratio (SNR) from the interval of -5 dB to 5 dB in order to randomly choose the target speaker or the interfering speaker to be in the foreground. We apply $+\frac{SNR}{2}$ to the target utterance and $-\frac{SNR}{2}$ to the interfering utterance to ensure zero mean. Lastly, the maximal amplitude of the two single utterances and the mixed signal is calculated and used in order to normalize the three audio signals and to avoid clipping. We consider the final mixtures as fully overlapped mixture speech.

3. EXPERIMENTAL SETUP

3.1. Data

We prepare the experimental data following the GlobalPhoneMCP corpus design concept [20]. The data set includes a predefined split of 5,172 training, 3,066 validation, and 2,510 test utterances, covering 22 different languages. While the training and validation sets comprise different utterances of the same speakers, the test set speakers are disjoint from these sets (open set condition). The data is based on the GlobalPhone 2000 Speaker Package[2], which is a subset of the comprehensively studied multilingual GlobalPhone corpus, introduced by Schultz [25] and Schultz et al. [26]. The GlobalPhone 2000 Speaker Package is distributed under commercial and research licenses by the European Language Resources Association (ELRA) [27]. The 22 languages cover 2,000 native speakers in total with roughly 40 seconds of speech per speaker.

Unlike to our previous experiments on TLE [20], here we process two-speaker mixtures only. However, the experimental setup could be easily extended to mixtures of more speakers. While we apply the dynamic language mixing concept (see Section 2.2) for training and validation, the testing utilizes fixed mixture sets to ensure comparability. To simulate the test sets, we define fixed speaker pairs and create the actual mixtures with a static version of the dynamic language mixing scripts (applying the original scripts, introduced by Isik et al. [24], would also be possible).

In this experimental setup, we select three fixed target languages. With German (GE) and Portuguese (PO) we cover two languages of the biggest family of languages and with Chinese Shanghai (WU)

[2]https://catalog.elra.info/en-us/repository/browse/ELRA-S0400/

we include one tonal language. For each of the target languages, we apply the following steps to create a fixed test set. First, we pick one of the remaining 21 interfering languages and repeat its utterances four times to have a meaningful amount of data. Second, we repeat the target language utterances to match the number of interfering utterances. Lastly, we randomly mix those utterances by applying the same procedure as for the creation of the training and validation mixtures, making sure that no utterance combination is mixed twice. We repeat this procedure for all individual 21 interfering languages. This leads to 21 language mixture test data sets for each target language. Overall, the test sets for GE, PO, and WU comprise 9,680, 9,580, and 9,720 utterances respectively.

3.2. Network architecture

All models in our experiments use the recent SepFormer [28] architecture, implemented in the SpeechBrain [29] toolkit[3]. Originally proposed to estimate multiple speaker masks for blind source separation, we adapt the SepFormer architecture to predict a single mask since we are only interested in extracting the target language signal. We refer this structure as SepFormer$_{\text{Single-Mask}}$ [20]. The rest of the model settings are adopted from the best-performing implementation in Subakan et al. [28]. In particular, the SepFormer$_{\text{Single-Mask}}$ block consists of eight Intra- and Inter-Transformer layers, repeated two times. The network has eight attention heads for the multi-head-attention layers and integrated feed-forward networks with 1024 dimensions.

For the Multi-TLE-Switch model, we introduce a language ID conditioning (see Section 2.1). Since our experiments cover only three fixed target languages, we apply a simple three-dimensional one-hot encoding, instead of a more complex trainable embedding. The one-hot encoding is concatenated with the encoder output along the feature dimension, as proposed by Deng et al. [30]. This causes a marginal increase in the number of trainable parameters from 25.613 million for the Single-TLE and the Multi-TLE-Fixed models to 25.614 million for the Multi-TLE-Switch model.

3.3. Training and evaluation

In our experiments, we train five different TLE models. Three Single-TLE models are trained on Chinese Shanghai (WU), German (GE), and Portuguese (PO) respectively, while one Multi-TLE-Fixed model and one Multi-TLE-Switch model are trained on all three target languages altogether. We remove three random languages from the training and validation sets to evaluate our models on open set conditions for speakers and languages. Hence, the languages French, Tamil, and Thai are only part of the test data sets.

To train the TLE models, we use Adam as optimizer with an initial learning rate of $1.5e^{-4}$ and halve the learning rate if the validation loss does not decrease in three subsequent epochs. The validation loss is also monitored by an early stopping mechanism which stops the training if the validation loss does not decrease within six consecutive epochs. We apply a batch size of two, set the clip norm to a value of five, and do not use any weight decay. The models are trained using the negative value of the scale-invariant signal-to-distortion ratio (SI-SDR) [31]. During evaluation, we calculate the SI-SDR improvement (dB) which is defined as the difference between the SI-SDR of the extracted target signal and the SI-SDR of the input mixture signal with respect to the target signal.

The Single-TLE, Multi-TLE-Fixed, and Multi-TLE-Switch models are trained with data tuples as described in Section 2.1. In each epoch, we apply the dynamic language mixing to create 20,000 and 10,000 data tuples for training and validation respectively. During training and validation, the input mixtures are split into chunks of four seconds. Mixtures with a length shorter than two seconds are discarded, while mixtures with a length between two and four seconds are zero-padded to match the chunk size. During evaluation, the model processes the entire input mixture at once.

[3]https://github.com/speechbrain/speechbrain

Table 1: Target language extraction results for three different target languages. The performance is reported in terms of SI-SDR improvement (dB) over all 21 language mixture data sets. Each Single-TLE model is trained on one target language only, while the Multi-TLE models are trained on three different target languages.

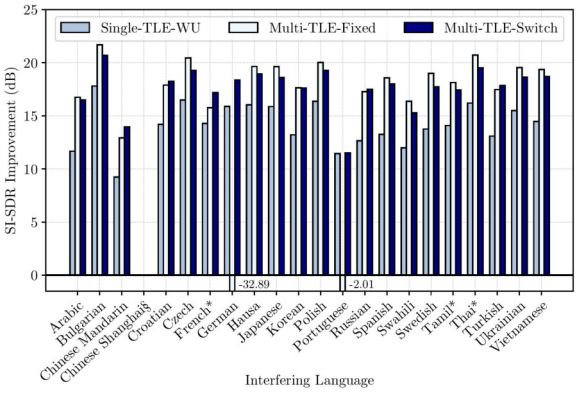
TLE model	Target language		
	WU	GE	PO
Single-TLE-WU	14.17	-	-
Single-TLE-GE	-	16.65	-
Single-TLE-PO	-	-	15.32
Multi-TLE-Fixed	14.95	16.93	12.10
Multi-TLE-Switch	**17.66**	**17.16**	**16.39**

4. RESULTS AND DISCUSSION

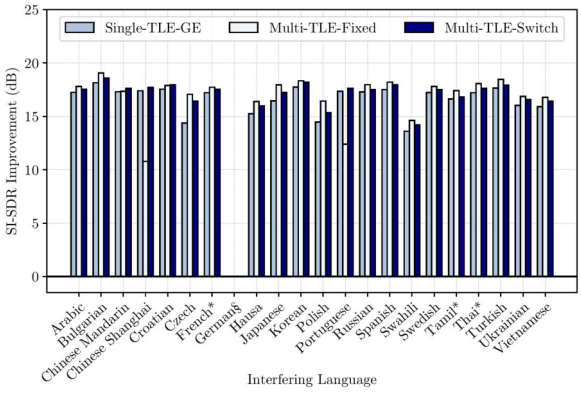
We report the target language extraction performance in terms of average SI-SDR improvement (dB) for three Single-TLE models and two Multi-TLE models. The details for each target language are shown in Figure 2a-c. The average SI-SDR improvements (dB) over all 21 language mixture data sets for each model are given in Table 1.

The three Single-TLE models show adequate and good extraction performances and provide the baselines for further comparisons, as depicted in Figure 2a-c and Table 1. For the majority of the language combinations, the Multi-TLE-Fixed model shows a better performance compared to its Single-TLE counterparts, illustrated in Figure 2a-c. These results indicate that a TLE model benefits from being trained on multiple target languages. This corroborates findings from other related research fields, such as acoustic modeling [32, 33, 34] and speech separation [22]. A possible explanation could be that, due to the task of being able to recognize and extract different target languages instead of a single target language during training, the model could learn more robust and distinguishable phonetic representations that improves the overall TLE performance.

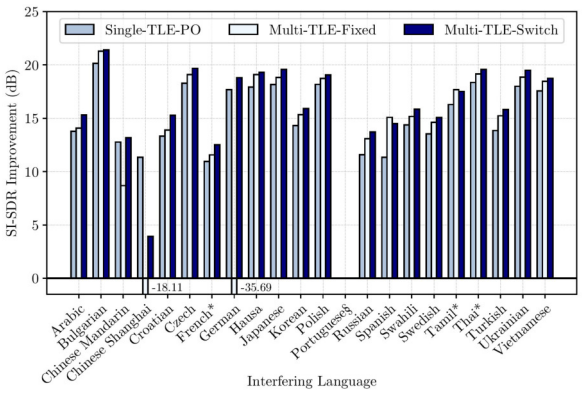
However, the Multi-TLE-Fixed model fails on conditions in which multiple trained target languages are present in the input mixture. In these scenarios, the model cannot determine which target language to extract, leading to a comparatively worse extraction performance. If the model is meant to extract GE while WU or PO are given as interfering language, as shown in Figure 2b, the SI-SDR improvement is substantially less compared to the Single-TLE-GE model. When either WU (Figure 2a) or PO (Figure 2c) is the to-be-extracted target language, the SI-SDR improvements are negative when any of the remaining target languages is present as interfering language in the mixture. This indicates that the extracted target signal contains more distortions than before the processing. Listening to the model output for mixtures that comprise WU and PO reveals that parts of both languages are present in the extracted signal. When GE is part of the mixtures as either target or



(a) TLE results for Chinese Shanghai (WU) as target language.



(b) TLE results for German (GE) as target language.



(c) TLE results for Portuguese (PO) as target language.

Fig. 2: Average SI-SDR improvement (dB) of different TLE models for test sets in which the target language is given as Chinese Shanghai (a), German (b), or Portuguese (c). The interfering languages consist of the respective 21 remaining languages. A star (*) denotes that a language was not used during training and is only used as interfering language during evaluation. A paragraph symbol (§) marks conditions in which the target and interfering languages match, such that no separation in terms of target language extraction can be performed.

interfering language, mostly GE is extracted, indicating a bias of the model towards this language.

Incorporating a language ID as additional input allows the Multi-TLE-Switch model to extract a trained target language even if another trained target language is given as interfering language. The extraction performance in those cases is mostly similar as for the Single-TLE models, except for the mixture comprising PO as target language and WU as interfering language. We analyzed the reconstructed audio files but could not find any clue on why the conditioning mechanism fails in this specific language combination. We assume that the model is biased towards WU. Comparing the results in Figure 2a-c for the Multi-TLE-Fixed and the Multi-TLE-Switch models indicates that conditioning a Multi-TLE model with a language ID does not consistently improve the separation performance. However, the average SI-SDR improvement over all language mixture data sets, given in Table 1, increases because the Multi-TLE-Switch model shows adequate extraction performance for mixtures that comprise two trained target languages. Finally, Figure 2a-c shows that all models work for unseen interfering language conditions, as French, Tamil, and Thai have not been used for training.

5. CONCLUSION AND FUTURE WORK

We proposed a TLE system to separate predefined target languages from multilingual cocktail parties using a single model, which is optimized by dynamically mixing utterances of multiple languages during training. Our results indicate that a TLE model, trained on multiple target languages, shows mostly a higher extraction performance compared to its monolingual counterparts. However, the unconditioned Multi-TLE-Fixed model fails to extract the target speech when multiple target languages are given in the input mixture. By conditioning the Multi-TLE-Switch model with an one-hot encoding of the language ID, it is able to extract the target speech signal from mixtures in which multiple target languages are given.

In our future work, we aim to extend Multi-TLE models to more challenging scenarios, in which more speakers and languages are included in the multilingual cocktail party. Furthermore, we plan to analyze the effect of the used target languages for training TLE models to get insights on, e.g., how many additional target languages yield the highest extraction gain, or whether the phonetic (dis-)similarity of the target languages influence the overall performance. In addition, we will study the behavior of TLE models in code-switching situations with respect to system performance and robustness.

6. REFERENCES

[1] E. C. Cherry, "Some Experiments on the Recognition of Speech, with One and with Two Ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, 1953.

[2] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep Clustering: Discriminative Embeddings for Segmentation and Separation," in *ICASSP*, 2016.

[3] D. Yu, M. Kolbæk, Z. Tan, and J. Jensen, "Permutation Invariant Training of Deep Models for Speaker-Independent Multi-talker Speech Separation," in *ICASSP*, 2017.

[4] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multi-talker Speech Separation with Utterance-level Permutation Invariant Training of Deep Recurrent Neural Networks," *IEEE/ACM TASLP*, vol. 25, no. 10, 2017.

[5] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE/ACM TASLP*, vol. 27, no. 8, 2019.

[6] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation," in *ICASSP*, 2020.

[7] N. Zeghidour and D. Grangier, "Wavesplit: End-to-End Speech Separation by Speaker Clustering," *arXiv preprint arXiv:2002.08933v2*, 2020.

[8] 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, 2019.

[9] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," in *INTERSPEECH*, 2019.

[10] T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito, K. Kinoshita, and S. Araki, "Listen to What You Want: Neural Network-based Universal Sound Selector," in *INTERSPEECH*, 2020.

[11] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "SpEx+: A Complete Time Domain Speaker Extraction Network," in *INTERSPEECH*, 2020.

[12] Z. Zhang, B. He, and Z. Zhang, "X-TaSNet: Robust and Accurate Time-Domain Speaker Extraction Network," in *INTERSPEECH*, 2020.

[13] K. Kinoshita, L. Drude, M. Delcroix, and T. Nakatani, "Listening to Each Speaker One by One with Recurrent Selective Hearing Networks," in *ICASSP*, 2018.

[14] T. von Neumann, C. Boeddeker, L. Drude, K. Kinoshita, M. Delcroix, T. Nakatani, and R. Haeb-Umbach, "Multi-talker ASR for an Unknown Number of Sources: Joint Training of Source Counting, Separation and ASR," in *INTERSPEECH*, 2020.

[15] Y. Luo and N. Mesgarani, "Separating Varying Numbers of Sources with Auxiliary Autoencoding Loss," in *INTERSPEECH*, 2020.

[16] S. Wisdom, H. Erdogan, D. P. W. Ellis, R. Serizel, N. Turpault, E. Fonseca, J. Salamon, P. Seetharaman, and J. R. Hershey, "What's All the FUSS About Free Universal Sound Separation Data?," in *ICASSP*, 2021.

[17] T. von Neumann, K. Kinoshita, C. Boeddeker, M. Delcroix, and R. Haeb-Umbach, "Graph-PIT: Generalized Permutation Invariant Training for Continuous Separation of Arbitrary Numbers of Speakers," in *INTERSPEECH*, 2021.

[18] N. Kanda, Y. Gaur, X. Wang, Z. Meng, and T. Yoshioka, "Serialized Output Training for End-to-End Overlapped Speech Recognition," in *INTERSPEECH*, 2020.

[19] S. Dovrat, E. Nachmani, and L. Wolf, "Many-Speakers Single Channel Speech Separation with Optimal Permutation Training," in *INTERSPEECH*, 2021.

[20] M. Borsdorf, H. Li, and T. Schultz, "Target Language Extraction at Multilingual Cocktail Parties," in *ASRU*, 2021.

[21] P. Appeltans, J. Zegers, and Hugo van Hamme, "Practical Applicability of Deep Neural Networks for Overlapping Speaker Separation," in *INTERSPEECH*, 2019.

[22] 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 *INTERSPEECH*, 2021.

[23] M. Delcroix, J. B. Vázquez, T. Ochiai, K. Kinoshita, and S. Araki, "Few-Shot Learning of New Sound Classes for Target Sound Extraction," in *INTERSPEECH*, 2021.

[24] Y. Isik, J. L. Roux, S. W. Z. Chen, and J. R. Hershey, "Scripts to Create wsj0-2 Speaker Mixtures," MERL Research, retrieved June 2, 2020, from https://www.merl.com/demos/deep-clustering/create-speaker-mixtures.zip, [Online].

[25] T. Schultz, "GlobalPhone: A Multilingual Speech and Text Database Developed at Karlsruhe University," in *ICSLP*, 2002.

[26] T. Schultz, N. T. Vu, and T. Schlippe, "GlobalPhone: A Multilingual Text & Speech Database in 20 Languages," in *ICASSP*, 2013.

[27] ELRA, "European Language Resources Association (ELRA)," ELRA Catalog, retrieved October 15, 2020, from http://catalog.elra.info, 2020, [Online].

[28] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention Is All You Need In Speech Separation," in *ICASSP*, 2021.

[29] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A General-Purpose Speech Toolkit," *arXiv preprint arXiv:2106.04624v1*, 2021.

[30] C. Deng, S. Ma, Y. Sha, Y. Zhang, H. Zhang, H. Song, and F. Wang, "Robust Speaker Extraction Network Based on Iterative Refined Adaptation," in *INTERSPEECH*, 2021.

[31] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-Baked or Well Done?," in *ICASSP*, 2019.

[32] T. Schultz and A. Waibel, "Multilingual and Crosslingual Speech Recognition," in *DARPA Workshop on Broadcast News Transcription and Understanding*, 1998.

[33] T. Schultz and A. Waibel, "Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition," *Speech Communication*, vol. 35, no. 1-2, 2001.

[34] T. Schultz and A. Waibel, "Experiments on Cross-language Acoustic Modeling," in *EUROSPEECH*, 2001.