TARGET LANGUAGE EXTRACTION AT MULTILINGUAL COCKTAIL PARTIES

Marvin Borsdorf\textsuperscript{1}, Haizhou Li\textsuperscript{2,1}, Tanja Schultz\textsuperscript{3}

\textsuperscript{1}Machine Listening Lab (MLL), University of Bremen, Germany
\textsuperscript{2}Department of Electrical and Computer Engineering, National University of Singapore, Singapore
\textsuperscript{3}Cognitive Systems Lab (CSL), University of Bremen, Germany

marvin.borsdorf@uni-bremen.de

ABSTRACT

Typically, target speaker extraction seeks to extract a target speaker’s contribution according to his or her individual voice characteristics. In a “multilingual cocktail party” however, listeners may desire to extract speaker contributions spoken in a particular language, regardless of the number of contributing speakers. In this paper, we propose a novel task called “target language extraction” (TLE) which extracts voices based on the spoken language rather than on individual speaker characteristics. We introduce a new database for TLE which simulates the multilingual cocktail party problem in mixtures of two and four speakers with German as the target language. The database is derived from the GlobalPhone 2000 Speaker Package and is called “GlobalPhone Multilingual Cocktail Party – German” (GlobalPhoneMCP-GE). Our experimental results show that our approach to TLE achieves very good performance regardless of the number of speakers in the mixture and that TLE generalizes well to unseen speakers and interfering languages. This work represents the first attempt at target language extraction.

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

1. INTRODUCTION

Human listeners can easily disentangle soundscapes into their individual sources and pay attention to a specific voice or sound. This ability is also referred to as “selective auditory attention”. However, emulating this ability on machines represents a long-lasting research problem since its introduction as “cocktail party problem”\cite{1}. In recent years, engineering solutions to selective auditory attention have been comprehensively studied in two ways.

In blind source separation (BSS), a model attempts to separate all sound sources in a given mixture from each other. The modern research in this field was leveraged by the deep clustering method, introduced by Hershey et al.\cite{2}, and the permutation invariant training (PIT), proposed by Yu et al.\cite{3}. Recently, new approaches such as TasNet\cite{4}, Conv-TasNet\cite{5}, DPRNN\cite{6}, Two Step Learning\cite{7}, Wavesplit\cite{8}, SepFormer\cite{9}, SDNet\cite{10}, DPTNet\cite{11}, GALR\cite{12}, and Sandglasset\cite{13} have advanced the state-of-the-art speech separation. BSS usually requires the prior knowledge of the number of sources in the mixture which limits the application in real-world conditions since this information is not always accessible. Some techniques have tried to address this constraint with e.g. the definition of a maximum number of separable sources\cite{14, 15, 16, 17, 18, 19, 20} or the application of iterative or recursive approaches\cite{21, 22, 23, 24}.

Target speaker extraction describes a mechanism which utilizes a reference signal of a desired speaker to guide a speech separation module to isolate this target speaker’s voice from all other sound sources in an audio mixture. The reference signal is usually given by an enrollment utterance of the target speaker but could also be based on other modalities such as video information. This signal is encoded into a speaker embedding and used to disentangle the target speaker from the other sources in the mixture. Recent state-of-the-art speaker extraction solutions include frequency domain methods such as SpeakerBeam\cite{25, 26, 27}, SBF-MTSAL/SBF-MTSAL-Concat\cite{28, 29}, VoiceFilter\cite{30}, VoiceFilter-Light\cite{31}, and Atss-Net\cite{32} as well as time domain approaches as e.g. SpEx\cite{33}, SpEx+\cite{34}, SpEx++\cite{35}, X-TaSNet\cite{36}, TD-SpeakerBeam\cite{37}, a universal sound selector\cite{38}, and a speaker conditioning mechanism\cite{39}.

Both BSS and target speaker extraction mechanisms use speaker characteristics as the attractor for separation and extraction. However, in everyday multilingual soundscapes (“multilingual cocktail party” scenarios), a human listener could desire to attend to speech that is spoken in a particular language, independent of the number of contributing speakers. This scenario is also motivated by the observation that humans usually perceive familiar languages, such as the mother tongue, much more prominent than unfamiliar languages. Practical examples for multilingual cocktail party scenes are crowded places with international character such as canteens, airports, train stations, conferences or touristic spots. For example, a traveler seeks to separate announcements at the airport or train station, which are given in the desired language, from the residual sources. This mechanism has to work in the presence of any interfering language, speaker, or sound, and for any number of sources in the soundscape. We strongly believe, that systems which can emulate this ability could promote the research on selective auditory attention and other related fields, such as automatic speech recognition (ASR) or machine translation, in multilingual cocktail party situations. This paper provides the following three contributions:

1. We propose a novel task called “target language extraction” (TLE). It isolates all speakers belonging to a target language from the residual sources in an audio mixture at once.

2. We create a TLE database, “GlobalPhone Multilingual Cocktail Party - German” (GlobalPhoneMCP-GE), to simulate multilingual cocktail party mixtures of two and four speakers with German as the target language. In total, 21 interfering languages are covered and individually mixed with German.

3. We perform a comprehensive study based on three model architectures and show that TLE works regardless of the num-
number of speakers and generalizes well to unseen speakers and languages. The achieved performances validate our design concept and provide first-time baselines for the TLE task.

The rest of the paper is organized as follows. In Section 2, we will introduce the TLE task more in detail and describe our method. In Section 3, we will specify the design of the GlobalPhoneMCP-GE database, followed by Section 4, in which we will explain and perform the experiments. We will discuss the results in Section 5 and conclude the study in Section 6. We will make all scripts and information publicly available\(^1\).

2. TARGET LANGUAGE EXTRACTION

In target speaker extraction, a reference signal of the desired speaker is used to encode this speaker’s specific voice characteristics into an embedding. The embedding guides a speech extraction model to filter only the target speaker’s voice out of a speech mixture of multiple sounds. While extracting one target speaker at a time, this mechanism is able to work for the extraction of multiple speakers according to the given reference signals as shown in Figure 1 (top).

The task of TLE is similar to that of target speaker extraction, i.e. the latter extracts all spoken contributions of a specific speaker, while the former extracts all speakers actively talking in a certain target language, independent of the number of speakers. We could consider to present a language embedding, which provides the target language specific characteristics such as tonal features or phonotactics, to TLE just like a speaker embedding to the target speaker extraction model. This language embedding could be used to guide the speech separation module to extract all target language speakers. That means, the predicted mask is not based on speaker characteristics but on language characteristics. In our case study, we are only interested in extracting a pre-defined target language in an arbitrary multilingual cocktail party scene.

We have two reasons to believe that for TLE it would be suboptimal to simply providing a language embedding in the same way as speaker embedding in speaker extraction. First, the number of target languages is rather limited in most applications since a typical user does not understand more than a handful languages. In this paper we build one TLE model for German, which mimics the case of a user who only speaks and understands German as the mother tongue. Second, in typical multiparty scenarios pockets of multiple speakers talk to each other in the same language. In this case, a language embedding is not the best to represent the overlapping multi-talk. We believe that a more natural method would be to train a model which can learn those characteristics and embed the information into its entire architecture and parameters. By doing this, we tailor the model to a specific language. The resulting model is ready to be plugged into any multilingual soundscape to extract all voices from the speakers who talk in the target language. Figure 1 (bottom) illustrates the proposed target language extraction concept.

The concept of having a mixture as target signal differs from the classical target speaker extraction approach in which the extracted signal contains only a single source. Until now, speaker extraction studies paid less attention on those scenarios. We are aware of only one work by Zhang et al. [36] in which a mixture signals is used as second supervision to improve the actual target signal extraction.

3. A NEW CORPUS FOR TLE: GLOBALPHONEMCP-GE

In our experiments, we covered only speech based signals and modeled two different multilingual cocktail party scenarios. The first scenario represents an easier task with one target language speaker and one interfering language speaker (1T-1I). The second one is more difficult but realistic, that consists of two target language speakers and two interfering language speakers (2T-2I).

The creation of the database was based on the audio data of the GlobalPhone 2000 Speaker Package\(^2\), which is a subset of the well-known multilingual GlobalPhone corpus proposed by Schultz [40] and Schultz et al. [41]. The GlobalPhone 2000 Speaker Package covers speech of 2000 native speakers from 22 different languages, recorded over several years in controlled environments. The amounts of speakers and utterances are slightly different among the languages but each individual speaker is given with roughly 40 seconds of total speech duration. The 22 different languages are represented by Arabic (AR), Bulgarian (BG), Chinese Mandarin (CH), Chinese Shanghai (WU), Croatian (CR), Czech (CZ), French (FR), German (GE), Hausa (HA), Japanese (JA), Korean (KO), Polish (PL), Portuguese (PO), Russian (RU), Spanish (SP), Swahili (SA), Swedish (SW), Tamil (TA), Thai (TH), Turkish (TU), Ukrainian (UA), and Vietnamese (VN). The European Language Resources Association (ELRA) [42] distributes the GlobalPhone 2000 Speaker Package under research and commercial licenses.

We established a new first of its kind corpus design concept, called “GlobalPhone Multilingual Cocktail Party” (GlobalPhoneMCP), which simulates the multilingual cocktail party in mixtures of two and four speakers. Each mixture contains a target language and an interfering language. All speakers in the database are native speakers. For our study, we created a database with German as fixed target language for all mixtures, while the interfering language varies between 21 different languages. We call this database “GlobalPhone Multilingual Cocktail Party - German” (GlobalPhoneMCP-GE).

The database design followed a strong belief, that a deep neu-

\(^1\)https://github.com/mborsdorf/TargetLanguageExtraction

\(^2\)https://catalog.elra.info/en-us/repository/browse/ELRA-S0400/
eral network is able to learn language characteristics such as tonal features or phonotactics directly from the data and can embed those information into its entire architecture and parameters. Therefore, no additional reference signals were considered. We wrapped this into a supervised learning scheme which led to tuples of input mixtures and target output mixtures. While the former comprised speech in the target and interfering languages, the latter contained the speech in the target language only (ground truth). We will describe the mixing details in the following.

Inspired by the work of Hershey et al. [2], we decided on a fixed number of utterances for the training, cross-validation and test sets for each language mixture dataset. We defined the total amounts of training, cross-validation, and testing data over all language mixture datasets to roughly match 30, 20, and 15 hours respectively.

For each language, we randomly selected 25 % of the speakers and assigned them to the test set. The remaining speakers were assigned to the training/cross-validation set. In this set, we performed another split to assign 80 % of the data to the training set and the rest to the cross-validation set. While the training and cross-validation sets share the same speakers but with different utterances, the test set speakers are completely disjoint from these sets (open set test condition).

We repeated this for all language mixture datasets. Afterwards, we sampled random signal-to-noise ratios (SNRs) from a uniform distribution within an interval of [-5 to 5] dB to randomly select either the target language speaker or the interfering language speaker to be the foreground speaker. In order to ensure a zero mean value, we applied \( +\frac{SNR}{2} \) to the target speaker and \( -\frac{SNR}{2} \) to the interfering speaker.

Next, we created speaker pair lists for the four talker (2T-2I) scenario. To simulate the target language speaker pairs, we combined each target utterance with a random second target utterance of another random target speaker. We performed this for the training, cross-validation, and test sets. We also made sure to avoid duplicates of combinations. We sampled random SNRs within a range of [-5 to 5] dB for those speaker pairs and applied them in the same manner as for the 1T-1I scenario. To create the speaker pairs for the interfering language, we randomly combined two utterances of different speakers until we matched the amount of target language pairs. We made sure to use all utterances at least once without duplicates in combinations. We balanced the utterances as far as possible. Afterwards, we assigned random SNR values in a range of [-5 to 5] dB to the speaker pairs, following the same method as for the target language speaker pairs. Finally, to create the four talker mixtures, we randomly combined the target language speaker pairs with the interfering language speaker pairs. We sampled random signal-to-noise ratios to define the fore- and background proportions in the same way as before. After this, the final SNRs between two speakers, within a 2T-2I scene, ranged in an interval of [-10 to 10] dB. The described mixing concept was applied to all language mixture datasets. We highlight that the 2T-2I scenario is more difficult than the 1T-1I scenario because of four speakers in the mixture and due to the possible larger negative SNRs. The latter could lead to much more prominent interfering language speakers, compared to the 1T-1I scenario.

The speaker pair lists were used to create the audio mixtures.

---

Fig. 2: Statistics of the GlobalPhoneMCP-GE database. The total audio durations for the training (tr), cross-validation (cv), and test (tt) sets for each language mixture dataset are given for 1T-1I (a) and 2T-2I (b) scenes. Both scenes involve the same amount of speakers, as reported in (c). While the training and cross-validation sets share the same speakers (denoted as tr-cv), the test set (tt) speakers are different.
We revised the scripts released by Isik et al. [43] for our purposes. Originally, those scripts were developed to create the wsj0-2mix and wsj0-3mix databases. For our experiments, we simulated the “min” version because we were interested in mixtures with highly overlapped speech. In this version, the longer utterance was truncated to the duration of the shorter one. We chose 8 kHz as sampling frequency since this has been frequently used in the research and to reduce computational costs during our experiments. The mixing scripts can also be used to create a database version with 16 kHz as sampling frequency or to simulate a “max” version in which the shorter utterance is padded with zeros to match the length of the longer utterance.

The final GlobalPhoneMCP-GE database comprises 30.45, 20.33, and 15.35 hours (1T-1I) and 29.41, 19.61, and 14.81 hours (2T-2I) of total audio data for training, cross-validation, and test sets respectively. Both scenarios involve the same amount of total utterances for training, cross-validation, and testing with 14,868, 9,912, and 7,560 respectively. The statistics for the individual language mixture datasets are illustrated in Figure 2.

4. EXPERIMENTAL SETUP

We conducted experiments considering three different network architectures in order to investigate the novel TLE task. To have meaningful results, we evaluated the trained networks on open set conditions for speakers and languages. The experiments have been implemented in Python using PyTorch as framework.

4.1. Network architectures

To perform the TLE task, we revise three existing network architectures, namely SpEx+ [34], Conv-TasNet [5], and SepFormer [9], by applying our design concept described in Section 2. All networks work in the time domain and the encoder-decoder pipelines are trained in a data driven manner. That means they learn the transformation to and from a latent representation directly from the data distribution and their characteristics. In contrast to speech separation solutions which explicitly use Fourier analysis, as e.g. in [44, 45], this method does not suffer from reconstruction errors due to inexact phase information. Additionally, it allows to easily embed a typical time domain metric, such as the signal-to-noise ratio (SNR), as the training objective. We now introduce the TLE network architectures.

4.1.1. SpEx+Ref-Free

The first TLE model was revised from the SpEx+ architecture\(^3\) which has been recently proposed by Ge et al. [34] to tackle the task of speaker extraction in the time domain. Since our TLE method does not require any additional reference signal, we removed the speaker encoder part of the SpEx+ architecture. The remaining structure of the model followed the original implementation by Ge et al. [34]. The adjusted model comprised 9.07 million trainable parameters in total. In our experiments, we refer this version as SpEx+Ref-Free.

4.1.2. Conv-TasNetSingle-Mask

The second TLE model was revised from the well-established Conv-TasNet, proposed by Luo and Mesgarani [5]. This model is typically used in blind source separation tasks and therefore estimates two or more masks to separate multiple sources in parallel. Since our TLE approach extracts a single target signal (mixture of one or multiple sources), we adapted the structure to estimate a single mask. Additionally, we changed the mask activation function to rectified linear unit (ReLU) since this led to better results compared to Sigmoid as activation function. The rest of the model was implemented according to the settings of the best model described in Luo and Mesgarani [5]. We refer this model as Conv-TasNetSingle-Mask with a total number of 4.98 million trainable parameters. To implement the model, we used “Asteroid” [46], a recently published toolkit for speech separation experiments\(^4\).

4.1.3. SepFormerSingle-Mask

The third TLE model was revised from the SepFormer architecture, recently introduced by Subakan et al. [9] for blind source separation tasks. The SepFormer achieves state-of-the-art speech separation performance on the widely studied wsj0-2mix dataset and on its extended three-speaker version, called wsj0-3mix. Both datasets were introduced with the deep clustering approach by Hershey et al. [2] and are based on the Wall Street Journal (WSJ0) corpus. This model represents the only transformer based architecture in our experiments. Similar to the Conv-TasNetSingle-Mask model, we modified the SepFormer to only estimate one target mask. Apart from that, the architecture followed the settings described in Subakan et al. [9]. We denoted our revised version for TLE as SepFormerSingle-Mask, which comprised 25.61 million trainable parameters in total. In our study we did not apply the concept of dynamic mixing, which was introduced with Wavesplit by Zeghidour and Grangier [8] and also applied to the SepFormer approach in Subakan et al. [9]. The model implementation was done with help of the recently proposed “SpeechBrain” [47] speech toolkit\(^5\).

4.2. Training and evaluation

In our experiments, we created two model instances of each network architecture. While we trained the first instances on 1T-1I only, we pre-trained the second instances on 1T-1I followed by a fine-tuning on both 1T-1I and 2T-2I altogether. This led to six different models. To investigate not only open set conditions for speakers but also for languages, we removed three random language mixture datasets from the training data composition, namely GE-FR, GE-AR, and GE-TH. Those language mixtures were only part of the test data. During the training, we split the data into chunks before feeding them to the models. The chunk size was defined as 32,000 samples, which reflects a duration of four seconds (since the data has been created as 8 kHz sampled version). Chunks with less than 32,000 samples were padded with zeros, if they contained at least 16,000 samples. Shorter chunks were discarded. In the evaluation, we processed the entire input utterance at once, independent of its length. We set the training parameters for the different models according to the descriptions in the respective paper and the given basis implementation (see Section 4.1). We will specify those in the following.

We applied Adam [48] as optimizer with an initial learning rate of 1e\(^{-3}\) for the SpEx+Ref-Free and Conv-TasNetSingle-Mask based models and 1.5e\(^{-4}\) for the models which utilized the SepFormerSingle-Mask architecture. We applied the same initial learning rates for the fine-tuning of the models since we used 1T-1I and 2T-2I altogether. The learning rates were scheduled and halved after three consecutive epochs without loss decrease. The minimum learning rate was defined as 1e\(^{-8}\). We used gradient clipping with a norm of five

\(^3\)https://github.com/gemengtju/SpEx_Plus

\(^4\)https://github.com/asteroid-team/asteroid

\(^5\)https://github.com/speechbrain/speechbrain
for the Conv-TasNet\textsubscript{Single-Mask} and the SepFormer\textsubscript{Single-Mask} models. A weight decay of $10^{-3}$ was utilized for the SpEx+Ref-Free based models.

We trained the SpEx+Ref-Free models for a maximum of 150 epochs and the Conv-TasNet\textsubscript{Single-Mask} and SepFormer\textsubscript{Single-Mask} models for a maximum of 200 epochs. An early stopping algorithm stopped the training if the validation loss did not decrease within six subsequent epochs. The batch sizes were set to eight, eight, and two for the SpEx+Ref-Free, Conv-TasNet\textsubscript{Single-Mask}, and SepFormer\textsubscript{Single-Mask} models respectively.

The scale-invariant signal-to-distortion ratio (SI-SDR), proposed by Le Roux et al. [49], was used to train and evaluate the three TLE models. The SI-SDR is given as

$$\text{SI-SDR} = 20 \log_{10} \left( \frac{\| \hat{x}^T s \|_2^2}{\| x^T s - \hat{\delta} \|_2^2} \right)$$

where $s$ denotes the ground truth signal and $\hat{s}$ represents the reconstructed target signal. The objective function to train the models has been derived by adding a small stabilization value $\varepsilon$ to prevent the equation from breaking. This led to:

$$\text{SI-SDR} \approx 20 \log_{10} \left( \frac{\| \hat{x}^T s \|_2^2 + \varepsilon}{\| x^T s - \hat{\delta} \|_2^2 + \varepsilon} \right)$$

In our experiments, we defined $\varepsilon = 10^{-8}$. To determine the final target language extraction quality for all 21 language mixture datasets in 1T-1I and 2T-2I scenes, we performed three steps. First, we calculated the mean SI-SDR for all extracted target signals of a given test set. Second, we repeated this for the unprocessed input mixtures of the test set. Third, we calculated the difference between those mean SI-SDRs which led to the SI-SDR improvement (dB) on the respective test set. The target language extraction quality was evaluated on the entire length of an utterance, following Eq. 1.

### 5. RESULTS AND DISCUSSION

We trained six models and evaluated their target language extraction performances on both 1T-1I and 2T-2I scenes, as described in Section 4.2. The mean performances over all language mixture datasets for 1T-1I and 2T-2I scenes are illustrated in Table 1. Figure 3a shows that models trained on 1T-1I demonstrate good extraction performances on almost all 1T-1I test sets (matching conditions). We identify worse extraction performances on GE-CZ, GE-HA, GE-PL, and GE-SA test sets. From the statistical perspective, those sets do not have any irregularities in the amount of speakers, utterances or input SNRs. We also cannot identify any relations between the performance drops and phonetic similarities between the target language and the interfering languages. We assume that the degradation may be based on specific differences in the recording conditions. In general, we find that the TLE models work for unseen speakers, since the speakers in the test sets are disjoint from the training sets. Additionally, they show high performances on test sets with unknown interfering languages, namely GE-FR, GE-TA, and GE-TH. While SpEx+Ref-Free and Conv-TasNet\textsubscript{Single-Mask} are mostly on pair, the SepFormer\textsubscript{Single-Mask} shows superior performance on all test sets within 1T-1I. This proofs our expectation, because the baseline SepFormer model shows new state-of-the-art performances on the wsj0-2mix and wsj0-3mix benchmark datasets, as described in Section 4.1.

Testing those models on mismatching four talker conditions (2T-2I) reveals poor extraction performances, as depicted in Figure 3b. However, the behavior of the performances across the individual language mixture datasets stays similar to Figure 3a. It is worth noting that the 2T-2I scene is not only more difficult because of four speakers in the mixtures but also due to the possible larger negative SNRs (see Section 3). Surprisingly, the Conv-TasNet\textsubscript{Single-Mask} mostly performs better than the other two models although it has by far the fewest amount of trainable parameters. This finding is also stated in Table 1. We hypothesize that this may be based on the way in which the Conv-TasNet\textsubscript{Single-Mask} constructs the target mask. Similar to the other two models, it mainly follows a feed forward concept with some encapsulated residual-connections and propagates the information through the following layers to the final mask estimation. However, the Conv-TasNet\textsubscript{Single-Mask} additionally applies skip-connections to combine all intermediate results with the final propagated output to estimate the mask. We believe that all those multi-level information may help to have a more robust TLE performance in mismatching conditions.

If the models are trained on both 1T-1I and 2T-2I scenes and tested on 1T-1I (see Figure 3c), the extraction performances slightly degrade compared to Figure 3a. We identify that the performance degradation of Conv-TasNet\textsubscript{Single-Mask} is only marginally. This observation is also given in Table 1. We suspect that this is also an effect of the incorporation of multi-level information in the Conv-TasNet\textsubscript{Single-Mask}. This may help to preserve the performance on the source task after the model has been adapted to a second task. Figure 3d illustrates high performance improvements on the 2T-2I test sets if the models have also been trained on 2T-2I scenes (matching conditions).

Comparing the Figures 2a-d and the results in Table 1, we see that models trained on 1T-1I and 2T-2I can preserve the performance on 1T-1I, compared to 1T-1I only models, while they strongly improve their performances on 2T-2I. This applies to all models although the SepFormer\textsubscript{Single-Mask} provides the best overall performance. Since those models can be plugged into both scenes, they work independent of the number of speakers. Additionally, they are robust to unseen interfering languages in all test scenarios. We observe that the performance on the 2T-2I scene is generally lower than on the 1T-1I scene. Finally, the results confirm that all TLE models can learn the target language’s characteristics and are capable of extracting the voice based on the language without any additional information of the language or speaker.

### 6. CONCLUSION AND FUTURE WORK

In this paper, we introduced a novel task called “target language extraction” (TLE) which attempts to extract voices based on the spoken language and regardless of the number of speakers. To simulate a “multilingual cocktail party” scenario, we proposed a new first of its kind database, called “GlobalPhone Multilingual Cocktail Party - German” (GlobalPhoneMCP-GE), which covers multiparty scenes with two and four speakers. Each scene comprises German as the target language and another interfering language. In total, scenes for 21 different interfering languages are provided. To investigate the feasibility of TLE, we studied three different model architectures under matching and mismatching conditions and applied open-set conditions for speakers and languages. The experimental results show that all TLE models can learn the target language’s characteristics...
Fig. 3: Target language extraction results for different models under matching and mismatching conditions. The results are reported in terms of SI-SDR improvement (dB) as mean value over all test samples for each language mixture dataset. The stars (*) indicate the language mixture datasets which have been left out during training and are only part of the evaluation to simulate open-set conditions for languages. Three models were trained on 1T-1I only and tested under matching conditions (a) and mismatching conditions (b). Three different models were trained on both 1T-1I and 2T-2I and tested under matching conditions only (c-d).

Table 1: Mean target language extraction results over all language mixture datasets for 1T-1I and 2T-2I test sets respectively. Additionally, the average performances over all test sets (1T-1I and 2T-2I altogether) are given. The results are reported in terms of SI-SDR improvement (dB).

<table>
<thead>
<tr>
<th>TLE Model</th>
<th>Trained on</th>
<th>SI-SDRi (dB)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1T-1I</td>
<td>2T-2I</td>
<td>1T-1I</td>
<td>2T-2I</td>
<td>Average</td>
</tr>
<tr>
<td>SpEx^Ref.-Free</td>
<td>✓</td>
<td>✓</td>
<td>15.04</td>
<td>4.98</td>
<td>10.01</td>
</tr>
<tr>
<td>Conv-TasNet_{Single-Mask}</td>
<td>✓</td>
<td>✓</td>
<td>14.41</td>
<td>10.07</td>
<td>12.24</td>
</tr>
<tr>
<td>SepFormer_{Single-Mask}</td>
<td>✓</td>
<td>✓</td>
<td>14.90</td>
<td>5.32</td>
<td>10.11</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>14.84</td>
<td>10.00</td>
<td>12.42</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>16.63</td>
<td>4.90</td>
<td>10.77</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>15.80</td>
<td>11.11</td>
<td>13.46</td>
</tr>
</tbody>
</table>

and are capable of extracting the voice based on the language without any additional information. The models perform this regardless of the number of speakers and generalize well to unseen speakers and interfering languages. Finally, our results confirm the proposed TLE method and also validate the design of the GlobalPhoneMCP-GE database.

In our future work, we will investigate more realistic multilingual cocktail party situations. This will include more than two active speakers per language and more than two languages within a scene. Moreover, a typical real-world soundscape does not only comprise speech but also other everyday sounds, such as music or sounds from animals, which we will consider in future experiments.

7. REFERENCES


