INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

GlobalPhone Mix-to-Separate out of 2: A Multilingual 2000 Speakers Mixtures Database for Speech Separation

Marvin Borsdorf [1], *Chenglin Xu*[2], *Haizhou Li*[2,1], *Tanja Schultz*[3]

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

marvin.borsdorf@uni-bremen.de

Abstract

Monaural speech separation has been well studied on various databases. However, these databases mostly concern English speech. Research in multi-speaker scenarios, such as speech recognition, speaker recognition, speaker diarization, and speech separation calls for speaker mixtures databases comprising multiple languages. In this paper, we propose a new extensive multilingual database for speech separation tasks derived from the GlobalPhone 2000 Speaker Package, called "GlobalPhone Mix-to-Separate out of 2" (GlobalPhoneMS2). We describe the construction of the database and conduct speech separation experiments in monolingual and multilingual as well as seen and unseen languages settings. When trained on a multilingual dataset, the networks improve their performances for unseen languages, and across almost all seen languages. We show that replacing a monolingual dataset with a trilingual one, while keeping the data size roughly the same, helps to improve the performance in most cases. We attribute this to a larger diversity in speech, language, speaker, and recording characteristics. Based on the GlobalPhoneMS2 database, speech separation results for two-speaker mixing scenarios are reported in 22 spoken languages for the first time.

Index Terms: GlobalPhone, speech separation, cocktail party problem, multilingual, crosslingual

1. Introduction

Humans are able to focus and follow on the desired (target) speaker's voice within a conversation between multiple speakers. This scenario was early described and introduced by Cherry [1] and is referred to as the "cocktail party problem". While the cocktail party problem is not difficult for humans, it is still an unsolved problem for machines. In recent years, the interplay of high-performing computational capacities and machine learning advancements has provided the basis to apply deep learning mechanisms to work on the solution of this challenging problem.

Traditional techniques and studies on auditory scene analysis contributed to the progress of monaural speech separation. Deep clustering introduced by Hershey et al. [2] represented a novel method to address the cocktail party problem. Further experiments on this approach were performed by Isik et al. [3]. Recently, new training procedures such as permutation invariant training (PIT) introduced by Yu et al. [4] and new network architectures such as Conv-TasNet by Luo and Mesgarani [5] have advanced the state-of-the-art in speech separation.

The scientific work requires suitable databases and many speech separation databases have been released over time. In the following we highlight a few of them: Hershey et al. [2] introduced the wsj0-2mix database which was based on the Wall Street Journal (WSJ0) corpus and contained clean speaker mixtures in English language. A couple of years later, wsj0-2mix was extended with ambient noise by Wichern et al. [6] and published as WHAM! database. The noise was collected in real-world scenarios such as in bars or restaurants and used as background audio in the speaker mixtures. Maciejewski et al. [7] proposed WHAMR!, an extended version of WHAM!, which also included reverberation for an even more realistic scenario.

Databases from other research areas have also been adapted for speech separation scenarios. LibriSpeech by Panayotov et al. [8] was released as an automatic speech recognition (ASR) corpus in English language, but later used to create LibriCSS by Chen et al. [9]. LibriMix introduced by Cosentino et al. [10] combined the speech data of LibriSpeech with the background noise data of WHAM! to provide a large amount of noisy speech mixtures. Further studies combined various databases to evaluate the robustness of speech separation systems as for example in Kadıoğlu et al. [11] and Maciejewski et al. [12].

While numerous databases for speech separation tasks have been developed, they mostly comprise a single language. We are aware of only one study by Appeltans et al. [13], in which the authors investigated speech separation in monolingual and multilingual setups. They studied two-speaker mixtures on seen and unseen languages for eight different languages. To create the two-speaker mixtures, they used the GlobalPhone corpus which was introduced by Schultz [14] and Schultz et al. [15].

Multilingual acoustic modeling research has a long history since its debut by Schultz and Waibel [16], that was further studied by Schultz and Waibel [17, 18], Gretter [19], Tachbelie et al. [20] and Pratap et al. [21]. Vu et al. [22] proposed a multilingual speech recognition system that can handle code-switching in conversational speech. Seki et al. [23] combined both a multilingual ASR system and a multi-speaker ASR system to one single end-to-end system, which is able to recognize the speech of two speakers talking at the same time in different languages.

So far, studies on multi-speaker speech recognition relied on databases, which provided both the mixed-speakers and the single-speaker soundtracks. Unfortunately, such information are not available in typical everyday scenarios such as at a cocktail party or in other tasks that involve speech separation and unseen languages. Consequently, there is a clear demand for multilingual speaker-mixed speech databases or mixing scripts for widespread multilingual single-speaker speech databases. Such data will foster the development of robust speech separation systems in the context of multi- and crosslingual setups.

In this work, we present a first of its kind multilingual database for speech separation, referred to as "GlobalPhone Mix-to-Separate out of 2" or in short "GlobalPhoneMS2". This database contains two-speaker mixtures produced in 22 lan-

http://dx.doi.org/10.21437/Interspeech.2021-1552

guages with a total number of 2000 different speakers. In the following we will describe the design and the features of GlobalPhoneMS2 in detail. All mixing scripts and information will be made publicly available[1]. In Section 2, we briefly describe the GlobalPhone 2000 Speaker Package followed by Section 3, which specifies the design of GlobalPhoneMS2. In Section 4, we conduct a set of monolingual and multilingual experiments. We discuss the experimental results in Section 5, and conclude the study in Section 6.

2. GlobalPhone 2000 Speaker Package

The GlobalPhone corpus proposed by Schultz [14] and Schultz et al. [15] is a multilingual database which contains speech and text sets of different languages from all over the world. It was collected over several years and used in numerous published studies. The GlobalPhone 2000 Speaker Package[2] is a subset of the GlobalPhone corpus and aims to provide a low-cost but high-quality multilingual speech database to the community. The database contains speech of 22 languages, namely 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 numbers of speakers and utterances vary slightly for each language and sum to a total of 2000 native speakers. Each speaker is represented by a total duration of roughly 40 seconds.

The database provides the speech data, the transcriptions, and the speaker and recording information. It is applicable to a variety of tasks and to speaker mixing for multi-talker speech in particular. The GlobalPhone 2000 Speaker Package is distributed by the European Language Resources Association (ELRA) [24] under research and commercial licenses.

3. GlobalPhone Mix-to-Separate Database

The creation of GlobalPhoneMS2 required to balance the size and richness of data with computational effort. If we would have used each utterance in a language only once, this would have led to a very small dataset. However, if we would have combined each utterance with all other speakers' utterances, we would have faced a combinatorial problem with tremendous durations for data simulation, training, and evaluation. Hershey et al. [2] used a split of 20000, 5000, and 3000 utterances which led to 30, 10, and 5 hours for training, cross-validation, and test data respectively. We decided to adopt this strategy and set the amounts of training and cross-validation data to 30 and 10 hours respectively and the test set size to a fixed number of 3000 utterances. We describe the exact procedure in the following.

Firstly, we randomly assigned 25 % of the speakers to the test set and the rest to the training/cross-validation set. The training/cross-validation data were split into 80 % for training and 20 % for cross-validation. We took care that training and cross-validation sets share the same speakers but with different utterances. The test set is disjoint from these sets (open set condition). Depending on the available numbers of speakers and utterances, we had to change the portions of the three sets to primarily match 3000 test utterances and secondary balance the training and cross-validation sets. Thus, for Chinese Shanghai and Tamil, there were not enough data left to match the defined training and cross-validation set sizes.

Secondly, we created speaker pair lists for each set. For GlobalPhoneMS2, only two-speaker mixtures have been considered. For each set we combined each utterance with all other speakers' utterances. Speaker pairs were sampled from these lists to create the final training, cross-validation and test sets. The algorithm sampled the utterances in a balanced way as long as enough speaker pairs of each speaker were available. Afterwards, additional scripts checked whether the speaker pairs were properly created. In the end, we randomly sampled signal-to-noise ratios (SNRs) from a uniform distribution within an interval of [0 - 5] dB for all speaker pairs. This formed the foreground and background speaker proportion.

Thirdly, to create the final speaker mixtures, we revised the scripts released by Isik et al. [25]. The scripts were originally developed to generate the wsj0-2 speaker mixtures to investigate the deep clustering method by Hershey et al. [2]. We focused on scenarios with highly overlapped speech and therefore simulated the so called "min" version. In other words, the longer utterance was truncated to the length of the shorter one when two utterances were mixed. We selected 8 kHz as sampling frequency because it has been frequently used by researchers and in addition we wanted to have the same information density as for wsj0-2mix (min, 8 kHz). The scripts can also be used to create a "max" version in which the shorter utterances is padded with zeros to match the length of the longer utterance as well as to set the sampling frequency to 16 kHz. This leads to four different combinations: (min, 8 kHz), (max, 8 kHz), (min, 16 kHz), and (max, 16 kHz). The statistics of the final GlobalPhoneMS2 corpus are reported in Table 1.

4. Speech Separation Experiments

4.1. Network architecture

As a case study, we trained a neural speech separation network on the GlobalPhoneMS2 dataset, following the Conv-TasNet architecture by Luo and Mesgarani [5]. The Conv-TasNet works in the time domain, that does not require any frequency analysis or transformation in the process. The encoder that processes the input speech and the decoder that creates the output speech are trained together with the source separation task. They "learn" to project the data to and from an appropriate latent space. The separation part extracts features out of the projected input and calculates a mask for each speaker. The masks are applied on the encoder's output. The masked projections are decoded in order to reconstruct the separated speech waveforms. Since Conv-TasNet is a time domain implementation, the phase information of the input mixture is not required to reconstruct the waveforms. In a way this reduces the reconstruction error compared to speech separation solutions implemented in frequency domain as e. g. in Isik et al. [3]. Finally, a signal metric such as signal-to-noise ratio (SNR) can easily be embedded as training objective to optimize the network based on the reconstructed target's signal quality. We implemented the Conv-TasNet architecture according to the best model's settings described in Luo and Mesgarani [5]. We only changed the mask activation function from Sigmoid to rectified linear unit (ReLU). The model comprised 5.1 million trainable parameters in total.

The code in our experiments was implemented in Python using PyTorch as framework. We utilized "Asteroid", a toolkit recently proposed by Pariente et al. [26], to design the data scheduling, the training procedure, and the evaluation step.

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

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

Table 1: *Statistics of the GlobalPhoneMS2 and the wsj0-2mix (WSJ) databases. For each language, denoted by its language code (LC), the amount of speakers (# Spks), the amount of utterances (# Utts), and the audio duration in hours (hrs) for training (tr), cross-validation (cv), and test (tt) sets are given. tr and cv sets share the same speakers, denoted as tr-cv under # Spks, but contain different utterances, thus denoted as tr and cv separately under # Utts. The audio duration measures the total length including speech and silence.*

LC	# Spks	# Utts	Audio (hrs)
	tr-cv / tt	tr / cv / tt	tr / cv / tt
AR	60 / 18	14285 / 4713 / 3000	30 / 10 / 6.63
BG	60 / 17	13758 / 4577 / 3000	30 / 10 / 6.58
CH	85 / 25	13816 / 4621 / 3000	30 / 10 / 6.42
WU	25 / 16	2772 / 1152 / 3000	5.95 / 2.48 / 6.27
CR	71 / 21	13477 / 4566 / 3000	30 / 10 / 6.52
CZ	79 / 23	13618 / 4561 / 3000	30 / 10 / 6.62
FR	77 / 23	13532 / 4517 / 3000	30 / 10 / 6.62
GE	59 / 18	13605 / 4533 / 3000	30 / 10 / 6.56
HA	80 / 23	30716 / 10025 / 3000	30 / 10 / 3.17
JA	85 / 25	13668 / 4575 / 3000	30 / 10 / 6.47
KO	77 / 23	13717 / 4568 / 3000	30 / 10 / 6.58
PL	77 / 22	13736 / 4556 / 3000	30 / 10 / 6.54
PO	77 / 23	13845 / 4578 / 3000	30 / 10 / 6.51
RU	81 / 24	13697 / 4572 / 3000	30 / 10 / 6.50
SP	78 / 22	14199 / 4897 / 3000	30 / 10 / 6.43
SA	54 / 16	19434 / 6364 / 3000	30 / 10 / 4.39
SW	76 / 22	13717 / 4567 / 3000	30 / 10 / 6.46
TA	13 / 16	737 / 222 / 3000	1.50 / 0.42 / 5.48
TH	76 / 22	13751 / 4588 / 3000	30 / 10 / 6.57
TU	78 / 22	13874 / 4629 / 3000	30 / 10 / 6.42
UA	85 / 25	23294 / 7702 / 3000	30 / 10 / 3.97
VN	78 / 23	19899 / 6496 / 3000	30 / 10 / 4.08
WSJ	101 / 18	20000 / 5000 / 3000	30 / 10 / 5

4.2. Training and evaluation

To introduce language diversity into our experiments, we opted for both tonal and non-tonal languages which are widely spoken in different continents and cover some of the major language families. We have chosen Chinese Mandarin, Hausa, Brazilian Portuguese, Polish, and German. In addition, we used wsj0-2mix (min version and 8 kHz sampling rate) to include English.

The networks were trained for a maximum of 200 epochs including an early stopping. An initial learning rate of $1e^{-3}$ was combined with a scheduler to reduce the learning rate on plateaus. Adam was used as optimizer to update the networks' weights based on the loss function which was defined as scale-invariant signal-to-distortion ratio (SI-SDR) proposed by Le Roux et al. [27]. We applied permutation invariant training (PIT) according to Yu et al. [4]. Six monolingual models were trained each on one single language. We also trained three multilingual models by sharing the training data across languages. For all models we used the same settings, as specified above. During training, we set the input segment length to four seconds except for the networks which were trained on Hausa (monolingual and multilingual). As the utterances of Hausa are very short, we had to change the input segment length to two seconds. According to the available GPU memory and the input segment length, we chose a batch size of four, eight or sixteen.

Based on the results of the monolingual GlobalPhoneMS2 models (see Section 5) we combined the training languages of the worst and the two best models. Thus, the first multilingual dataset comprised German, Chinese Mandarin, and Portuguese. This dataset only contained 33.33 % of the amount of data of each language (denoted as 33 % model) to match the dataset sizes of the monolingual counterparts. The idea was to study the influence of having more languages while the total amount of training data stays similar. The second multilingual dataset covered the full amount of data of the three languages to investigate the impact of an increase in training data while preserving the number of languages. The third multilingual dataset additionally included the languages Hausa and Polish (full amount of data) to further analyze the effects of a growing number of languages and an increase of the total amount of data.

We evaluated the speech reconstruction quality on the entire length of each test utterance. Each model was tested on all 23 languages (GlobalPhoneMS2 and wsj0-2mix). We determined the quality of a reconstructed speech signal by calculating the SI-SDR (dB). The overall separation performance on each test set was reported in terms of SI-SDR improvement (dB), which is defined as the difference between the SI-SDR of a separated signal and the SI-SDR of the signal before the separation. A high value means the reconstructed speech waveform contains less distortions and is closer to the target's clean speech (ground truth). A small value identifies worse separation performance.

5. Results and Discussion

We trained six monolingual and three multilingual speech separation networks and evaluated them on all 23 test sets. We individually list the results for those languages on which we have also trained monolingual models, while the rest is reported as mean SI-SDR improvement over 17 unseen test sets (eval_17). With eval_all we report the mean speech separation performance of each network over all 23 languages (GlobalPhoneMS2 and wsj0-2mix together). Table 2 shows the results.

5.1. Monolingual GlobalPhoneMS2 models

The monolingual models perform well on the respective seen training language, as indicated by the bold diagonal, but occasionally sharply degrade in performance when tested on unseen languages. Some results show negative SI-SDR values, i. e. the model cannot handle the audio data of the specific test language and corrupts the speech instead. We mainly attribute this to different speech, language, and speaker characteristics between training and test data as reflected e. g. in the mismatching sound systems, tonal features, and phonotactics.

While the GlobalPhone data collection followed a well determined setup to minimize differences between the recording sessions, there might be varying environmental noise artifacts within languages, and recording variations across languages. Such artifacts could adversely affect the neural network since its components learn from training data in order to generalize to unseen data. According to Table 1 and Table 2, the performance degradation does not seem to be linked to the number of speakers since models which have been trained on less speakers than other models do not always perform worse and vice versa.

5.2. Multilingual GlobalPhoneMS2 models

The lower part of Table 2 below the dashed line shows that the multilingual models are able to outperform their monolingual counterparts on all conditions, i.e. on seen and unseen languages. This is observed partly for the CH+GE+PO (33 %) model and fully for the CH+GE+HA+PL+PO (full) model. We find that the CH+GE+PO (33 %) model, replacing a monolin-

Table 2: *Speech separation results for monolingual and multilingual (above and below dashed line) trained speech separation networks evaluated on seen and unseen languages. The results are reported in SI-SDR improvement (dB). All training languages are individually listed. The results for the remaining 17 other languages of GlobalPhoneMS2 are summarized as mean SI-SDR improvement (eval_17). eval_all provides the mean SI-SDR improvement of a trained network over all 23 languages (GlobalPhoneMS2 and wsj0-2mix).*

Test Train	CH	GE	HA	PL	PO	wsj0-2mix	eval_17 (mean)	eval_all (mean)
CH	**14.36**	-1.112	-0.3624	0.7931	12.55	9.951	8.044	7.519
GE	1.532	**12.06**	2.483	4.142	1.083	4.700	0.9965	1.867
HA	1.236	3.391	**9.747**	4.566	2.515	4.861	1.404	2.182
PL	4.722	7.125	4.021	**10.42**	6.616	9.457	4.325	5.038
PO	11.79	-1.215	-1.495	0.1551	**14.28**	9.804	8.063	7.408
wsj0-2mix	10.90	10.43	4.882	8.169	12.37	**16.22**	8.231	8.822
CH+GE+PO (33 %)	14.29	12.63	3.847	5.565	14.63	13.27	10.42	10.49
CH+GE+PO (full)	**15.61**	13.69	4.645	5.424	15.83	13.92	11.32	11.37
CH+GE+HA+PL+PO (full)	15.33	**13.97**	**11.64**	**12.40**	**15.87**	**15.60**	**11.81**	**12.42**

gual dataset with a trilingual dataset, improves the crosslingual performance while the total amount of training data is roughly preserved. This is not only shown in the individual results on each unseen language but also given by the mean value over 17 unseen languages denoted as eval_17. The results for eval_all also confirm that the multilingual models show better performances in average over seen and unseen languages.

We assume the superior performance of the multilingual networks can be primarily attributed to the additional variety in speech, language, and speaker information in the training data. These observations would corroborate the findings from other acoustic modeling studies by Schultz and Waibel [16, 17, 18]. The multilingual training also covers a broader diversity of channel and recording conditions which potentially help to enhance the robustness on unseen test data as shown e. g. by Maciejewski et al. [12] and Pandey and Wang [28]. In our experiments, the trainable encoder and decoder also benefit from this. An increase just in the amount of training data shows less performance improvements than the aforementioned considerations. The best multilingual model shows an average improvement of 3.7 dB over the best monolingual GlobalPhoneMS2 model (PO) and 3.6 dB over the monolingual wsj0-2mix model on unseen data depicted by eval_17.

Zegers and van Hamme [29] showed that long short-term memory (LSTM) based speech separation models seem to pay more attention to phoneme level and long-time speaker information than to phonotactic or lexical levels. Appeltans et al. [13] found related results and assumed that LSTM networks rather form a language independent modeling. Our experiments do not apply LSTMs and we use a trainable data driven encoder-decoder pair. This makes it hard to assess our model's attention to short- and long-time speech and language information compared to the aforementioned LSTM based results.

We are convinced that such information to some extent matter especially for monolingual models and multilingual models with just a little number of different languages. We also think that models exposed to multilingual training (and therefore to speakers of different languages) could form better speaker characterizations. Based on our results we feel that speech separation models could form a language independent modeling when trained under comprehensive multilingual conditions. This could equip a model with all the required knowledge about diversity in speech, language, and speaker information as well as recording characteristics to develop a robust and general modeling. Appeltans et al. [13] found that replacing parts of a monolingual dataset with other languages can degrade the performance on seen languages. We identify a performance degradation for the CH+GE+PO (33%) model when tested on CH and just little improvements when tested on GE and PO compared to the monolingual counterparts. However, a further increase in multilingual training data strengthens the model's performance on seen languages. Finally, in accordance with Appeltans et al. [13] we identified that multilingual training in most cases can help to improve the performance on unseen languages.

5.3. Monolingual wsj0-2mix model

Although the monolingual wsj0-2mix model is trained on a different corpus it shows a reasonable performance when applied to unseen languages. This is also indicated by eval_17 and eval_all. Mostly, the model can attain superior crosslingual performance over the other monolingual models. It also outperforms the trilingual models on HA and PL. The wsj0-2mix data was recorded in a clean, controlled and homogeneous setup and contains a slightly higher number of speakers. This could be an explanation for the acceptable results of the monolingual network when tested on unseen languages of a different corpus.

6. Conclusions

We introduced GlobalPhoneMS2 a new database of 22 languages for speech separation. To demonstrate its usefulness, we trained various monolingual and multilingual networks based on the Conv-TasNet architecture. Our experimental results on seen and unseen languages show that multilingual trained networks outperform monolingual ones on unseen languages and also denote performance improvements on almost all seen languages. In most cases this can be attained by a trilingual model trained with the same size of data. We therefore attribute the gains of the multilingual network to a higher robustness resulting from a more comprehensive training phase which includes a larger diversity in speech, language, and speaker information and also faces the model with more different recording characteristics. Moreover, the multilingual training itself represents an advantage over monolingual training, because it enables new ways to increase the amount of data to train deep neural networks for speech separation tasks. Finally, the reported results also validate the design concept of GlobalPhoneMS2.

7. Acknowledgements

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

8. 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, pp. 975–979, 1953.

[2] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep Clustering: Discriminative Embeddings for Segmentation and Separation," in *Proc. ICASSP*, 2016, pp. 31–35.

[3] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-Channel Multi-Speaker Separation using Deep Clustering," in *Proc. INTERSPEECH*, 2016, pp. 545–549.

[4] D. Yu, M. Kolbæk, Z. Tan, and J. Jensen, "Permutation Invariant Training of Deep Models for Speaker-Independent Multi-talker Speech Separation," in *Proc. ICASSP*, 2017, pp. 241–245.

[5] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," in *Proc. IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, 2019, pp. 1256–1266.

[6] 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*, 2019, pp. 1368–1372.

[7] M. Maciejewski, G. Wichern, and J. Le Roux, "WHAMR!: Noisy and Reverberant Single-Channel Speech Separation," in *Proc. ICASSP*, 2020, pp. 696–700.

[8] 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.

[9] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, "Continuous Speech Separation: Dataset and Analysis," in *Proc. ICASSP*, 2020, pp. 7284–7288.

[10] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "LibriMix: An Open-Source Dataset for Generalizable Speech Separation," *arXiv preprint arXiv:2005.11262v1*, 2020.

[11] B. Kadıoğlu, M. Horgan, X. Liu, J. Pons, D. Darcy, and V. Kumar, "An Empirical Study of Conv-TasNet," in *Proc. ICASSP*, 2020, pp. 7264–7268.

[12] 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 *Proc. WASPAA*, 2019, pp. 165–169.

[13] P. Appeltans, J. Zegers, and Hugo van Hamme, "Practical Applicability of Deep Neural Networks for Overlapping Speaker Separation," in *Proc. INTERSPEECH*, 2019, pp. 1353–1357.

[14] T. Schultz, "GlobalPhone: A Multilingual Speech and Text Database Developed at Karlsruhe University," in *Proc. ICSLP*, 2002, pp. 345–348.

[15] T. Schultz, N. T. Vu, and T. Schlippe, "GlobalPhone: A Multilingual Text & Speech Database in 20 Languages," in *Proc. ICASSP*, 2013, pp. 8126–8130.

[16] T. Schultz and A. Waibel, "Multilingual and Crosslingual Speech Recognition," in *Proc. of the DARPA Broadcast News Transcription and Understanding*, 1998, pp. 259–262.

[17] T. Schultz and A. Waibel, "Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition," in *Speech Communication*, vol. 35, 2001, pp. 31–51.

[18] T. Schultz and A. Waibel, "Experiments on Cross-language Acoustic Modeling," in *Proc. EUROSPEECH*, 2001, pp. 2721–2724.

[19] R. Gretter, "Euronews: A Multilingual Benchmark for ASR and LID," in *Proc. INTERSPEECH*, 2014, pp. 1603–1607.

[20] M. Y. Tachbelie, S. T. Abate, and T. Schultz, "Analysis of GlobalPhone and Ethiopian Languages Speech Corpora for Multilingual ASR," in *Proc. LREC*, 2020, pp. 4152–4156.

[21] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A Large-Scale Multilingual Dataset for Speech Research," in *Proc. INTERSPEECH*, 2020, pp. 2757–2761.

[22] N. T. Vu, D.-C. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, E.-S. Chng, T. Schultz, and H. Li, "A First Speech Recognition System For Mandarin-English Code-Switch Conversational Speech," in *Proc. ICASSP*, 2012, pp. 4889–4892.

[23] H. Seki, T. Hori, S. Watanabe, J. Le Roux, and J. R. Hershey, "End-to-End Multilingual Multi-Speaker Speech Recognition," in *Proc. INTERSPEECH*, 2019, pp. 3755–3759.

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

[25] 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].

[26] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: The PyTorch-based Audio Source Separation Toolkit for Researchers," in *Proc. INTERSPEECH*, 2020, pp. 2637–2641.

[27] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-Baked or Well Done?" in *Proc. ICASSP*, 2019, pp. 626–630.

[28] A. Pandey and D. Wang, "Learning Complex Spectral Mapping for Speech Enhancement with Improved Cross-Corpus Generalization," in *Proc. INTERSPEECH*, 2020, pp. 4511–4515.

[29] J. Zegers and H. van Hamme, "Memory Time Span in LSTMs for Multi-Speaker Source Separation," in *Proc. INTERSPEECH*, 2018, pp. 1477–1481.