BCN2BRNO: ASR System Fusion for Albayzin 2020 Speech to Text Challenge

Martin Kocour[1], Guillermo Cámbara[2,3], Jordi Luque[2], David Bonet[2], Mireia Farrús[4],
Martin Karafiát[1], Karel Veselý[1] and Jan "Honza" Černocký[1]

[1]Brno University of Technology, Speech@FIT, Czechia
[2]Telefónica I+D, Research
[3]Universitat Pompeu Fabra
[4]Universitat de Barcelona

ikocour@fit.vutbr.cz

Abstract

This paper describes the joint effort of BUT and Telefónica Research on the development of Automatic Speech Recognition systems for the Albayzin 2020 Challenge. We compare approaches based on either hybrid or end-to-end models. In hybrid modelling, we explore the impact of a SpecAugment layer on performance. For end-to-end modelling, we used a convolutional neural network with gated linear units (GLUs). The performance of such model is also evaluated with an additional n-gram language model to improve word error rates. We further inspect source separation methods to extract speech from noisy environments (i.e. TV shows). More precisely, we assess the effect of using a neural-based music separator named Demucs. A fusion of our best systems achieved 23.33 % WER in official Albayzin 2020 evaluations. Aside from techniques used in our final submitted systems, we also describe our efforts in retrieving high-quality transcripts for training.

Index Terms: fusion, end-to-end model, hybrid model, semi-supervised, automatic speech recognition, convolutional neural network.

1. Introduction

This paper describes the BCN2BRNO's team Automatic Speech Recognition (ASR) system for the IberSPEECH-RTVE 2020 Speech to Text Transcription Challenge, a joint collaboration between Speech@FIT research group, Telefónica Research (TID) and Universitat Pompeu Fabra (UPF). Our goal is to develop two distinct ASR systems, one based on a hybrid model [1] and the other one on an end-to-end approach [2], and complement each other through a joint fusion.

We submitted one primary system and one contrastive system. The primary system – Fusion B – is a word-level ROVER fusion of hybrid ASR models and end-to-end models. It achieved 23.33 % WER on official evaluation dataset. However, the same result was accomplished by the contrastive system – Fusion A–, a fusion which comprises only hybrid ASR models. In this paper we describe both ASR systems and a post-evaluation analysis and experiments that lead to a better performance of the primary fusion. We also discuss the effect of speech enhancement techniques such as background music removal or speech denoising.

2. Data

The Albayzin 2020 challenge comes with two databases: *RTVE2018* and *RTVE2020*. RTVE2018 is the main source of training and development data, while the RTVE2020 database is used for the final evaluation of submitted systems. RTVE2018 database [3, 4] comprises 15 different TV programs broadcast by the Spanish public television Radiotelevisión Española (RTVE). The programs contain a great variety of speech scenarios from read speech to spontaneous speech, live broadcast or political debates. The database consists of 569 hours of audio data, from which 468 hours are provided with subtitles (train set), and 109 hours are human-revised (dev1, dev2 and test sets). Both hybrid and end-to-end models utilize dev1 and train sets for training, while dev2 and test sets serve as validation datasets. RTVE2020 database [5] consists of 70 hours of manually annotated TV shows broadcast by the RTVE.

In addition, three Linguistic Data Consortium corpora were used for training the language model in the hybrid ASR system: *Fisher Spanish Speech* [6], *CALLHOME Spanish Speech* [7] and *Spanish Gigaword Third Edition* [8]. The audio recordings from CALLHOME and Fisher Spanish Speech corpus were used only for training tiny ASR model for transcript retrieval.

The end-to-end model is trained on Fisher Spanish Speech, Mozilla's Common Voice Spanish corpus and Telefónica's Call Center in-house data (23 hours). The Mozilla's Common Voice Spanish [9] corpus is an open-source dataset that consists of recordings from volunteer contributors pronouncing scripted sentences, recorded at 48kHz rate. The sentences come from original contributor donations and public domain movie scripts. The version of Common Voice corpus used for this work is 5.1, which has 521 hours of recorded speech. However, we have kept only speech validated by the contributors, i.e. 290 hours.

2.1. Transcript retrieval

The training data from RTVE2018 database includes many hours of subtitled speech. However, the captions contain several errors. In most cases, captions are shifted by a few seconds, so a segment with correct transcript corresponds to a different portion of audio. This phenomenon also occurs in human-revised development and test sets. Another problem with subtitled speech is "partly-said" captions. This issue involves misspelled and unspoken words of the transcription.

Since the training procedure of the hybrid ASR system is quite error-prone in case of misaligned labels, we decided to apply a transcript retrieval technique developed by Manohar, et al. [10]: the closed-captions related to the same audio, i.e., the whole TV show, are first concatenated according to the original timeline. This creates a small text corpus containing a few hundreds of words. The text corpus is used for training a biased N-gram language model (LM) with $N = 7$, so the model is biased only on the currently processed captions. During decoding, the weight of the acoustic model (AM) is significantly smaller than

the weight of LM, because we believe that the captions should occur in hypotheses. Then, the "winning" path is retrieved from the hypothesis lattice as the path that has a minimum edit cost w.r.t. the original transcript. Finally, the retrieved transcripts are segmented using the CTMs obtained from the oracle alignment (previous step). The segments, which do not correspondent to original transcripts, are discarded in this step. More details can be found in [11, 10].

Table 1: *2-pass transcript retrieval.*

Cleaning	Train	Dev1	Dev2	Test
Original	468	60.6	15.2	36.8
1-pass	99.4	21	7.5	-
2-pass	234.2	55.1	14.3	33.7
Recovered	50 %	91 %	94 %	92 %

The transcript retrieval technique is applied twice. First, we train an initial ASR system on out-of-domain data, e.g., Fisher and CALLHOME. This system is used to obtain hypothesis in the 1st pass of transcript retrieval. Then, a new system is trained from scratch on clean data from 1st pass and the whole process of transcript retrieval is repeated again. Table 1 shows how this 2-pass cleaning leads to the recovery of almost all the manually annotated development data and half of the subtitled training data.

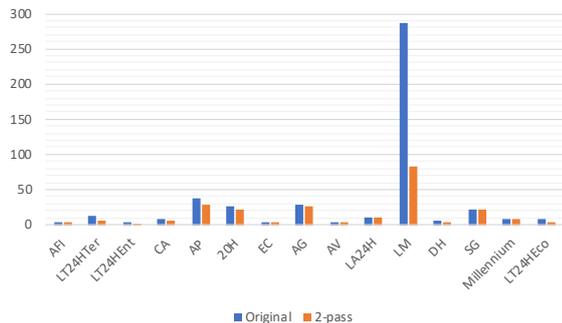Figure 1: *Amount of cleaned audio per TV-show, in hours.*

Figure 1 depicts how many hours have been recovered in individual TV programs. It also shows how data is distributed in the database. Most speech comes from La-Mañana (LM) TV program. We discarded most data in this TV program after 2-pass data cleaning. It happened because this particular TV show was quite challenging for our ASR model.

3. Hybrid speech recognition

3.1. Acoustic Model

In all our experiments, the acoustic model was based on a hybrid Deep Neural Network – Hidden Markov Model architecture trained in Kaldi [12]. The NN part of the model contains 6 convolutional layers followed by 19 TDNN layers with semi-orthogonal factorization [1] (CNN-TDNNf). The input consists of 40-dim MFCCs concatenated with speaker dependent 100-dim i-vectors. Whole model is trained using LF-MMI objective function with bi-phone acoustic units as the targets.

In order to make our NN model training more robust, we introduced feature dropout layer into the architecture. This

prevents the model from overfitting on training data. In fact, it turned the overfitting problem into an underfitting problem. Thus, it leads to a slower convergence during training. This is solved by increasing the number of epochs from 6 to 8 to balance the underfitting in our system. This technique is also known as Spectral Augmentation. It was first suggested for multi-stream hybrid NN models in [13] and fully examined in [14].

3.2. Language Model

We trained three different 3-gram language models: Alb, Wiki and Giga. The names suggest which text corpus was used during training. Albayzin LM was trained on dev1 and train sets from RTVE2018. This text mixture contains 80 thousand unique words in 0.5 million sentences. This small training text is not optimal to train N-gram LM, which is able to generalize well. So we also included larger text corpora: Wikipedia and Spanish Gigaword. These databases were further processed to get rid of unrelated text such as advertisements, emojis and urls. This resulted into more than 2.5 million fine sentences in Wikipedia and 20 million sentences in Spanish Gigaword. We experimented with 4 combinations of interpolation: Alb, Alb+Wiki, Alb+Giga, Alb+Wiki+Giga.

Our vocabulary consists of words from the RTVE2018 database and from the Santiago lexicon[1]. The pronunciation of Spanish words was extracted using a public TTS model called E-speak [15]. The vocabulary was then extended by auxiliary labels for noise, music and overlapped speech. The final lexicon contains around 110 thousand words.

3.3. Voice Activity Detection

Voice activity detection (VAD) was applied on evaluation data in order to segment the audio into smaller chunks. The VAD approach used in this system is based on a feed-forward neural network with two outputs. It expects 15-dimensional filterbank features with additional 3 Kaldi pitch features [16] as the input. Features are normalized with cepstral mean normalization. More details can be found in [17].

4. End-to-end speech recognition

4.1. Acoustic Model

The end-to-end acoustic model is based on a convolutional architecture proposed by Collobert et al. [2] that uses gated linear units (GLUs). Using GLUs in convolutional approaches helps avoiding vanishing gradients, by providing them linear paths while keeping high performances. Concretely, we have used the model from wav2letter Wall Street Journal (WSJ) recipe. This model has approximately 17M parameters with dropout applied after each of its 17 layers. The WSJ dataset contains around 80 hours of audio recordings, which is smaller than the magnitude of our data (∼600 hours). The LibriSpeech recipe (∼1000 hours) provides a deeper ConvNet GLU based architecture, however we decided to use the WSJ one in order to reduce computational time and improve hyper-parameter fine-tuning of the network.

All data samples are resampled at 16kHz, and the system is trained with wav2letter++ framework. Mel-frequency spectral coefficients (MFSCs) are extracted from raw audio, using 80 filterbanks, and the system is trained using the Auto Segmentation criterion (ASG) [2] with batch size set to 4. The learning

---

[1]https://www.openslr.org/34/

rate starts at 5.6 and is decreased down to 0.4 after 30 epochs, where training is finished since no significant WER gains are achieved. From epochs 22 to 28 the system is trained also with the same train set, but adding the RTVE2018 train and dev1 samples with the background music cleaned by Demucs module [18]. The last two epochs, from epoch 28 to epoch 30, are done incorporating further samples with background noise removed by Demucs and denoised by a neural denoiser [19]. This way, data augmentation with samples without background music and noise is done, to aid the network at training with samples with difficult acoustic conditions. Besides, the network is more likely to generalize audio artifacts caused by the denoiser and music separator networks, which is useful when using these to clean test audio.

4.2. Language Model

Regarding the lexicon, we extract it from the train and validation transcripts, plus Sala lexicon [20]. The resulting lexicon is a grapheme-based one with 271k words. We use the standard Spanish alphabet as tokens, plus the "ç" letter from some Catalan words found in the training data, plus the vowels with diacritical marks, making a total of 37 tokens.

The LM is a 5-gram model trained with KenLM [21] using only transcripts from the training sets: RTVE2018 train and dev1, plus Common Voice, Fisher and Call Center. The resulting LM is described in this paper as *Alb+Others*.

Fine-tuning of decoder hyperparameters is done via grid-search with the RTVE2018 dev2 set. The best results are achieved with an LM weight of 2.25, a word score of 2.25 and a silence score of -0.35. This same configuration is then applied for evaluation datasets from RTVE2018 and RTVE2020.

5. Experiments

5.1. Data cleaning

Data cleaning by means of 2-pass transcript retrieval improves the performance of our models the most. Table 2 shows the effect of each pass. The 2nd pass helped to improve the accuracy by almost 2 % in terms of WER. We also ran a 3rd pass, but that did not help anymore. We simply did not retrieve more cleaned data from the original transcripts, just 3 hours more.

Table 2: *Effect of 2-pass transcript cleaning evaluated on RTVE2018 test set.*

AM	LM	Training data	WER [%] Test
CNN-TDNNf	Alb	1-pass	17.2
		2-pass	15.5
		3-pass	15.5

5.2. Speech Enhancement

It is very common to find background music on TV programs, which can confuse our recognizer if it has a notorious presence. This brought us the idea of processing the audio through a Music Source Separator called Demucs [18]. It separates the original audio into voice, bass, drums and others. By keeping only the voice component, we managed to significantly eliminate the background music, while maintaining relatively good quality in the original voice.

We enhanced both validation sets in order to assess possible WER reductions. As seen in Table 4, this approach yielded a small increase in WER. We also tried applying a specialized denoiser [19] after background music removal, but the WER for dev2 increased in an absolute 1.6%, compared to original system without enhancement. None of these two approaches (Demucs and Demucs+Denoiser) provided WER improvements at first, so we did not apply them for the end-to-end model used in the fusion. Nevertheless, the end-to-end, end-to-end + Demucs and end-to-end + Demucs + Denoiser models were submitted as separate systems by the UPF-TID team (see Table 5 for details).

Our hypothesis is that not all the samples contain background music. Speech enhancement for samples that are clean already is detrimental because it causes slight degradations in the signal. Hence, we have evaluated the effects of applying music source separation to samples under certain SNR ranges, measured with the WADA-SNR algorithm [22]. The application of music separation on RTVE dataset is optimal for SNR ranges between -5 and 5 or 8 as it is shown in Table 3. Looking at Figure 2, best improvements are found at TV shows with higher WER (thus harder/noisier speech), e.g., AV, where most of the time speakers are in a car, or LM and DH, where music and speech often overlap. Other shows have slighter benefits, since these already contain good quality audio. The exception is AFI show, which is reported to have poor quality audio, so further audio degradation from Demucs might cause worse performance.

Figure 2: *Variation of the mean WER per TV show between using Demucs-cleaned or original samples on RTVE's 2018 test set. Negative values represent Demucs improvements. Note that only samples with SNR between -5 and 8 are enhanced.*
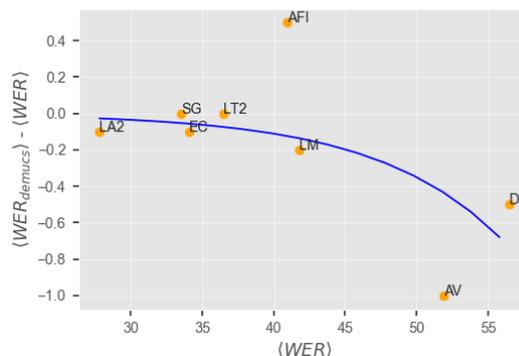

Table 3: *WER impact of cleaning speech signals between certain SNR ranges, using a music source separator. End-to-end ConvNet GLU model is used without LM, and percentage of cleaned samples are reported.*

SNR	Cleaned Samples [%]		Test WER [%]	
	2018	2020	2018	2020
$(-\infty, \infty)$	100	100	37.50	53.53
$(-\infty, 10)$	25.97	34.22	-0.05	-0.87
$(-5, 10)$	25.84	31.33	-0.05	-0.88
$(-5, 5)$	5.14	11.88	-0.07	**-1.03**
$(-5, 8)$	14.95	22.11	**-0.08**	-0.97

5.3. Spectral augmentation

Table 4 shows compared models with and without spectral augmentation. The technique helps quite significantly. All models with feature dropout layer outperformed their counterparts with a quite constant 0.4% absolute WER improvement on RTVE2018 test set and around 0.6% on RTVE2018 dev2 set.

5.4. Model fusion

We also fuse the output of our best systems to further improve the performance. Overall results of our systems considered for the fusion are depicted in Table 4. Since the models with spectral augmentation performed significantly better, we decided to fuse only these systems. We analyzed two different approaches: a pure hybrid model fusion (Fusion A) and hybrid and end-to-end model fusion (Fusion B).

Considering that the end-to-end model does not provide word-level timestamps, we had to force-align the transcripts with the hybrid ASR system in order to obtain CTM output. The original word-level fusion was done using ROVER toolkit [23]. Fusion B with end-to-end models performed slightly better than its counterpart Fusion A, despite the fact that the end-to-end models achieved worse results. This somehow proves the idea that the fusion can benefit from different modeling approaches.

6. Final systems

Table 5 depicts the results on RTVE2020 test set. For the end-to-end ConvNet GLU model, the performance drops around a 15% WER when compared with previous results on development sets. Since the TV shows in such sets are also present in training dataset, our hypothesis is that the model slightly overfits to them. Therefore, when facing different acoustic conditions, voices, background noises and music types included in the RTVE2020 test set, the WER increases noticeably. Enhancing the test samples with Demucs or with Demucs+Denoiser yields a worse WER score, probably due to an inherent degradation of the signal. A deeper analysis about more efficient ways to apply such enhancements has been done in Section 5.2.

Also, note that the submitted systems had a leak of dev2 stm transcripts in the LM, causing an hyperparameter overfitting during LM tuning. This caused a WER drop in all end-to-end systems, yielding WERs of 41.4%, 42.3% and 58.6%. Table 5 also displays the results of these systems with the leakage and LM tuning corrected in post-evaluation analysis.

7. Conclusions

In this paper we described two different ASR model architectures and their fusion. We focused on improving the original subtitled data in order to train our models on high quality target labels. We also improved the N-gram language model by incorporating publicly available text data from Wikipedia and Spanish Gigaword corpus from LDC. We have also successfully incorporated the spectral augmentation into our AM architecture. Our best system achieved 13.3% and 23.24% WER on RTVE2018 and RTVE2020 test sets respectively.

The performance of our hybrid system can be further improved by using lattice-fusion with Minimum Bayes Risk decoding [24]. Another space for improvement is offered by

---

2 Primary system of UPF-TID team.
3 First contrastive system of UPF-TID team.
4 Second contrastive system of UPF-TID team.

Table 4: *Overall results on RTVE2018 dataset with various language models and fusions.*

	AM	LM	WER [%] Dev2	Test
1		Alb	14.1	15.5
2	CNN-TDNNf	Alb + Wiki	13.6	14.9
3		Alb + Giga	13.6	15.1
4		Alb + Wiki + Giga	13.5	15.0
5		Alb	13.4	15.0
6	+ SpecAug	Alb+Wiki	12.9	14.5
7		Alb+Giga	13.0	14.7
8		Alb+Wiki+Giga	12.9	14.6
9	ConvNet GLU	None	36.1	37.5
10		Alb + Others	20.8	20.7
11	+ Demucs	None	36.4	37.5
12		Alb + Others	21.1	20.8
13	Fusion A	*(row 5-8)*	12.9	13.7
14	Fusion B	*(row 5-8 and 10)*	**12.8**	**13.3**

Table 5: *Official and post-evaluation final results on RTVE2020 eval set for the submitted systems.*

Model	WER [%] Official	Post-eval
CNN-TDNNf	-	24.3
+ SpecAug	-	23.5
ConvNet GLU	41.4[2]	36.2
+ Demucs	42.3[3]	37.9
+ Demucs + Denoiser	58.6[4]	40.0
Fusion A	23.33	23.38
Fusion B	**23.33**	**23.24**

adding an RNN-LM lattice rescoring. Our end-to-end model shows relatively competitive performance on the RTVE2018 test set in comparison with its hybrid counterpart. However, its performance on the RTVE2020 evaluation set exposes that the model was not able to generalize very well since this database turns out to contain slightly different acoustic conditions. Despite of this fact, the model still managed to improve the results in the final fusion with hybrid systems. An exploration on background music removal shows that it yields the best results for lower SNR ranges, thus having a different impact depending on the acoustic conditions of each TV show.

8. Acknowledgements

9. References

[1] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks," in *Proceedings of Interspeech*, 09 2018, pp. 3743–3747.

[2] R. Collobert, C. Puhrsch, and G. Synnaeve, "Wav2letter: an end-to-end convnet-based speech recognition system," *CoRR*, vol. abs/1609.03193, 2016. [Online]. Available: http://arxiv.org/abs/1609.03193

[3] E. Lleida, A. Ortega, A. Miguel, V. Bazán, C. Pérez, M. Zotano, and A. De Prada, "RTVE2018 Database Description," 2018. [Online]. Available: http://catedrartve.unizar.es/reto2018/RTVE2018DB.pdf

[4] E. Lleida, A. Ortega, A. Miguel, V. Bazán-Gil, C. Pérez, M. Gómez, and A. de Prada, "Albayzin 2018 evaluation: the IberSPEECH-RTVE challenge on speech technologies for Spanish broadcast media," *Applied Sciences*, vol. 9, no. 24, p. 5412, 2019.

[5] E. Lleida, A. Ortega, A. Miguel, V. Bazán-Gil, C. Pérez, M. Góméz, and A. De Prada, "RTVE2020 Database Description," 2020. [Online]. Available: http://catedrartve.unizar.es/reto2020/RTVE2020DB.pdf

[6] D. Graff, S. Huang, I. Cartagena, K. Walker, and C. Cieri, "Fisher Spanish Speech," *LDC2010S01. DVD. Philadelphia: Linguistic Data Consortium*, 2010.

[7] A. Canavan and G. Zipperlen, "CALLHOME Spanish Speech," *LDC96S35. Web Download. Philadelphia: Linguistic Data Consortium*, 1996.

[8] Ângelo Mendonça, D. Jaquette, D. Graff, and D. DiPersio, "Spanish Gigaword Third Edition," *LDC2011T12. Web Download. Philadelphia: Linguistic Data Consortium*, 2011.

[9] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," 2019.

[10] V. Manohar, D. Povey, and S. Khudanpur, "JHU Kaldi system for Arabic MGB-3 ASR challenge using diarization, audio-transcript alignment and transfer learning," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, vol. 2018-. IEEE, 2017, pp. 346–352.

[11] M. Kocour, "Automatic Speech Recognition System Continually Improving Based on Subtitled Speech Data," Diploma thesis, Brno University of Technology, Faculty of Information Technology, Brno, 2019, technical supervisor Dr. Ing. Jordi Luque Serrano. supervisor Doc. Dr. Ing. Jan Černocky.

[12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[13] S. H. R. Mallidi and H. Hermansky, "A Framework for Practical Multistream ASR," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, N. Morgan, Ed. ISCA, 2016, pp. 3474–3478.

[14] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*. ISCA, 2019, pp. 2613–2617.

[15] J. Duddington and R. Dunn, "eSpeak text to speech," 2012. [Online]. Available: http://espeak.sourceforge.net

[16] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. Florence, Italy: IEEE, May 2014.

[17] O. Plchot, P. Matějka, O. Novotný, S. Cumani, A. D. Lozano, J. Slavíček, M. S. Diez, F. Grézl, O. Glembek, M. V. Kamsali, A. Silnova, L. Burget, L. Ondel, S. Kesiraju, and A. J. Rohdin, "Analysis of BUT-PT Submission for NIST LRE 2017," in *Proceedings of Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 47–53.

[18] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," *arXiv preprint arXiv:1911.13254*, 2019.

[19] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," 2020.

[20] A. Moreno, O. Gedge, H. Heuvel, H. Höge, S. Horbach, P. Martin, E. Pinto, A. Rincón, F. Senia, and R. Sukkar, "SpeechDat across all America: SALA II," 2002.

[21] K. Heafield, "KenLM: Faster and Smaller Language Model Queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, ser. WMT '11. USA: Association for Computational Linguistics, 2011, p. 187–197.

[22] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[23] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," in *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, 1997, pp. 347–354.

[24] P. Swietojanski, A. Ghoshal, and S. Renals, "Revisiting hybrid and GMM-HMM system combination techniques," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, 2013.