IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Speech Enhancement for Wake-Up-Word detection in Voice Assistants

David Bonet[1,5], *Guillermo Cámbara*[2,5], *Fernando López*[4], *Pablo Gómez*[4], *Carlos Segura*[5], *Mireia Farrús*[3], *Jordi Luque*[5]

[1]Universitat Politècnica de Catalunya, [2]Universitat Pompeu Fabra, [3]Universitat de Barcelona, [4]Telefónica I+D, Digital Home
[5]Telefónica I+D, Research, Spain

jordi.luque@telefonica.com

Abstract

Keyword spotting and in particular Wake-Up-Word (WUW) detection is a very important task for voice assistants. A very common issue of voice assistants is that they get easily activated by background noise like music, TV or background speech that accidentally triggers the device. In this paper, we propose a Speech Enhancement (SE) model adapted to the task of WUW detection that aims at increasing the recognition rate and reducing the false alarms in the presence of these types of noises. The SE model is a fully-convolutional denoising auto-encoder at waveform level and is trained using a log-Mel Spectrogram and waveform reconstruction losses together with the BCE loss of a simple WUW classification network. A new database has been purposely prepared for the task of recognizing the WUW in challenging conditions containing negative samples that are very phonetically similar to the keyword. The database is extended with public databases and an exhaustive data augmentation to simulate different noises and environments. The results obtained by concatenating the SE with a simple and state-of-the-art WUW detectors show that the SE does not have a negative impact on the recognition rate in quiet environments while increasing the performance in the presence of noise, especially when the SE and WUW detector are trained jointly end-to-end.
Index Terms: keyword spotting, speech enhancement, wake-up-word, deep learning, convolutional neural network

1. Introduction

Voice interaction with devices is becoming ubiquitous. Most of them use a mechanism to avoid the excessive usage of resources, a trigger word detector. This ensures the efficient use of resources, using a Speech-To-Text tool only when needed and with the consequent start of a conversation. It is key to only start this conversation when the user is addressing the device, otherwise the user experience is notoriously degraded. Thus, the wake-up-word detection system must be robust enough to avoid wake-ups with TV, music, speech and sounds that do not contain the key phrase.

A common approach to reduce the impact of this type of noise in the system is the adoption of speech enhancement algorithms. Speech enhancement consists of the task of improving the perceptual intelligibility and quality of speech by removing background noise [1]. Its main application is in the field of mobile and internet communications [2] and related to hearing aids [3], but SE has also been applied successfully to automatic speech recognition systems [4, 5, 6].

Traditional SE methods involved a characterization step of the noise spectrum which is then used to try reduce the noise from the regenerated speech signal. Examples of these approaches are spectral subtraction [3], Wiener filtering [7] and subspace algorithms [8]. One of the main drawbacks of the classical approaches is that they are not very robust against non-stationary noises or other type of noises that can mask speech, like background speech. In the last years, Deep Learning approaches have been widely applied to SE at the waveform level [9, 10] and spectral level [6, 11]. In the first case, a common architecture falls within the encoder-decoder paradigm. In [12], authors proposed a fully convolutional generative adversarial network architecture structured as an auto-encoder with U-Net like skip-connections. Other recent work [13] proposes a similar architecture at the waveform level that includes a LSTM between the encoder and the decoder and it is trained directly with a regression loss combined with a spectrogram domain loss.

Inspired by these recent models, we propose a similar SE auto-encoder architecture in the time domain that is optimized not only by minimizing waveform and Mel-spectrogram regression losses, but also includes a task-dependent classification loss provided by a simple WUW classifier acting as a Quality-Net [14, 15]. This last term serves as a task-dependent objective quality measure that trains the model to enhance important speech features that might be degraded otherwise. The WUW detection is performed by concatenating the SE model with the classifier.

2. Speech Enhancement

Speech enhancement is interesting for triggering phrase detection since it tries to remove noise that could trigger the device, and at the same time improves speech quality and intelligibility for a better detection. In this case, we try to tackle the most common noisy environments where voice assistants are used: TV, music, background conversations, office noise and living room noise. Some of these types of background noise, such as TV and background conversations, are the most likely to trigger the voice assistant and are also the most challenging to remove.

2.1. Model

Our model has a fully-convolutional denoising auto-encoder architecture with skip connections (Fig. 1), working end-to-end at waveform level. Similar designs have proven to be very effective in SE tasks [12, 13, 16]. In training, we input a noisy audio $\boldsymbol{x} \in \mathbb{R}^T$, comprised of clean speech signal $\boldsymbol{y} \in \mathbb{R}^T$ and background noise $\boldsymbol{n} \in \mathbb{R}^T$ so that $\boldsymbol{x} = \lambda \boldsymbol{y} + (1 - \lambda)\boldsymbol{n}$.

The encoder compresses the input signal and expands the number of channels. It is composed of six convolutional blocks (ConvBlock1D), each consisting of a convolutional layer, followed by an instance normalization and a rectified linear unit (ReLU). Kernel size $K = 4$ and stride $S = 2$ are used, except in the first layer where $K = 7$ and $S = 1$. The compressed sig-

10.21437/IberSPEECH.2021-9

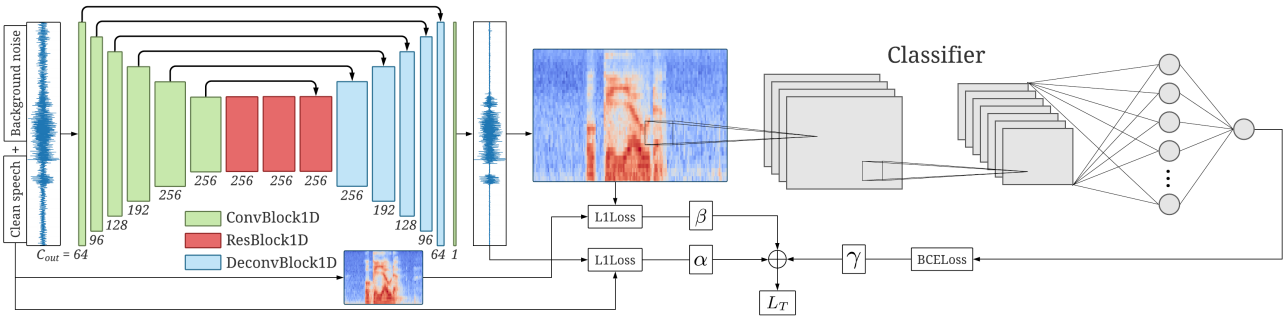

Figure 1: *End-to-end SE model at waveform level concatenated with a classifier. The Log-Mel Spectrogram and waveform reconstruction losses of the SE model can be used together with the task-dependent loss (BCE Loss) of the classifier acting as a Quality-Net [15] to train the model. The latter term aims at enhancing relevant speech features for the WUW detection task.*

nal goes through an intermediate stage where the shape is preserved, consisting of three residual blocks (ResBlock1D), each formed by two ConvBlock1D with $K = 3$ and $S = 1$ where a skip connection is added from the input of the residual block to its output. The last stage of the SE model is the decoder, where the original shape of the raw audio is recovered at the output, and the enhanced signal can serve as input to a WUW classifier. Its architecture follows the inverse structure of the encoder, where deconvolutional blocks (DeconvBlock1D) replace the convolutional layers of the ConvBlock1D with transposed convolutional layers. Skip connections from the encoder blocks to the decoder blocks are also used to ensure low-level detail when reconstructing the waveform.

We use a regression loss function (L1 loss) at raw waveform level together with another L1 loss over the log-Mel Spectrogram as proposed in [17] to reconstruct a "cleaned" signal $\hat{\boldsymbol{y}}$ at the output. Finally, we include the classification loss (BCE Loss) when training the SE model jointly with the classifier or concatenating a pretrained classifier at its output. Thus, we also try to optimize the SE model to the specific task of WUW classification. Our final loss function is defined as a linear combination of the three losses:

$$L_T = \alpha L_{raw}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \beta L_{spec}(S(\boldsymbol{y}), S(\hat{\boldsymbol{y}})) + \gamma L_{BCE} \quad (1)$$

where α, β and γ are hyperparamters weighting each loss term and $S(\cdot)$ denotes the log-Mel Spectrogram of the signal, which is computed using 512 FFT bins, a window of 20 ms with 10 ms of shift and 40 filters in the Mel Scale.

3. Methodology

3.1. Databases

The database used for conducting the experiments here presented consists of WUW samples labeled as positive, and other non-WUW samples labeled as negative. Since the chosen keyword is "OK Aura", which triggers Telefónica's home assistant, Aura, positive samples are drawn from company's in-house databases. Some of the negative samples have been also recorded in such databases, but we also add speech and other types of acoustic events from external data sets, so the models gain robustness with further data augmentation. Information about all data used is detailed in this section.

3.1.1. OK Aura Database

In a first round, around 4300 WUW samples from 360 speakers have been collected, resulting in 2.8 hours of audio. Furthermore, office ambient noise has been recorded as well, with the aim of having samples for noise data augmentation. The second data collection round has been done in order to study and improve some sensitive cases where WUW modules typically underperform. For instance, such dataset contains rich metadata about positive and negative utterances, like room distance, speech accent, emotion, age or gender. Furthermore, the negative utterances contain phonetically similar words to "OK Aura", since these are the most ambiguous to recognize for a classifier. Detailed information about data acquisition is explained in the following subsection.

3.1.2. Data acquisition

A web-based Jotform form[1] has been designed for data collection. Readers are invited to contribute to the dataset while the form is still open. Until the date of this work, 1096 samples from 80 speakers have been recorded, which consists of 1.2 hours of audio[2]. Volunteers are asked to pronounce various scripted utterances at a close distance and also at two meters from the device mic. The similarity levels are the following:

1. Exact WUW, in an isolated manner: *OK Aura.*
2. Exact WUW, in a context: *Perfecto, voy a mirar qué dan hoy. OK Aura.*
3. Contains "Aura": *Hay un aura de paz y tranquilidad.*
4. Contains "OK": *OK, a ver qué ponen en la tele.*
5. Contains similar word units to "Aura": *Hola Laura.*
6. Contains similar word units to "OK": *Prefiero el hockey al baloncesto.*
7. Contains similar word units to "OK Aura": *Porque Laura, ¿qué te pareció la película?*

3.1.3. External data

General negative examples have been randomly chosen from the publicly available Spanish Common Voice (CV) corpus [18] that currently holds over 300 hours of validated audio. However, we keep a 10:1 ratio between negative and positive samples, since such ratio proves to yield good results in [19], thus avoiding bigger ratios that lead to increasing computational times. Final database collects a CV partition consisting of 55h for training, 7h for development and 7h for testing.

Background noises were selected from various public datasets according to different use case scenarios. Living room

[1] https://form.jotform.com/201694606537056

[2] The AURA-WUW dataset, including audio and alignments, is available upon request from the authors and agreement of EULA for research purposes.

background noise (HOME-LIVINGB) from the QUT-NOISE Database [20], TV audios from the IberSpeech-RTVE Challenge [21], and music[3] and conversations[4] from free libraries.

3.1.4. Data processing

All the audio samples are monoaural signals stored in Waveform Audio File Format (WAV) with a sampling rate of 16kHz. The speech data that has been collected was processed with a Speech Activity Detection (SAD) module producing timestamps where speech occurs. For this purpose the tool from pyannote.audio [22] has been used, which has been trained with the AMI corpus [23]. This helped us to only use the valid speech segments of the audios we collected.

As features to train the classifiers we mainly used two, as we wanted to maintain the architectures with the originally proposed features [24, 25]: Mel-Frequency Cepstral Coefficients (MFCCs) and log-Mel Spectrogram. The MFCCs were constructed first filtering the audio with a band pass filter (20Hz to 8kHz) and then, extracting the first thirteen coefficients with 100 ms of windows size and frame shifting of 50 ms. The procedure to extract the log-Mel Spectrogram ($S(\cdot)$) is detailed in §2.1.

Train, development and test partitions are split ensuring that neither speaker nor background noise is repeated between partitions, trying to maintain a 80-10-10 proportion, respectively. Total data, containing internal and external datasets, consists of 50.737 non-WUW samples and 4.651 WUW samples.

3.2. Data augmentation

Several Room Impulse Responses (RIR) were created based on the Image Source Method (ISM) [26], for a room of dimensions (L_x, L_y, L_z) where $2 \leq L_x \leq 4.5, 2 \leq L_y \leq 5.5, 2.5 \leq L_z \leq 4$ meters, with microphone and source randomly located at any (x, y) point within a height of $0.5 \leq z \leq 2$ meters. Every TV and music original recordings were convolved with different RIRs to simulate the signal picked up by the microphone of the device in the room.

The main data augmentation technique used in this work is background noise addition. Different noise scenarios, like TV, music, conversations, office and living room, have been combined with clean samples within a wide range of SNRs. It aims at improving the performance of the models against noisy environments. In each epoch, we create different noisy samples by randomly selecting a sample of background noise for each speech event and combining them with a randomly chosen SNR in a specified range. We tried data augmentation techniques like time stretching and pitch shifting, but we discarded them since no significant changes were achieved in the noisy regions.

3.3. Wake-Up Word Detection Models

With the aim of assessing the quality of the trained SE models, we use several trigger word detection classifier models, reporting the impact of the SE module at WUW classification performance. The WUW classifiers used here are a `LeNet`, a well-known standard classifier, easy to optimize [27]; `Res15`, `Res15-narrow` and `Res8` based on a reimplementation by Tang and Lin [28] of Sainath and Parada's Convolutional Neural Networks (CNNs) for keyword spotting [29], using residual learning techniques with dilated convolutions [30]; a `SGRU` and

[3] https://freemusicarchive.org/

[4] http://www.podcastsinspanish.org/

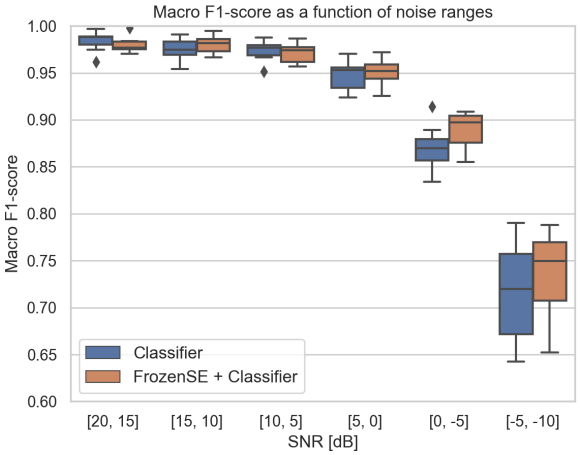

Figure 2: *Macro F1-score box plot for different SNR ranges. Classifiers trained with low noise ($[5, 30]$ dB SNR).*

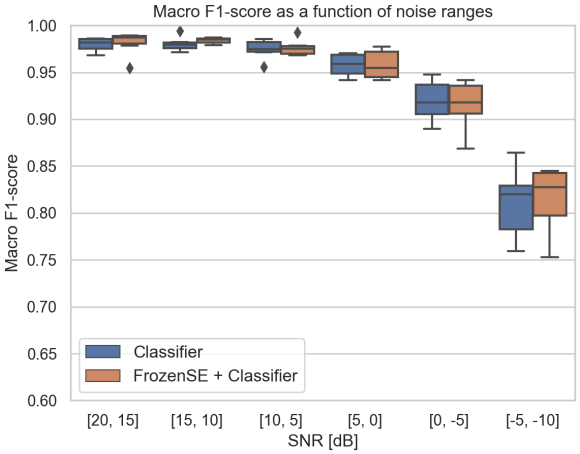

Figure 3: *Macro F1-score box plot for different SNR ranges. Classifiers trained with a very wide range of noise ($[-10, 50]$ dB SNR).*

`SGRU2`, two Recurrent Neural Network (RNNs) models, based on the open source tool named Mycroft Precise [24], which is a lightweight wake-up-word detection tool implemented in TensorFlow. These are two bigger variations that we have implemented in PyTorch. We also use a `CNN-FAT2019`, a CNN architecture adapted from a kernel [25] in Kaggle's FAT 2019 competition [31], which has shown good performance in tasks like audio tagging or detection of gender, identity and speech events from pulse signal [32].

3.4. Training

Speech signals and background noises are combined randomly following the procedure explained in 3.2 with a given SNR range. The SE model is trained to cover a wide SNR range of $[-10, 50]$ dBs, whereas WUW models are trained to cover two scenarios: a classifier trained with the same SNR range as the SE model, and a classifier less aware of noise with a narrower SNR range of $[5, 30]$ dBs. This way, it is possible to study the impact of the SE model regarding if the classifier has been trained with more or less noise.

Data imbalance is addressed balancing the classes in each batch using a weighted sampler. We use a fixed window length of 1.5 s based on the annotated timestamps for our collected database, and random cuts for the rest of the CV samples.

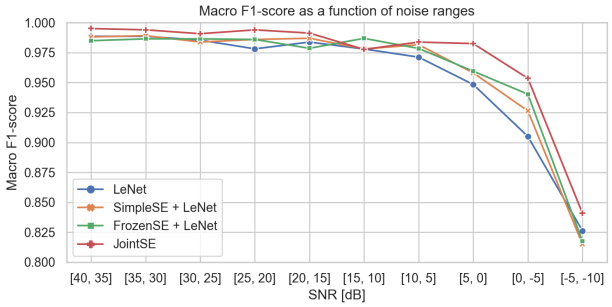

Figure 4: *Comparison of different training methods for the SE models and LeNet classifier, in terms of the macro F1-Score for different SNR ranges. All models trained in the range of* $[-10, 50] dB$ *SNR.*

All the models are trained with early stopping based on the validation loss with 10 epochs of patience. We use the Adam optimizer with a learning rate of 0.001 and a batch size of 50. Loss (1) allows to train the models in multiple ways and we define different SE models and classifiers based on the loss function used:

a) Classifier: we remove the auto-encoder from the architecture (Fig. 1) and train any of the classifiers using the noisy audio as input: $\alpha = \beta = 0$ and $\gamma = 1$
b) SE model (SimpleSE): we remove the classifier from the architecture and optimize the auto-encoder based on the reconstruction losses only: $\alpha = \beta = 1$ and $\gamma = 0$
c) SE model + frozen classifier (FrozenSE): operations of the classifier are dropped from the backward graph for gradient calculation, optimizing only the SE model for a given pretrained classifier (LeNet). $\alpha = \beta = \gamma = 1$
d) SE model + classifier (JointSE): auto-encoder and LeNet are trained jointly using the three losses: $\alpha = \beta = \gamma = 1$

3.5. Tests

All the models take as input windows of 1.5 s of audio, to ensure that common WUW utterances are fully within it, since the average "OK Aura" is about 0.8 s long. Therefore, we perform an atomic test evaluating if a single window contains the WUW or not. Both negative and positive samples are assigned a background noise sample with which they are combined with a random SNR between certain ranges, as described in §3.4.

Given the output scores of the models, the decision threshold is chosen as the one yielding the biggest difference between true and false positive rates, based on Youden's J statistic [33]. Once the threshold is decided, macro F1-score is computed in order to balance WUW/non-WUW proportions in the results. We average such scores across all WUW classifiers described in §3.3, for every SNR range.

4. Results

Figure 2 illustrates the improvement of the WUW detection in noisy scenarios by concatenating our FrozenSE model with all WUW classifiers described in §3.3 trained with low noise ($[5, 30]$ dB SNR), which we could find in simple voice assistant systems. Applying SE in quiet scenarios maintains fairly good results, and improves them in lower SNR ranges.

If we train the classifiers with more data augmentation ($[-10, 50]$ dB SNR), results using the FrozenSE do not decrease but the improvement in ranges of severe noise is not as large as in Figure 2, see Figure 3.

Table 1: *Macro F1-score enhancing the noisy audios with SOTA SE models and using a LeNet as a classifier.*

SNR $[dB]$		**No SE**	**SEGAN**	**Denoiser**	**JointSE**
$[20, 10]$	Clean	0.980	0.964	0.980	**0.990**
$[10, 0]$	Noisy	0.969	0.940	0.955	**0.972**
$[0, -10]$	Very noisy	0.869	0.798	0.851	**0.902**

Table 2: *Macro F1-score percentage difference between JointSE and LeNet without SE module, for different background noise types. Positive values mean that the JointSE score is bigger than the single LeNet's.*

SNR $[dB]$		**Music**	**TV**	**Office**	**Living Room**	**Conversations**
$[20, 10]$	Clean	1.0	-0.9	1.4	0.4	**2.3**
$[10, 0]$	Noisy	0.0	-1.2	0.8	0.4	**1.9**
$[0, -10]$	Very noisy	0.5	3.9	**11.2**	3.1	3.8

In §3.4 we have defined the parameters of the loss function (1) to train a classifier (case a)), and different approaches to train the SE model, either standalone (b), c)) or in conjunction with the classifier (d)). In Figure 4 we can see how JointSE performs better than all the other cases in almost every SNR range. From 40 dB to 10 dB of SNR, the results are very similar for the 4 models. In contrast, in the noisiest ranges we can see how the classifier without SE model is the worst performer, followed by the SimpleSE case where only the waveform and spectral reconstruction losses are used. We found that the FrozenSE case, which includes the classification loss in the training stage, improves the results for the wake-up-word detection task. However, the best results are obtained with the JointSE case where the SE model + LeNet are trained jointly using all three losses.

We compared the WUW detection results of our JointSE with other SOTA SE models (SEGAN [12] and Denoiser [13]), followed by a classifier (data augmented LeNet) in different noise scenarios. In Table 1, it can be observed how when training the models together with the task loss, the results in our setup are better than with other more powerful but more general SE models, since there is no mismatch between the SE and classifier in the end-to-end and it is also more adapted to common home noises. JointSE improves the detection over the no SE model case, especially in scenarios with background conversations, loud office noise or loud TV, see Table 2.

5. Conclusions

In this paper we proposed a SE model adapted to the task of WUW in voice assistants for the home environment. The SE model is a fully-convolutional denoising auto-encoder at waveform level and it is trained using a log-Mel Spectrogram and waveform regression losses together with a task-dependent WUW classification loss. Results show that for clean and slightly noisy conditions, SE in general does not bring a substantial improvement over a classifier trained with proper data augmentation. In very noisy conditions, SE does improve the results, especially when the SE model and WUW detector are trained jointly end-to-end, performing better than a general-purpose SE model.

6. Acknowledgments

This work has been partly funded by the INGENIOUS project within the European Union's Horizon 2020 Research and Innovation Programme under GA No 833435 and by the Agencia Estatal de Investigación (AEI), Ministerio de Ciencia, Innovación y Universidades and the Fondo Social Europeo (FSE) under grant RYC-2015-17239 (AEI/FSE, UE).

7. References

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[2] C. K. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, "A scalable noisy speech dataset and online subjective test framework," *arXiv preprint arXiv:1909.08050*, 2019.

[3] L.-P. Yang and Q.-J. Fu, "Spectral subtraction-based speech enhancement for cochlear implant patients in background noise," *The journal of the Acoustical Society of America*, vol. 117, no. 3, pp. 1001–1004, 2005.

[4] C. Zorilă, C. Boeddeker, R. Doddipatla, and R. Haeb-Umbach, "An investigation into the effectiveness of enhancement in ASR training and test for chime-5 dinner party transcription," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 47–53.

[5] A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[6] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

[7] J. Meyer and K. U. Simmer, "Multi-channel speech enhancement in a car environment using wiener filtering and spectral subtraction," in *1997 IEEE international conference on acoustics, speech, and signal processing*, vol. 2. IEEE, 1997, pp. 1167–1170.

[8] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on speech and audio processing*, vol. 3, no. 4, pp. 251–266, 1995.

[9] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.

[10] H. Phan, I. V. McLoughlin, L. Pham, O. Y. Chén, P. Koch, M. De Vos, and A. Mertins, "Improving GANs for speech enhancement," *arXiv preprint arXiv:2001.05532*, 2020.

[11] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *arXiv preprint arXiv:1609.07132*, 2016.

[12] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[13] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *arXiv preprint arXiv:2006.12847*, 2020.

[14] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm," *arXiv preprint arXiv:1808.05344*, 2018.

[15] S.-W. Fu, C.-F. Liao, and Y. Tsao, "Learning with learned loss function: Speech enhancement with quality-net to improve perceptual evaluation of speech quality," *IEEE Signal Processing Letters*, vol. 27, pp. 26–30, 2019.

[16] J. Llombart, D. Ribas, A. Miguel, L. Vicente, A. Ortega, and E. Lleida, "Progressive loss functions for speech enhancement with deep neural networks," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 1–16, 2021.

[17] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[18] 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," *arXiv preprint arXiv:1912.06670*, 2019.

[19] J. Hou, Y. Shi, M. Ostendorf, M.-Y. Hwang, and L. Xie, "Mining effective negative training samples for keyword spotting," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7444–7448.

[20] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," *Proceedings of Interspeech 2010*, 2010.

[21] 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.

[22] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "pyannote.audio: neural building blocks for speaker diarization," in *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020.

[23] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[24] M. D. Scholefield, "Mycroft Precise," https://github.com/MycroftAI/mycroft-precise, 2019.

[25] M. H. "mhiro2", "Freesound Audio Tagging 2019: Simple 2D-CNN Classifier with PyTorch," https://www.kaggle.com/mhiro2/simple-2d-cnn-classifier-with-pytorch/, 2019.

[26] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[27] Y. LeCun *et al.*, "Lenet-5, convolutional neural networks," *URL: http://yann. lecun. com/exdb/lenet*, vol. 20, no. 5, p. 14, 2015.

[28] R. Tang and J. Lin, "Honk: A PyTorch reimplementation of convolutional neural networks for keyword spotting," *CoRR*, vol. abs/1710.06554, 2017. [Online]. Available: http://arxiv.org/abs/1710.06554

[29] T. N. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[30] R. Tang and J. Lin, "Deep residual learning for small-footprint keyword spotting," *CoRR*, vol. abs/1710.10361, 2017. [Online]. Available: http://arxiv.org/abs/1710.10361

[31] E. Fonseca, M. Plakal, F. Font, D. P. Ellis, and X. Serra, "Audio tagging with noisy labels and minimal supervision," *arXiv preprint arXiv:1906.02975*, 2019.

[32] G. Cámbara, J. Luque, and M. Farrús, "Detection of speech events and speaker characteristics through photo-plethysmographic signal neural processing," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7564–7568.

[33] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.