WAWENETS: A NO-REFERENCE CONVOLUTIONAL WAVEFORM-BASED APPROACH TO ESTIMATING NARROWBAND AND WIDEBAND SPEECH QUALITY

Andrew A. Catellier and Stephen D. Voran

Institute for Telecommunication Sciences, 325 Broadway, Boulder, Colorado, 80305, USA
{acatellier, svoran}@ntia.gov

ABSTRACT

Building on prior work we have developed a no-reference (NR) waveform-based convolutional neural network (CNN) architecture that can accurately estimate speech quality or intelligibility of narrowband and wideband speech segments. These Wideband Audio Waveform Evaluation Networks, or WAWEnets, achieve very high per-speech-segment correlation ($\rho_{seg} \geq 0.92$, RMSE ≤ 0.38) to established full-reference quality and intelligibility estimators (PESQ, POLQA, PEMO, STOI) based on over 17 hours of speech from 127 previously unseen talkers speaking in 13 different languages; just 10% of our total data. NR correlations at this level across such a broad scope are unprecedented. This achievement was made possible by using full-reference estimates as training targets so that WAWEnets could learn implicit undistorted speech models and exploit them to produce accurate NR estimates.

***Index Terms*—** convolutional neural net, no-reference, speech intelligibility, speech quality, wideband

1. INTRODUCTION

Speech quality and intelligibility estimators are critical to the design, optimization, and maintenance of telecommunication networks. Current popular estimators include Perceptual Evaluation of Speech Quality (PESQ) [1], Perceptual Objective Listening Quality Analysis (POLQA) [2], PEMO [3] and Short-Time Objective Intelligibility Measure (STOI) [4]. These full-reference (FR) estimators give values that correlate well with human assessments but require the undistorted reference signal as well as the distorted test signal. This requirement limits the applicability of FR estimators and motivates the development of no-reference (NR) estimators.

NR estimators produce quality or intelligibility values without access to the reference signal. To do this they must explicitly or implicitly embody very broad yet detailed models for undistorted speech. This is an active research area and many different approaches have been proposed over the years. The most recent contributions include novel signal processing techniques [5] and numerous applications of artificial neural networks (ANN) and deep learning techniques.

Most recent work utilizes features derived from traditional speech processing techniques (e.g. magnitude spectrogram, Mel-spectral or Mel-cepstral features, pitch values, voice activity) as inputs to machine learning (ML) processes in order to generate quality or intelligibility estimates [6–16].

Quality-Net [8] uses magnitude spectra as input. In [13] a convolutional neural network (CNN) is applied to 2-D arrays of Mel-cepstral coefficients and other derived features. NISQA accepts super-wideband speech, targets MOS and POLQA [14], multiple subjective dimensions, and gives average per-condition correlations as high as 0.90 [16]. In [16], NISQA builds features from Mel-processed spectrograms (eliminating the need for DCT features used in the earlier NISQA version [14]). NIC-STOI [5] does not use any ANN. It estimates reference speech characteristics to achieve an impressive $\rho = 0.940$ with human intelligibility scores but only for highly constrained data (one language, no speech coding) which is understandable given the high cost of human scores.

Our method and the work in [12] are the only NR evaluation networks where a raw waveform is used as input *instead of calculated features*. Our method achieves superior performance across a much broader scope (see Section 5).

Traditional speech processing techniques have enabled speech compression, noise suppression, and frame erasure concealment. But using those same techniques at the front-end of a quality or intelligibility estimator forces the estimator to operate under the same assumptions as the devices under test, including specific models of speech and distortion as well as limited computational power and memory.

ANNs have limitations as well and are certainly not suited for every application. But CNNs implement a traditional signal processing technique—large banks of arbitrary, customized moving average (MA) filters. This makes CNNs of particular interest and importance for evaluating speech waveforms. We present a 1-D convolutional architecture that can be viewed as a multi-channel signal processor. Each section (other than the first and last) has f_n input and output channels, $f_n \subset \mathbb{Z}_{>0}$. Each input channel is split into f_n copies and each copy is individually filtered by f_n MA filters. The f_n^2 resulting signals are then mixed down to f_n output signals and passed through a non-linear distortion and a non-conventional sample-rate reduction.

This paper will be presented at and appear in the Proceedings of ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 331-335. ICASSP Proceedings are Copyright © IEEE.

S	section type	$\hat{f}_s$ (Hz)	l_{in}	s_l (ms)	l_{out}
S_1	Conv A-2	16,000	48,000	0.0625	8,000
S_2	Conv M-4	8,000	24,000	0.125	6,000
S_3	Conv M-2	2,000	6,000	0.5	3,000
S_4	Conv M-4	1,000	3,000	1	750
S_5	Conv M-3	250	750	4	250
S_6	Conv M-2	83.3	250	12	125
S_7	P Conv M-2	41.7	128	24	64
S_8	Conv M-2	20.8	64	48	32
S_9	Conv A-32	10.4	32	96	1
S_{10}	Dense	—	f_n	—	1

Table 1. WAWEnet architecture: sections S_1–S_9 are composed of one of the three section types listed in Table 2. Number of input and output samples per channel are given by l_{in} and l_{out}, effective sample rate by $\hat{f}_s$, and effective sample spacing by s_l. The dense layer S_{10} maps f_n scalar outputs from S_9 to the final output.

name	Conv A-k	Conv M-k	P Conv M-k
layers			Pad(1, 2)
	C-f_n-f_l	C-f_n-f_l	C-f_n-f_l
	BatchNorm	BatchNorm	BatchNorm
	PReLU-f_n	PReLU-f_n	PReLU-f_n
	AvgPool-k	MaxPool-k	AvgPool-k

Table 2. WAWEnet section types. Each section contains a 1-D convolution layer C-f_n-f_l with f_n filters per channel and f_n channels, filter length $f_l = 3$, stride of 1, and zero padding $\lfloor \frac{f_l}{2} \rfloor = 1$. Padding layer Pad$(a, b)$ prepends a zeros and appends b zeros to the input vector. PReLU-f_n [18] indicates a PReLU activation with f_n parameters. k denotes pooling layer kernel size for Max or Average Pooling layers.

It is prudent to explore signal processing on the CNN's terms and leverage efficiencies afforded by ML technologies to accomplish signal processing tasks—in this case, to learn the proper features to estimate speech quality and intelligibility. In this work we expand on a prior exploration in this area [17] where we constructed a no-reference convolutional waveform-based architecture to mimic narrowband (NB) speech FR quality and intelligibility estimators. The NB Audio Waveform Evaluation network (NAWEnet) learned its own features and estimates quality and intelligibility with high accuracy: per-speech-segment correlation $\rho_{seg} > 0.92$. In Section 2 we describe the architecture of the Wideband Audio Waveform Evaluation Network (WAWEnet), a network that produces quality or intelligibility estimates for NB or wideband (WB) speech. Section 3 describes the speech corpus used for training and evaluation, Section 4 our training methodology, Section 5 our results.

2. NETWORK DESIGN

In [17] we described an architecture framework and designed a specific convolutional architecture, NAWEnet, to accept three-second long NB speech waveforms and estimate quality or intelligibility consistent with one of three FR targets: PESQ, POLQA, or STOI. Naturally, extending NAWEnet to process WB data requires adding a convolutional section that accepts WB input ($16{,}000$ smp/s $\times$ 3 s $= 48{,}000$ smp) to the top of the architecture. This approach would require ≈ 40.5 million parameters, a small increase from the NAWEnet architecture in [17] which has ≈ 40.1 million parameters.

Improving upon this natural extension, we created a much more efficient network by removing the large dense network from the bottom of the architecture and replacing it with three additional convolutional sections and one very small dense section for a total of 10 sections. In addition we trimmed sections with two convolutional layers to have just one convolutional layer. The resulting architecture for WAWEnet is shown in Table 1. Table 2 describes the three section configurations used in S_1–S_9. We successfully reduced the filter length to $f_l = 3$ in all sections and we used $f_n = 96$ channels with $f_n = 96$ filters per channel resulting in a total of 225,025 parameters, about 0.5% of the original number.

3. DATA CORPORA

We used high-quality WB speech recordings recorded in our lab and taken from a variety of other sources including [19–28]; some are conveniently grouped at openslr.org.

We divided these recordings into 100,681 three-second segments (83.9 hours total) called reference segments. These segments represent 13 languages and 1230 different talkers. Mandarin, Spanish, and North American English (NAE) each account for 29% of the reference segments, Korean 6%, African-accented French 3%, Japanese 2%, and the remaining 2% contains Italian, French, German, Portuguese, Hindi, British English, Finnish, Dutch, and Polish.

Each reference segment has an active speech level of 26 ± 0.2 dB below the overload point and speech activity factor of 50% or greater, both determined by the P.56 speech voltmeter found in [29]. Any segment has at most 50% (1.5 sec) content in common with any other segment.

We applied a WB impairment to each of the 100,681 reference segments to produce a corresponding test segment. The set of WB impairments includes 47 different WB speech coding modes taken from EVS, AMR, G.711.1, G.722.1, and G.722. They also include the addition of noise (coffee shop, party, bus, street at 5 or 15 dB SNR) followed with noise suppression by removal of time-frequency components falling below a threshold. We varied the threshold and the processing frame length to achieve noise suppression results ranging from low quality (many artifacts and/or much unsuppressed noise) to moderate quality (few artifacts and modest noise).

This paper will be presented at and appear in the Proceedings of ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 331-335.

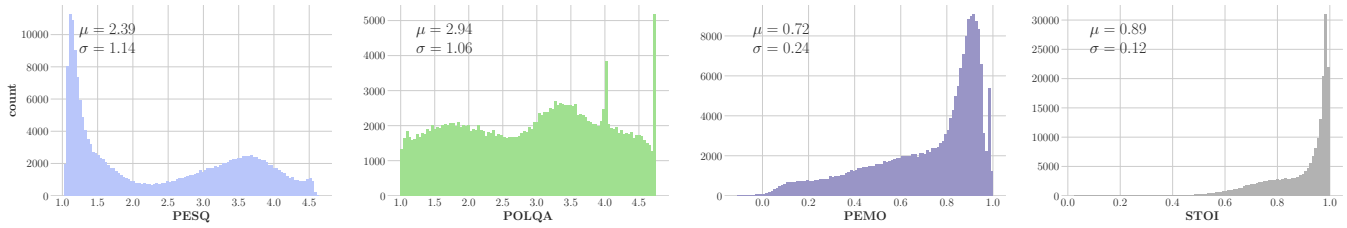

Fig. 1. Histograms, means, and stds. of FR targets over all available data. Lower values indicate lower quality or intelligibility.

Additionally, we applied the high-quality low-pass filter option in [29] to each WB test segment (post-impairment) to create an NB version (with sample rate 16,000 smp/s) resulting in a total of 201,362 test segments.

Because the required amount of subjectively-labeled data was not available, we used FR estimates as training targets to teach WAWEnets to generate NR estimates. Thus we applied FR estimators to each pair of reference-test segments. We used speech-quality estimators PESQ and POLQA, the audio-quality estimator PEMO (software available via [30]), and the intelligibility estimator STOI (developed using $f_s = 10,000$ smp/s but successfully applied to WB speech elsewhere).

Thus each test segment received four target values (PESQ, POLQA, PEMO, and STOI). In the case of an NB test segment, the WB reference segment was used as the reference for each FR estimator. Fig. 1 shows a histogram of each target across all speech segments.

4. TRAINING METHODOLOGY

Starting with 201,362 three-second segments (167.8 hours of speech), we generated an unseen dataset by reserving 10% of each language, to the extent possible. The 127 talkers associated with the unseen dataset do not contribute any segments to the remaining data and are only used for evaluation. The unseen dataset contains 20,782 segments (17.3 hours) of speech. We split the remaining segments into training, testing, and validation datasets (50%, 40%, and 10%, resp.). Thus the training dataset contains 90,299 segments (75.2 hours); the testing dataset contains 72,232 segments (60.2 hours); and the validation dataset contains 18,049 segments (15.0 hours).

As in [17], we performed inverse phase augmentation (IPA) to augment all datasets and to train WAWEnet to learn invariance to waveform phase inversion. This resulted in 335.6 hours of speech data total.

We used affine transformations to map PESQ values ($[1.02, 4.64]$), POLQA values ($[1, 4.75]$), PEMO values ($[0, 1]$), and STOI values ($[0.45, 1]$), to $[-1, 1]$ before use as targets. We used the training process described in [17] with one exception: the mini-batch size used was 60 segments per batch. The Adam optimizer was configured identically as was the learning rate scheduler. After training for 30 epochs, we evaluated the network on the test and unseen datasets separately. The training and testing processes were performed four times to generate one WAWEnet model instance for each target: PESQ, POLQA, PEMO, and STOI. We used PyTorch to construct our datasets and to train and test WAWEnets. Training was done with an NVIDIA GeForce GTX 1070.[1]

5. RESULTS

Training a WAWEnet on one target for 30 epochs with $f_n = 96$ takes about 10.7 hours. This is a 33% reduction from NAWEnet training time [17] in spite of the fact that WAWEnet training uses 26% more speech data.

Table 3 gives the per-speech segment Pearson correlations, ρ_{seg}, and RMSEs for completely unseen data for each WAWEnet on each language, as well as on the combined unseen and test datasets. Fig. 2 shows a 2-D histogram of target quality or intelligibility scores vs. quality or intelligibility scores estimated by the corresponding WAWEnet on unseen data. The ρ_{seg} values range from 0.92 to 0.97 showing that WAWEnets can be trained to agree with quality or intelligibility targets across a broad range of WB and NB speech conditions. These results are superior to any previously reported (see Section 1). In addition, results for different languages within a target are similar and this demonstrates WAWEnet's language robustness. The unseen dataset contains only unseen talkers. RMSE and ρ_{seg} values for unseen and test are very close and this demonstrates robustness to unseen talkers.

We built a WAWEnet with $f_n = 16$ to efficiently test sensitivity to random initialization and random batch selection. We trained for the PESQ target as described in Section 4 a total of 14 times. When evaluated on the unseen dataset the 14 ρ_{seg} values ranged from 0.911 to 0.926 and the RMSE values ranged from 0.434 to 0.473. This demonstrates that the training process is stable but parameter initialization and batch selection can have a measurable impact on performance. In addition, it is impressive that reducing f_n from 96 to 16 still produced $\rho_{seg} > 0.91$ and RMSE < 0.48.

To learn more about combined effects of multichannel filtering (convolution), non-linear processing (PReLU), and non-conventional sample-rate conversion (pooling) we installed probes to measure signal characteristics throughout

[1]Certain products are mentioned in this paper to describe the experiment design. The mention of such entities should not be construed as any endorsement, approval, recommendation, prediction of success, or that they are in any way superior to or more noteworthy than similar entities not mentioned.

This paper will be presented at and appear in the Proceedings of ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 331-335.

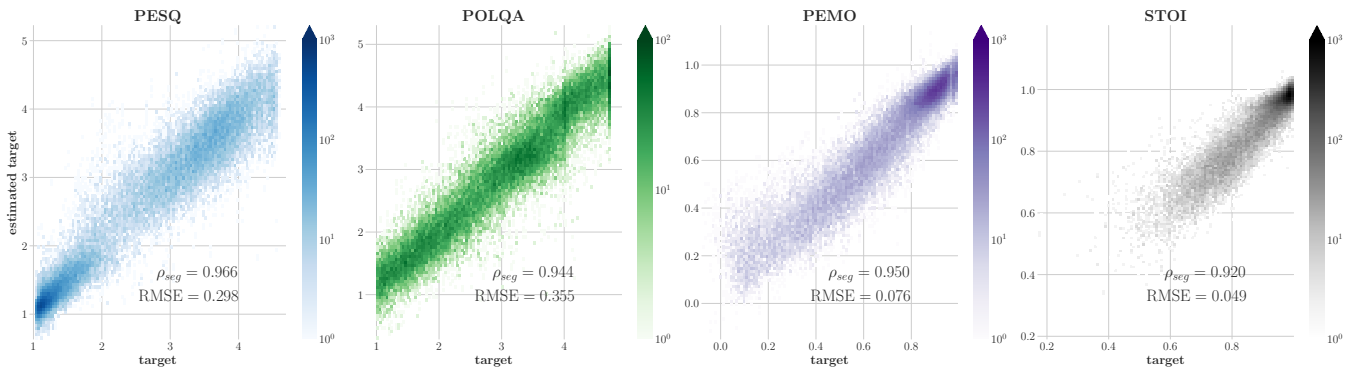

Fig. 2. Two-dimensional histograms showing target vs. estimated values for PESQ, POLQA, PEMO, and STOI when evaluated on the unseen dataset. Number of segments per bin is given by the scale at the right.

		Mandarin	Spanish	NAE	Korean	AA French	Japanese	others	unseen comb.	test comb.
PESQ	ρ_{seg}	0.972	0.968	0.959	0.966	0.977	0.960	0.950	0.966	0.968
	RMSE	0.261	0.299	0.324	0.304	0.272	0.310	0.379	0.298	0.282
POLQA	ρ_{seg}	0.942	0.947	0.947	0.949	0.960	0.909	0.922	0.944	0.950
	RMSE	0.345	0.356	0.344	0.356	0.330	0.407	0.430	0.355	0.337
PEMO	ρ_{seg}	0.952	0.948	0.952	0.950	0.949	0.951	0.949	0.950	0.951
	RMSE	0.072	0.079	0.072	0.092	0.090	0.084	0.081	0.076	0.074
STOI	ρ_{seg}	0.925	0.934	0.909	0.895	0.884	0.947	0.801	0.920	0.916
	RMSE	0.054	0.041	0.048	0.058	0.062	0.044	0.076	0.049	0.049

Table 3. Per-segment Pearson correlation and RMSE achieved on unseen data after training WAWEnet to target PESQ, POLQA, PEMO and STOI separately. Results in "combined" columns reflect evaluation on the aggregated unseen or test dataset.

WAWEnets. Signal power spreads as signals flow through the sections; 99% of signal power is in 71 of 96 channels at the output of Section 1 and is in 87 of 96 channels at the output of Section 9, suggesting that 96 channels is sufficient.

As these 96 signals flow through the nine sections some channels continually accumulate information about the speech-like nature of the input signal and other channels accumulate information about the non-speech-like or distorted nature of the input signal. Signal levels in these "speech channels" are positively correlated to the targets (quality or intelligibility values) while signal levels in the "distortion channels" are negatively correlated to the targets. In all sections after Section 1, roughly half the channels serve each role. The largest correlation magnitudes increase monotonically section-by-section. Example approximate values are 0.50, 0.60, 0.71, 0.82, and 0.94 at the outputs of Sections 1, 3, 5, 7, and 9, respectively. S_9 has 96 outputs and many of these show strong negative or positive correlation to the target. The final WAWEnet output is an optimized linear combination of these (plus bias) and thus has even higher correlation to the target.

WAWEnets successfully target four different FR quality or intelligibility estimators. Given the WAWEnet performance on these different tasks (the four targets have very different distributions), it stands to reason that given appropriate training data, WAWEnets could also directly target mean opinion score (MOS) or specific dimensions of speech quality, such as noisiness, coloration, and discontinuity. We have demonstrated extending a WEnet from NB to WB and we expect that extending to super wideband and fullband can follow suit. Labeled data would be the main constraint.

6. CONCLUSION

We adapted the WEnets framework described in [17] to provide speech quality and intelligibility estimates for WB speech. This required adding a convolutional section and reorganizing two lower sections, reducing parameter count by $> 99\%$. Per-segment correlations between 0.92 and 0.97 demonstrate that NR WAWEnets accurately follow FR targets across 13 languages, over 17 hours of NB and WB speech from 127 talkers that was completely unseen during training, validation, and testing, and could likely follow a true MOS target. Future opportunities include further dissection of the inner workings of both NAWEnets and WAWEnets and publishing implementations at https://github.com/NTIA/WEnets. Extending WAWEnets to super-wideband and fullband speech applications as well as synthetic speech applications should be possible given commensurate training data. It may be possible to further prune WAWEnets parameters by following the successes of image-targeted convolutional architectures or by learning auto-regressive moving-average filters.

This paper will be presented at and appear in the Proceedings of ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 331-335.

7. REFERENCES

[1] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ)," Geneva, 2001.

[2] ITU-T Recommendation P.863, "Perceptual objective listening quality analysis," Geneva, 2018.

[3] R. Huber and B. Kollmeier, "PEMO-Q – A new method for objective audio quality assessment using a model of auditory perception," *IEEE Trans. ASLP*, vol. 14, no. 6, pp. 1902–1911, Nov. 2006.

[4] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. ASLP*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[5] C. Sørensen, J. B. Boldt, and M. G. Christensen, "Validation of the non-intrusive codebook-based short time objective intelligibility metric for processed speech," in *Proc. Interspeech 2019*, pp. 4270–4274.

[6] T. H. Falk and W. Chan, "Single-ended speech quality measurement using machine learning methods," *IEEE Trans. ASLP*, vol. 14, no. 6, pp. 1935–1947, Nov. 2006.

[7] M. H. Soni and H. A. Patil, "Novel deep autoencoder features for non-intrusive speech quality assessment," in *Proc. EUSIPCO 2016*, pp. 2315–2319.

[8] S. Fu, Y. Tsao, H. Hwang, and H. Wang, "Quality-Net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," in *Proc. Interspeech 2018*, pp. 1873–1877.

[9] H. Salehi, D. Suelzle, P. Folkeard, and V. Parsa, "Learning-based reference-free speech quality measures for hearing aid applications," IEEE Trans. ASLP, vol. 26, no. 12, pp. 2277–2288, Dec. 2018.

[10] C. Spille, S. D. Ewert, B. Kollmeier, and B. T. Meyer, "Predicting speech intelligibility with deep neural networks," *Computer Speech & Language*, vol. 48, pp. 51 – 66, Mar. 2018.

[11] R. Huber, M. Krüger, and B. T. Meyer, "Single-ended prediction of listening effort using deep neural networks," *Hearing Research*, vol. 359, pp. 40 – 49, Mar. 2018.

[12] P. Seetharaman, G. J. Mysore, P. Smaragdis, and B. Pardo, "Blind estimation of the speech transmission index for speech quality prediction," in *Proc. ICASSP 2018*, pp. 591–595.

[13] H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, "Intrusive and non-intrusive perceptual speech quality assessment using a convolutional neural network," in *Proc. 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 85–89.

[14] G. Mittag and S. Möller, "Non-intrusive speech quality assessment for super-wideband speech communication networks," in *Proc. ICASSP 2019*, pp. 7125–7129.

[15] J. Ooster and B. T. Meyer, "Improving deep models of speech quality prediction through voice activity detection and entropy-based measures," in *Proc. ICASSP 2019*, pp. 636–640.

[16] G. Mittag and S. Möller, "Quality degradation diagnosis for voice networks — Estimating the perceived noisiness, coloration, and discontinuity of transmitted speech," in *Proc. Interspeech 2019*, pp. 3426–3430.

[17] A. A. Catellier and S. D. Voran, "WEnets: A Convolutional Framework for Evaluating Audio Waveforms," *arXiv e-prints*, arXiv:1909.09024, Sep. 2019.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on Imagenet classification," in *Proc. 2015 IEEE Int. Conf. on Computer Vision*, pp. 1026–1034.

[19] ITU-T "P series supplement 23 speech database," Geneva, 1998.

[20] ITU-T Recommendation P.501, "Test signals for use in telephonometry," Geneva, 2017.

[21] Telecommunications and Signal Processing Laboratory Speech Database, http://www-mmsp.ece.mcgill.ca/Documents/Data/.

[22] J. S. Garofolo, et. al., "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," 1993.

[23] A. Rousseau , P. Deléglise, and Y. Estève , "TED-LIUM: An automatic speech recognition dedicated corpus," in *Proc. Eighth Int. Conf. on Language Resources and Evaluation*, May 2012.

[24] C. D. Hernandez-Mena, "TEDx Spanish Corpus. Audio and transcripts in Spanish taken from the TEDx Talks," 2019, http://openslr.org/67.

[25] D. Wang, X. Zhang, and Z. Zhang, "THCHS-30: A free Chinese speech corpus," 2015. http://openslr.org/18.

[26] Y. Choi and B. Lee, "Pansori: ASR corpus generation from open online video contents," 2018, http://openslr.org/58.

[27] Recordings of African Accented French speech, http://openslr.org/57.

[28] Open Speech Repository,
https://www.voiptroubleshooter.com/.

[29] ITU-T Recommendation G.191, "Software tools for speech and audio coding standardization," Geneva, 2005.

[30] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. ASLP*, vol. 19, no. 7, pp. 2046–2057, Sep. 2011.

This paper will be presented at and appear in the Proceedings of ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 331-335.