INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Towards achieving robust universal neural vocoding

Jaime Lorenzo-Trueba[1], *Thomas Drugman*[1], *Javier Latorre*[1*], *Thomas Merritt*[1], *Bartosz Putrycz*[1], *Roberto Barra-Chicote*[1], *Alexis Moinet*[1], *Vatsal Aggarwal*[1]

[1]Amazon.com, Cambridge, United Kingdom

{truebaj, drugman, jlatorre, thommer, bartosz, rchicote, amoinet, agvatsal}@amazon.com

Abstract

This paper explores the potential universality of neural vocoders. We train a WaveRNN-based vocoder on 74 speakers coming from 17 languages. This vocoder is shown to be capable of generating speech of consistently good quality (98% relative mean MUSHRA when compared to natural speech) regardless of whether the input spectrogram comes from a speaker or style seen during training or from an out-of-domain scenario when the recording conditions are studio-quality. When the recordings show significant changes in quality, or when moving towards non-speech vocalizations or singing, the vocoder still significantly outperforms speaker-dependent vocoders, but operates at a lower average relative MUSHRA of 75%. These results are shown to be consistent across languages, regardless of them being seen during training (e.g. English or Japanese) or unseen (e.g. Wolof, Swahili, Ahmaric).

Index Terms: Neural Vocoder, Text-to-speech, Scalability, Statistical Waveform Speech Synthesis

1. Introduction

Statistical parametric speech synthesis (SPSS) has seen a paradigm change recently, mainly thanks to the introduction of a number of autoregressive models [1, 2, 3, 4, 5, 6], turning into what can be termed statistical speech waveform synthesis (SSWS) [5]. This change has closed the gap in naturalness between statistical text to speech (TTS) and natural recordings whilst maintaining the flexibility of statistical models.

In the case of traditional vocoding [7, 8, 9, 10], approaches commonly relied on simplified models (e.g. source-filter model [11]) that were defined by acoustic features such as voicing decisions, the fundamental frequency (F0), mel-generalized cepstrum (MGC) or band aperiodicities. The quality of those traditional vocoders was limited by the assumptions made by the underlying model and the difficulty to accurately estimate the features from the speech signal [12, 13].

Traditional waveform generation algorithms, while capable of generating speech from their spectral representation such as Griffin-Lim [14], are not capable of generating speech with acceptable naturalness. This is due to the lack of phase information in the short-time Fourier transform (STFT).

Neural vocoders are a data-driven method where neural networks learn to reconstruct an audio waveform from acoustic features [1, 2, 15, 6]. They allow us to overcome the shortcomings of traditional methods [16] at a very significant cost in computation power and data requirements. However, due to sparsity (it is unlikely that we will ever be able to cover all possible human-generated sounds in the training data) the neural vocoder models are prone to over-fit to the training speaker characteristics and have poor generalization capabilities [17]. Several recent studies attempted to improve the adaptation capabilities of such models [18, 19], commonly using explicit speaker information (either as a onehot encoding or some other form of speaker embedding) [20]. There are however reports in literature of initial successes training neural vocoders without providing explicit speaker information [21, 22], however the investigation either did not provide significant improvements in terms of robustness or did not cover the details on how the model handles changes in domain or unseen speakers.

This contributions of this paper are: 1) we demonstrate that a speaker encoding is not required to train a high-quality Speaker-Independent (SI) WaveRNN-based [2] neural vocoder; 2) our SI neural vocoder can effectively synthesise speakers that were unseen during training, which is not possible with vocoders trained with explicit speaker information or with a speaker-dependent approach; 3) we study the robustness and potential universality of our SI neural vocoder on a large diversity of unseen conditions (e.g. language, phonation, noise or speaking style).

2. System description

Even though CNN-based systems have been thoroughly researched and real-time implementations have been proposed [4, 23], it is known that they are prone to instabilities [24] which occasionally affect perceptual quality. RNN-based systems, on the other hand, can be expected to provide a more stable output due to the persistence of the hidden state, at least when vocoding, in which context is not critical beyond the closest spectrograms (a known characteristic of RNNs).

The structure of the neural vocoder system used in this paper (heavily inspired by WaveRNN [2], only with minor changes in the conditioning network) is described in Figure 1. We refer to this system as RNN_MS. The autoregressive side consists of a single forward GRU (hidden size of 896) and a pair of affine layers followed by a softmax layer with 1024 outputs, predicting the 10-bit mu-law samples for a 24 kHz sampling rate. The conditioning network consists of a pair of bidirectional gated recurrent units (GRUs) with a hidden size of 128. The mel-spectrograms used for conditioning the network were extracted using Librosa library [25], with 80 coefficients and frequencies ranging from 50 Hz to 12 kHz.

We trained system RNN_MS in 4 different configurations, whose details are shown in Table 1. First three SD systems were trained on American English speakers, two female (F1 & F2) and 1 male (M1) from our internal corpora.

We also trained 3 multi-speaker vocoders, one with all the training data from the 3 SD voices ($3Spk$), another one with 7 American English speakers ($7Spk$) comprising 4 females, 2 males and 1 child but with restricted amounts of training data per speaker (5000 utterances). This $7Spk$ neural vocoder aims

*: Work performed while at Amazon.com, currently associated with Apple Inc., UK.

http://dx.doi.org/10.21437/Interspeech.2019-1424

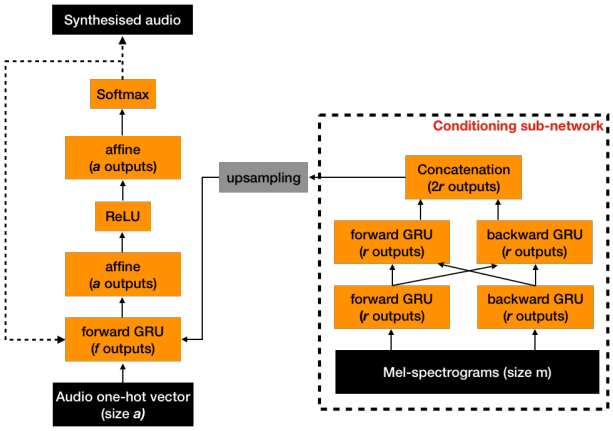

Figure 1: *Block diagram of system RNN_MS*

Table 1: *Summary of the training data of the different RNN-based vocoders.*

Vocoder	Speakers	Utterances	Language
F1 (SD)	1	22000	US English
F2 (SD)	1	15000	US English
M1 (SD)	1	15000	US English
3spk	3	52000	US English
7spk	7	35000	US English
Univ	74	149134	Multiple (17)

to check whether variability or data (i.e. $3Spk$) are more important for robustness in general. Finally we trained what is introduced as our universal neural vocoder with 74 different voices, 22 males and 52 females, extracted from 17 languages, with approx. 2000 utterances per speaker. This neural vocoder was designed with the expectation of being generalizable to any incoming speaker regardless of whether it was seen during training or not.

3. Experimental protocol

To properly characterize the generalization capabilities of the different vocoders in terms of naturalness we considered a number of scenarios, but always considering oracle spectrograms directly extracted from recordings. First of all a topline scenario in which we generated speech from speakers present in the training data of all the vocoders, but with utterances not seen during training (section 4.1). Then, we also generated speech in scenarios partially out-of-domain from the training data: a mixture of male and female neutral speakers extracted from VCTK [26] for English or from the NITech Japanese samples database [27]. We also considered audiobook speech extracted from Blizzard2016 development set [28], which was out of domain in terms of speaker, speaking style but as in all previous cases, recorded with studio-quality.

Finally we considered a number of out-of-domain scenarios ranging from: i) different voice qualities [29], ii) irregular recording conditions (i.e. background noise [30], reverberation [31], or both [32]), iii) unseen languages (Ahmaric, Swahili and Wolof) recorded in sub-optimal recording situations [33] (i.e. significant reverberation, or poor quality audio), iv) singing extracted from publicly available music corpora [34], v) non-speech vocalizations [35]. The naturalness perceptual evaluation was designed as a MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) test [36], where the participants were presented with the systems being evaluated side-by-side, asked

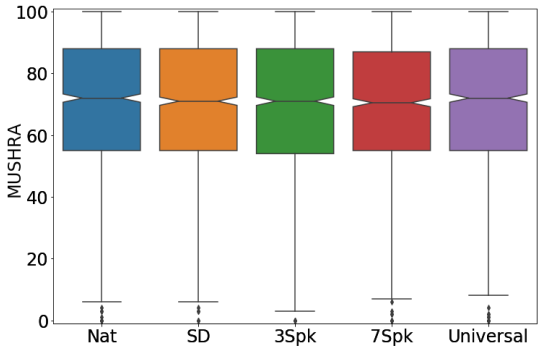

Figure 2: *MUSHRA evaluation for the in-domain speakers.*

to rate them in terms of naturalness from 0 (very poor) to 100 (completely natural), but modified so as not to force at least one 100 rated system. The test consisted of 200 randomly-selected utterances, not included in the training data. Evaluations were conducted with self-reported native American English speakers using Amazon Mechanical Turk. 50 listeners participated in each evaluation, balanced so that every utterance was rated by 5 listeners, each rating 20 screens.

Paired Student T-tests with Holm-Bonferroni correction were used to validate the statistical significance of the differences between systems, considering it validated when $p - value < 0.01$. We use the ratio between the mean MUSHRA score of a system and natural speech, we refer to this as 'relative MUSHRA', to illustrate the gap with the reference.

4. Results

4.1. In-domain speakers and style

This evaluation considered 2 female and 1 male speaker (the ones used to train the $3Spk$ vocoder). The results in Figure 2 show that there is no significant difference in terms of evaluated naturalness when using any of the trained vocoders as long as the speakers were part of the training data. This is a strong result for the proposed universal vocoder, as it showed no degradation when compared to the highly specific SD neural vocoder. Moreover, while there was a statistically significant difference between vocoded and natural naturalness scores, it was minimal (98.5% relative MUSHRA). It must be noted that while there were inter-speaker differences, those did not affect the rank-order, so results are presented as averages.

4.2. Robustness to unseen and out-of-domain speakers

In this evaluation, we considered out of domain speakers for which some of the defining aspects were still part of the training corpus. That is, out of domain speakers but recorded in a studio scenario, stretching it further by considering a children audiobook scenario but from a professional voice talent [28].

In this scenario SD vocoders were not available. As such, results are expectedly poor in comparison to some of the more general neural vocoders. They were included as a bottom anchor and selected by looking for the one trained with the speaker most similar to the target speaker. Similarity was measured by training a number of multi-variate Gaussian Mixture Models (GMMs) of the training data of the different vocoders and of the target speaker, then obtaining the Kullback-Leibler divergence (KLD) between the GMMs.

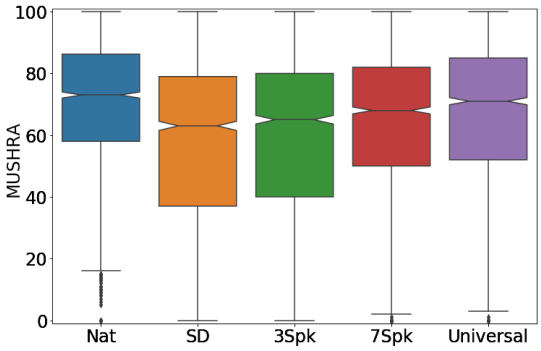

Figure 3: *MUSHRA evaluation for the English, neutral, out-of-domain speakers.*

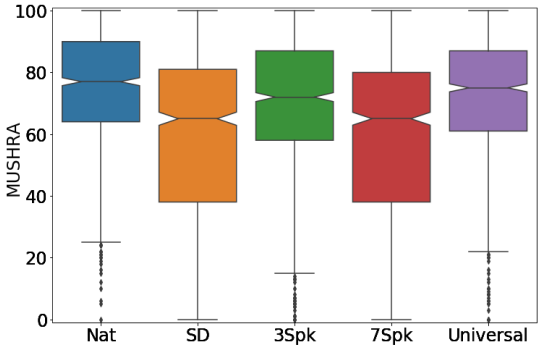

Figure 4: *MUSHRA evaluation for the audiobook data.*

4.2.1. English speakers

Results (Figure 3) show that the more variety in number of training speakers the better the quality, to the point where $Univ$ is capable of providing practically the same relative MUSHRA score as for in-domain speakers (98% vs. 98.5%). This speaks very strongly about the generalization capabilities of such a system. Moreover, we can see how the vocoder trained with more speakers but with less training data (7Spk) is capable of providing better quality than the other two systems (SD and 3Spk), suggesting that variability is more important than quantity for generalization.

4.2.2. Japanese speakers

We carried out an evaluation with out-of-domain Japanese speakers, which is an in-domain language, extracted from the NITech Japanese samples database [27]. Results were similar to those in English (98% relative MUSHRA).

4.2.3. Audiobook style speaker

In the case of highly expressive data, including disfluencies and onomatopoeias, (see Figure 4) the universal vocoder is still capable of proving steady quality, once again maintaining the relative MUSHRA scores at 98%. Both SD and 7Spk show comparatively poor performance, while 3Spk breaks the trend. This is confirmed by the KLD between the audiobook speaker and those of the vocoders (2.64 against Univ, 5.42 against 3Spk, 14.45 against 7Spk and 14.62 against SD). All in all reinforcing the hypothesis that the dissimilarity between training and testing speakers is critical for performance.

Table 2: *Summary of the results for the unseen scenarios. 'Rev.' stands for reverberation and 'Vocal.' for vocalizations.*

	SD	Univ	Nat	SD Rel.	Univ Rel.
Breathy	38.4	61.9	67.6	56.8%	91.6%
Pressed	30.9	63.4	70.9	43.5%	89.5%
Noisy (N)	37.5	58.2	73.4	51.1%	79.4%
Rev. (R)	35.5	56.2	73.6	48.2%	76.4%
N+R.	23.0	39.7	68.7	33.5%	57.8%
African	34.5	55.4	70.9	48.6%	78.1%
Singing	41.2	52.3	72.6	56.8%	72.0%
Vocal.	24.9	48.0	73.9	33.7%	64.9%

4.3. Robustness to unseen scenarios

For the additional evaluations, we did not consider all possible vocoders and restricted the exploration to a lower anchor (SD systems, selected as in Section 4.2), an upper anchor (natural speech) and the proposed $Univ$ system.

Table 2 summarizes the results over the various unseen scenarios. It can be observed that the $Univ$ model significantly ($p < 0.01$) improves over the SD vocoder, with a relative MUSHRA gain varying between 15% and 45%. Despite this improvement, the proposed SI vocoder is not yet capable of providing a consistently high fidelity across all unseen scenarios, with a relative MUSHRA score falling down to 58%.

4.3.1. Robustness to voice quality

Results in terms of voice quality (Breathy and Pressed in Table 2) appear to be relatively robust, with 91.6% and 89.5% relative MUSHRA respectively. This is a slight degradation compared to the normal phonation style provided with the corpus [29] (96.3%, not shown in Table 2), but to a much lesser extent than for the SD model. The drop in relative MUSHRA compared to the clean recordings in Section 4.2 most likely happens due to the data having been recorded at 16kHz, and due to an overall lower quality in the source material, with some clicks appearing in the end of recordings that are amplified in the re-synthesis process.

4.3.2. Robustness to signal quality

Performance falls to about 78% relative MUSHRA in the noisy or reverberant conditions (Noisy and Reverb. in Table 2 respectively), and even lower (58%) in a combination of both (N+R). The $Univ$ system however provides a comparable quality for noisy recordings regardless of them being in English (79%) or in unseen African languages (78%), which suffered from either reverberation or considerable background noise due to the poor recording conditions.

The degradation in quality seems to be caused by distortion appearing in the re-synthesised material, as the vocoder did not seem to have learned how to properly render non-human sounds such as background noise or echo. This distortion ranges from a strong vibrato-like effect appearing in the case of reverberating samples, to distorted speech when attempting to generate the background noise.

4.3.3. Robustness to singing

The vocoder was capable of handling singing re-synthesis (Singing in Table 2) with an average performance of 72% relative MUSHRA. A closer analysis of the results show some

significant trend and differences depending on the style: clean singing (e.g. songwriter music, ballads...) performed at an average of 94.5% relative MUSHRA, comparable to the results achievable with clean speech. Conversely, singing styles that rely on distortion (e.g. rock, pop) perform at a much lower quality (39.3%). This correlates with the results achieved for conventional speech, suggesting that the underlying issue is voice quality rather than style. An additional observation is that tracks with multiple simultaneous voices are rendered with lower quality compared to a single voice.

4.3.4. Robustness to non-speech vocalizations

The results, summarized by Vocal. in Table 2, vary significantly with the kind of vocalization. While sounds of anger or achievement, represented as grunts or shouts in this dataset, perform with a poor average relative MUSHRA of 47.7%, sounds of disgust or pleasure got an average of 77.9%. This is probably due to the energy bursts present in the grunts and shouts, which are generated as heavily distorted sounds.

5. Discussion

Our experimental results in Section 4 have highlighted a few shortcomings to overcome. The $Univ$ system is not yet robust to noise or reverberation in the source materials, is sensitive to extreme energy bursts (shouts or grunts) and is not capable of properly generating spectrograms with multiple overlapping speakers. In these unseen scenarios, the proposed vocoder is capable of significantly outperforming a SD system (between 15% and 45% higher relative MUSHRA), but also introduces some distortion which substantially impairs the quality compared to that achieved in clean situations. Nonetheless, it is worth emphasizing that in studio-quality recordings, the proposed $Univ$ vocoder achieved a high fidelity of 98% relative MUSHRA consistently across seen or unseen languages and styles. Those are promising clues showing that the generalization capabilities of the model can go way beyond simply replicating training conditions.

6. Conclusions

We have introduced a robust neural vocoder conditioned on mel-spectrograms, without any form of speaker encoding. The system was evaluated with an exhaustive framework, attempting to cover a very diverse range of in and out-of domain scenarios.

Our results suggest that the proposed vocoder, trained on varied materials (74 speakers and 17 languages, all recorded in studio conditions) can significantly outperform speaker-dependent vocoders in clean unseen scenarios (relative MUSHRA score of 98%). This is likely due to the variety seen during training, allowing the vocoder to generalize better to unseen scenarios, including singing, non-speech vocalizations or low-quality signals, achieving an average relative MUSHRA score of 72%.

Achieving a truly universal neural vocoder would allow for future work to focus on spectrogram estimation from text to any new speaker, language or style without being constrained by training a specific neural vocoder. But there is still room for improvement in terms of training data diversity and model expressiveness before we can claim the universality goal is achieved. The path towards that goal goes through understanding what training material will teach our vocoding systems to universally generalize.

7. References

[1] A. van den Oord, S. Dieleman, H. Zen, et al., "Wavenet: A generative model for raw audio," *CoRR abs/1609.03499*, 2016.

[2] N. Kalchbrenner, E. Elsen, K. Simonyan, et al., "Efficient neural audio synthesis," *arXiv preprint arXiv:1802.08435*, 2018.

[3] J. Shen, R. Pang, R. J. Weiss, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," *arXiv preprint arXiv:1712.05884*, 2017.

[4] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," *arXiv preprint arXiv:1807.07281*, 2018.

[5] T. Merritt, B. Putrycz, A. Nadolski, et al., "Comprehensive evaluation of statistical speech waveform synthesis," in *SLT*, 2018.

[6] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," *arXiv preprint arXiv:1811.00002*, 2018.

[7] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," *Proc. MAVEBA*, pp. 13–15, 2001.

[8] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE transactions on information and systems*, vol. E99-D, no. 7, pp. 1877–1884, 2016.

[9] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *IEEE TASLP*, vol. 20, no. 3, pp. 968–981, 2012.

[10] M. W. Macon and M. A. Clements, "Speech concatenation and synthesis using an overlap-add sinusoidal model," in *Proc. ICASSP*, 1996, ICASSP '96, pp. 361–364.

[11] G. Fant and Q. Liljencrants, J.and Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, no. 1985, pp. 1–13, 1985.

[12] T. Merritt, T. Raitio, and S. King, "Investigating source and filter contributions, and their interaction, to statistical parametric speech synthesis," in *Interspeech*, 2014.

[13] T. Merritt, J. Latorre, and S. King, "Attributing modelling errors in HMM synthesis by stepping gradually from natural to modelled speech," in *ICASSP*, 2015.

[14] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[15] Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu, "Fftnet: A real-time speaker-dependent neural vocoder," in *Proc. ICASSP 2018*. IEEE, 2018, pp. 2251–2255.

[16] X. Wang, J. Lorenzo-Trueba, S. Takaki, et al., "A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis," in *Proc. ICASSP*. IEEE, 2018, pp. 4804–4808.

[17] S. Ö. Arık, H. Jun, and G. Diamos, "Fast spectrogram inversion using multi-head convolutional neural networks," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 94–98, 2019.

[18] X. Wu, Y. Cao, M. Wang, et al., "Rapid style adaptation using residual error embedding for expressive speech synthesis," *Proc. Interspeech 2018*, pp. 3072–3076, 2018.

[19] B. Sisman, M. Zhang, and H. Li, "A voice conversion framework with tandem feature sparse representation and speaker-adapted wavenet vocoder," *Proc. Interspeech 2018*, pp. 1978–1982, 2018.

[20] L. Liu, Z. Ling, Y. Jiang, et al., "Wavenet vocoder with limited training data for voice conversion," *Proc. Interspeech 2018*, pp. 1983–1987, 2018.

[21] T. Hayashi, A. Tamamori, K. Kobayashi, et al., "An investigation of multi-speaker training for wavenet vocoder," in *Proc. ASRU 2017*. IEEE, 2017, pp. 712–718.

[22] Y. Jia, Y. Zhang, R. Weiss, et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in Neural Information Processing Systems*, 2018, pp. 4485–4495.

[23] A. van den Oord, Y. Li, I. Babuschkin, et al., "Parallel wavenet: Fast high-fidelity speech synthesis," *arXiv preprint arXiv:1711.10433*, 2017.

[24] Y. Wu, K. Kobayashi, T. Hayashi, et al., "Collapsed speech segment detection and suppression for wavenet vocoder," *Proc. Interspeech 2018*, pp. 1988–1992, 2018.

[25] B. McFee, C. Raffel, D. Liang, et al., "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, pp. 18–25.

[26] J. Yamagishi and K. Edwards, "Voice cloning toolkit for festival and hts," 2010.

[27] H. Zen, T. Nose, J. Yamagishi, et al., "The hmm-based speech synthesis system (hts) version 2.0.," in *SSW*, 2007, pp. 294–299.

[28] S. King and V. Karaiskos, "Blizzard Challenge 2016," 2016, http://www.festvox.org/blizzard/.

[29] M. Airas and P. Alku, "Comparison of multiple voice source parameters in different phonation types," in *Proc. Interspeech*, 2007.

[30] C. Valentini-Botinhao et al., "Noisy speech database for training speech enhancement algorithms and tts models," 2017.

[31] C. Valentini-Botinhao et al., "Reverberant speech database for training speech dereverberation algorithms and tts models," 2016.

[32] C. Valentini-Botinhao et al., "Noisy reverberant speech database for training speech enhancement algorithms and tts models," 2017.

[33] E. Gauthier, L. Besacier, S. Voisin, et al., "Collecting resources in sub-saharan african languages for automatic speech recognition: a case study of wolof," in *Proc. LREC 2016*, 2016.

[34] A. Liutkus, F.-R. Stöter, Z. Rafii, et al., "The 2016 signal separation evaluation campaign," in *Proc. LVA/ICA 2015*, Cham, 2017, pp. 323–332.

[35] C. F. Lima, S. L. Castro, and S. K. Scott, "When voices get emotional: a corpus of nonverbal vocalizations for research on emotion processing," *Behavior research methods*, vol. 45, no. 4, pp. 1234–1245, 2013.

[36] I. Recommendation, "Bs. 1534-1. method for the subjective assessment of intermediate sound quality (mushra)," *International Telecommunications Union, Geneva*, 2001.