ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095923

THE MBSTOI BINAURAL INTELLIGIBILITY METRIC USING A CLOSE-TALKING MICROPHONE REFERENCE

Pierre Guiraud, Alastair H. Moore, Rebecca R. Vos, Patrick A. Naylor, Mike Brookes

Department of Electrical and Electronic Engineering, Imperial College London,
London SW7 2AZ, United Kingdom

ABSTRACT

Intelligibility metrics are a fast way to determine how comprehensible a target signal is in a noisy situation. Most metrics however rely on having a clean reference signal for computation and are not adapted to live recordings. In this paper the deep correlation modified binaural short time objective intelligibility metric (Dcor-MBSTOI) is evaluated with a single-channel close-talking microphone signal as the reference. This reference signal inevitably contains some background noise and crosstalk from non-target sources. It is found that intelligibility is overestimated when using the close-talking microphone signal directly but that this overestimation can be eliminated by applying speech enhancement to the reference signal.

***Index Terms*—** intelligibility metric, MBSTOI, Machine-Learning, close-talking microphone

1. INTRODUCTION

Determining the intelligibility of a speaker in a noisy situation comes naturally to a human listener but is challenging to estimate computationally. Computed intrusive intelligibility metrics are a fast alternative to in-person tests but rely on having a clean reference signal available to compare with the noisy signal [1, 2, 3, 4, 5, 6].

One of the most widely used single-channel intrusive intelligibility metrics is the short time objective intelligibility metric (STOI) [7]. It has been further developed over the years [8, 9], and its latest implementation called modified binaural STOI (MBSTOI) [10] extends its applicability to binaural signals by adding an equalisation-cancellation (EC) stage [11]. However, the requirement of a clean binaural reference signal makes it challenging to use MBSTOI except in controlled scenarios.

In recent years machine learning (ML) and deep learning (DL) based techniques [12, 13] have enabled the creation of new end-to-end metrics [14, 15, 16] and the improvement of existing ones [17, 18, 19]. Recently, deep correlation MBSTOI (Dcor-MBSTOI) [20, 21] allowed the computation of MBSTOI when only a clean single-channel reference was available.

In this paper Dcor-MBSTOI is evaluated using a close-talking microphone (CM) to provide the reference signal to determine whether this metric is applicable to live recordings. Sec. 2 describes the metrics MBSTOI, Dcor-MBSTOI and the dataset used for evaluation. The performance of Dcor-MBSTOI is presented in Sec. 3 and conclusions are drawn in Sec. 4.

2. METRICS AND SIMULATION

2.1. MBSTOI and Dcor-MBSTOI

The calculation of MBSTOI and Dcor-MBSTOI is illustrated in Fig. 1. MBSTOI uses a clean binaural reference signal containing the target speech components and compares it with a noisy binaural signal to predict an intelligibility value between 0 and 1 [10]. To take into account binaural information, the EC stage [11] uses binaural cues to find parameters that align and cancel undesired localised interference. These parameters represent interaural level and time differences (ILD and ITD). In each time analysis frame and third-octave band, a grid-search is performed to determine the parameter values that maximize the correlation between the reference and noisy signals.

Dcor-MBSTOI uses the same noisy binaural signal but only requires a clean single-channel reference signal to be available. It aims to reproduce the value of MBSTOI as shown in Fig. 1. In Dcor-MBSTOI, the EC stage is replaced by a deep neural network comprising two 2D convolutional layers, of sizes 8 and 16, and two linear layers, of sizes 256 and 1, which directly estimates the correlation coefficients in every time analysis frame and third-octave band. More details about the structure and training hyperparameters can be found in [21].

The preprocessing stage of MBSTOI shown in Fig. 1 applies an energy-based voice-activity detector (VAD) to the reference signal in order to identify and remove analysis frames in which the target talker is silent. To ensure that all versions of the metric use the same set of analysis frames, we have used the VAD from MBSTOI in all evaluations. This VAD is based on the clean binaural signal and so, in practice, the VAD would instead use either the clean single-channel

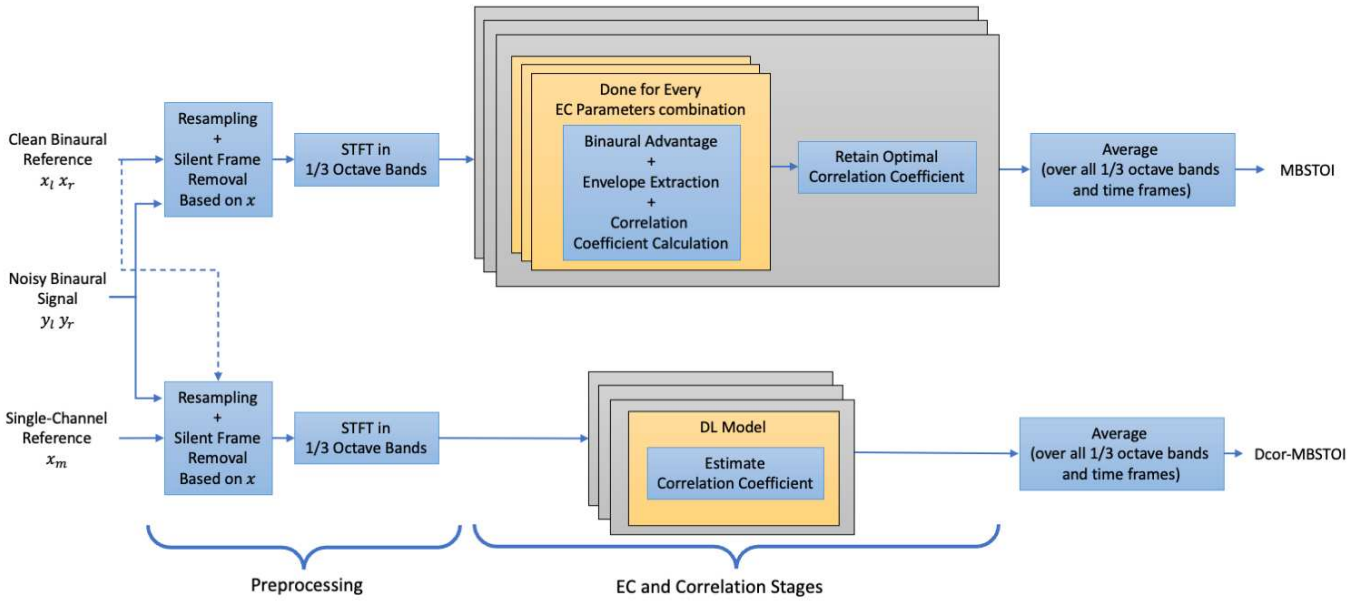

Fig. 1. MBSTOI and Dcor-MBSTOI computation block diagram. In preprocessing, Dcor-MBSTOI can use the clean binaural signal to have perfect silent frame removal. Dcor-MBSTOI uses DL to directly estimate the correlation coefficient of the current analysis time frame and third-octave band.

signal or the close-talking microphone signal. Since both these signals would normally have a high signal to noise ratio (SNR), we do not expect that their use would have a significant effect on the VAD decisions.

2.2. Simulated scenes

Table 1. Simulated scene parameters

	Target	Interferer
Distance (m)	1	1
Level (dB)	60	[50, 51, ..., 70]
Azimuth (degrees)	[-30, 0, 30]	[0, 22.5, ..., 180]
Elevation (degrees)	[-45, 0, 45]	[-45, 0, 45]

In order to test the metrics, 250 sentences were generated using the `tascar` software package [22]. The listener is located at the origin of the spatial coordinate system and is facing forwards at zero azimuth and elevation. The target speaker is located in front of the listener at one of the nine positions listed in Table 1. In addition to diffuse babble noise, a localised interferer is located somewhere around the listener. Potential locations of target and interferer are listed in Table 1. This takes place in empty rooms of varying size with reverberation times of up to 3 s. The reverberation time is below 0.42 s in over 50 % of cases and above 1 s in 20 % of cases.

The target speech signal is randomly chosen from the IEEE speech corpus [23], UK recordings. The interferer can be another sentence from the same corpus or else a localised noise from the PNL100 non-speech noise corpus [24]. Diffuse noise consist of babble noise recordings from crowded bar. As described in [21], the DL stage in Dcor-MBSTOI was trained with data generated from the same set of scenarios using a clean single-channel reference and was not retrained for the experiments discussed below.

The scene includes a binaural microphone pair and two single-channel omnidirectional microphones. The binaural microphone represents the listener and incorporates the main features of a measured head related transfer function (HRTF). One of the omnidirectional microphones is positioned at the origin, coincident with the listener, while the other is placed 10 cm in front of the target speech location in the direction of the listener. We refer to the signals from these microphones as the single-channel and close-talking (CM) signals respectively. Each scene is simulated twice: a clean version which includes only the anechoic target talker and a noisy version which additionally includes the interferer, the diffuse noise and reverberation. As shown in Fig. 1, MBSTOI compares the noisy binaural signal with the clean binaural signal as a reference. In contrast, Dcor-MBSTOI compares the noisy binaural signal with one of the single-channel signals as a reference. The frequency weighted segmental SNR (fwSegSNR) of the noisy CM is ranging from -4 dB to 8 dB. The effect of choosing different references for Dcor-MBSTOI is discussed in Sec. 3.

3. ANALYSIS

3.1. MBSTOI reproduction

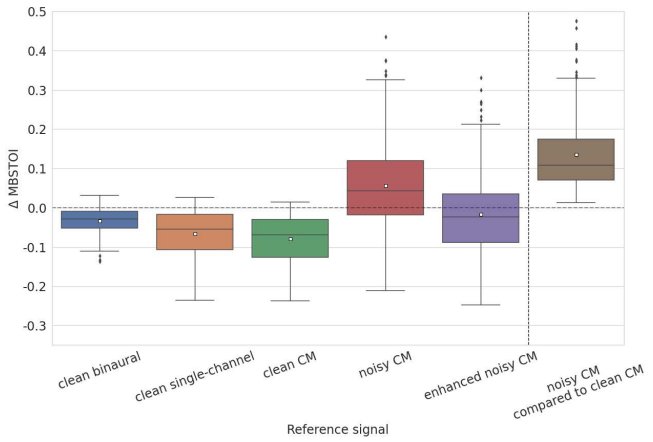

Fig. 2. Box plot representing ΔMBSTOI, the difference between Dcor-MBSTOI and MBSTOI, when using different reference signals for Dcor-MBSTOI. The rightmost column shows the difference between using the "noisy CM" with respect to "clean CM" signals as the Dcor-MBSTOI reference.

Table 2. Mean absolute deviation (MAD) and standard deviation (SD) of ΔMBSTOI using different reference for Dcor-MBSTOI. The root mean square error (RMSE) between Dcor-MBSTOI and MBSTOI is also calculated.

Reference signal	MAD	SD	RMSE
clean binaural	0.037	0.034	0.048
clean single-channel	0.069	0.058	0.088
clean CM	0.080	0.059	0.099
noisy CM	0.095	0.113	0.126
enhanced noisy CM	0.083	0.104	0.105
noisy CM/clean CM	0.135	0.088	0.162

The leftmost five columns of Figure 2 show the distribution of ΔMBSTOI = Dcor-MBSTOI − MBSTOI on the simulated sentences when Dcor-MBSTOI uses different references. In each case we calculate the mean absolute deviation (MAD), the standard deviation (SD) and the root mean squared error (RMSE) whose values are given in Table 2. Best performance is achieved using the clean binaural reference (using a Dcor-MBSTOI version adapted to binaural references [21]), followed by the clean single-channel reference. Notably, the clean CM reference performs very similarly to the clean single-channel reference, even though it incorporates a small bulk time offset due to the different microphone positions. Dcor-MBSTOI is thus robust to small time delays. As previously observed in [21], Dcor-MBSTOI generally underestimates MBSTOI when using clean reference.

Using the noisy CM reference, Dcor-MBSTOI now often overestimates MBSTOI. Each individual sentence always estimates a higher Dcor-MBSTOI than when a clean CM reference is used. This is demonstrated in the rightmost column of Fig. 2 which plots $\Delta\text{MBSTOI} = \text{Dcor-MBSTOI}_{\text{noisy CM}} - \text{Dcor-MBSTOI}_{\text{clean CM}}$. The better performing sentences of Dcor-MBSTOI with a noisy CM reference have been investigated in terms of interferer position relative to target, signal to noise ratio, global levels, or reverberation time but no significant trend was found.

We hypothesize that, when using a noisy reference, the correlation between signal and reference is artificially increased leading to higher Dcor-MBSTOI. If this is the case, it should be possible to improve the estimation by applying noise reduction to the reference. Noise reduction in python using spectral gating [25] is therefore applied to noisy CM, resulting in a new reference called enhanced noisy CM. Results are shown in Fig. 2 second plot from the left. Results improved from noisy CM with lower MAD, SD and RMSE, but do not reach clean CM performance.

3.2. DNN performances

Figure 3 displays scatter plots of the correlation coefficients estimated by Dcor-MBSTOI compared to the target MBSTOI correlation value in a single, randomly-chosen sentence for which MBSTOI equals 0.289 and fwSegSNR at the CM equals $2.3\,\text{dB}$. Every dot corresponds to a single analysis time frame and each plot to a single third-octave band. Color denotes the reference signal used in Dcor-MBSTOI.

Black dots correspond to Dcor-MBSTOI with a clean binaural reference. Good correlation is observed in high frequency bands and more deviation is seen at lower frequencies. The average Pearson correlation coefficient across all frequencies is 0.896 with Dcor-MBSTOI value of 0.270.

Blue dots correspond to Dcor-MBSTOI with a clean single-channel reference. At low frequency, estimated coefficients match those using a binaural reference whereas at higher frequencies blue dots are seen to underestimate high correlation values. This leads to a lower Dcor-MBSTOI value of 0.202. Nevertheless, mean Pearson correlation remain high at 0.820. Similar performances are obtained with clean CM and are not displayed here.

Orange dots correspond to Dcor-MBSTOI with a noisy CM reference. While at very high and very low frequencies estimated values closely match those with a clean single-channel reference, Dcor-MBSTOI overestimates the correlation coefficients in most third-octave bands. This leads to a higher Dcor-MBSTOI of 0.336 with a mean Pearson correlation of 0.416. As noted in Sec. 2.1, an oracle VAD is provided to Dcor-MBSTOI so that every time frame contains target speech. Nevertheless, the noisy CM reference carries noise information which leads to higher estimated correlation coefficients.

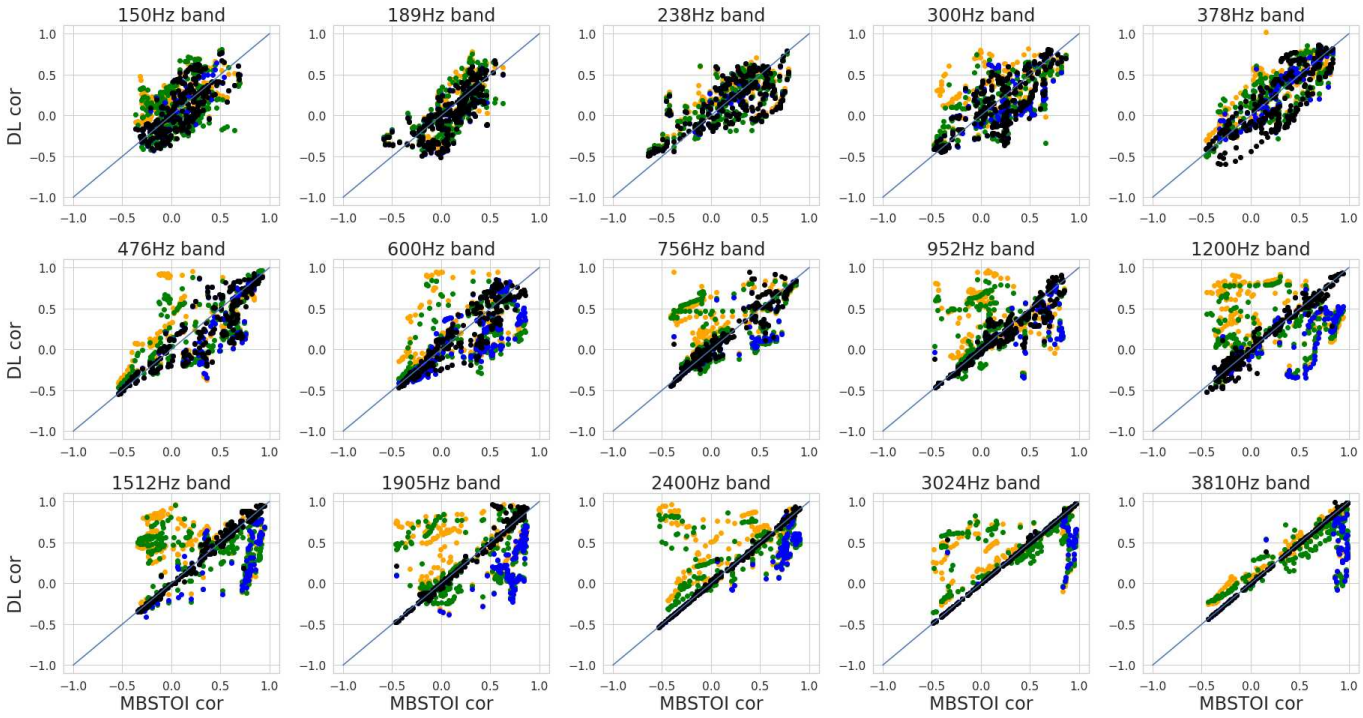

Fig. 3. Scatter plot of the estimated correlation coefficient from Dcor-MBSTOI compared to the target value from MBSTOI using various reference signal. Binaural reference in black, clean single-channel reference in blue, noisy close-talking reference in orange and enhanced close-talking reference in green. Displayed in each third-octave band for a single sentence, each dot being a single time frame estimation.

Lastly, green dots correspond to Dcor-MBSTOI with an enhanced noisy CM reference. It is observed that many frames still overestimate the correlation coefficients but less often than with the raw noisy CM. Dcor-MBSTOI of 0.267 is then closer to the target value and the mean Pearson correlation increases to 0.514. Notably in this example, using the enhanced noisy signal led to the Dcor-MBSTOI estimation closest to MBSTOI. However, given the poor average Pearson correlation coefficient and the high RMSE seen in Table 2, an alternative enhancement algorithm might result in more consistent estimation.

4. CONCLUSION

In this work, the machine learning hybrid version of MBSTOI called Dcor-MBSTOI, which uses only a clean single-channel reference, has been tested with a close-talking microphone reference which includes some noise and reverberation.

The results showed that when a noisy CM reference was used, Dcor-MBSTOI generally overestimated the value of the MBSTOI metric with an RMSE of 0.126. By applying a speech enhancement algorithm to the CM reference, the RMSE was reduced to 0.105. This still exceeds the RMSE of 0.088 obtained by using a clean single-channel reference, so it may be that an alternative enhancement algorithm would give better performance.

Nevertheless, this work has shown that it is feasible to have an accurate binaural intelligibility metric that uses a close-talking microphone signal as the reference in circumstances where a clean reference is unavailable.

5. ACKNOWLEDGMENTS

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/S035842/1].

6. REFERENCES

[1] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, 1947.

[2] ANSI, "Methods for the calculation of the speech intelligibility index," ANSI Standard S3.5-1997 (R2007), American National Standards Institute (ANSI), 1997.

[3] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[4] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI)," *Speech Commun.*, vol. 65, pp. 75–93, 2014.

[5] S. Jørgensen and T. Dau, "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," *J. Acoust. Soc. Am.*, vol. 130, no. 3, pp. 1475–1487, Sept. 2011.

[6] S. Jørgensen, S. D. Ewert, and T. Dau, "A multi-resolution envelope-power based model for speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 436–446, July 2013.

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

[8] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

[9] A. H. Andersen, J. M. de Haan, Z. H. Tan, and J. Jensen, "Predicting the intelligibility of noisy and nonlinearly processed binaural speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 1908–1920, Nov. 2016.

[10] A. H. Andersen, J. M. de Haan, Z. H. Tan, and J. Jensen, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Commun.*, vol. 102, pp. 1–13, Sept. 2018.

[11] N. I. Durlach, "Equalization and Cancellation Theory of Binaural Masking-Level Differences," *J. Acoust. Soc. Am.*, vol. 35, no. 8, pp. 1206–1218, Aug. 1963.

[12] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," *arXiv:1603.04467 [cs.DC]*, 2016.

[13] A. Paszke, S. Gross, F. Massa, A. Lerer, et al., "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc, 2019.

[14] M. B. Pedersen, A. H. Andersen, S. H. Jensen, and J. Jensen, "A Neural Network for Monaural Intrusive Speech Intelligibility Prediction," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2020.

[15] B. Cauchi, K. Siedenburg, J. F. Santos, T. H. Falk, et al., "Non-Intrusive Speech Quality Prediction Using Modulation Energies and LSTM-Network," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 7, pp. 1151–1163, July 2019.

[16] J. Rosbach, S. Rottges, C. F. Hauth, T. Brand, and B. T. Meyer, "Non-Intrusive Binaural Prediction of Speech Intelligibility Based on Phoneme Classification," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*. May 2021, pp. 396–400.

[17] D. S. Kim and A. Tarraf, "ANIQUE+: A new american national standard for non-intrusive estimation of narrowband speech quality," *Bell Labs Tech. J.*, vol. 12, pp. 221–236, 2007.

[18] R. E. Zezario, S.-W. Fu, C.-S. Fuh, Y. Tsao, and H.-M. Wang, "STOI-Net: A Deep Learning based Non-Intrusive Intelligibility Assessment Model," *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conf. (APSIPA ASC)*, 2020.

[19] A. H. Andersen, J. M. de Haan, Z. H. Tan, and J. Jensen, "Nonintrusive Speech Intelligibility Prediction Using Convolutional Neural Networks," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 26, no. 10, pp. 1925–1939, Oct. 2018.

[20] P. Guiraud, A. Moore, R. Vos, P. Naylor, and M. Brookes, "Machine learning for parameter estimation in the MBSTOI binaural intelligibility metric," in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, Sept. 2022.

[21] P. Guiraud, A. Moore, R. Vos, P. Naylor, and M. Brookes, "Using a single-channel reference with the MBSTOI binaural intelligibility metric," *Speech Communication*, submitted.

[22] G. Grimm, J. Luberadzka, and V. Hohmann, "A toolbox for rendering virtual acoustic environments in the context of audiology," *Acta Acustica united with Acustica*, vol. 105, no. 3, pp. 566–578, 2019.

[23] E. H. Rothauser, W. D. Chapman, N. Guttman, M. H. L. Hecker, et al., "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio and Electroacoust.*, vol. 17, no. 3, pp. 225–246, 1969.

[24] G. Hu and D. L. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 8, pp. 2067–2079, Nov. 2010.

[25] T. Sainburg, M. Thielk, and T. Q. Gentner, "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires," *PLoS computational biology*, vol. 16, no. 10, pp. e1008228, 2020.