Evaluating Multichannel Speech Enhancement Algorithms at the Phoneme Scale Across Genders

Nasser-Eddine Monir, Paul Magron, Romain Serizel

Université de Lorraine, CNRS, Inria, LORIA

F-54000 Nancy, France

{nasser-eddine.monir, paul.magron}@inria.fr, romain.serizel@loria.fr

Abstract—**Multichannel speech enhancement algorithms are essential for improving the intelligibility of speech signals in noisy environments. These algorithms are usually evaluated at the utterance level, but this approach overlooks the disparities in acoustic characteristics that are observed in different phoneme categories and between male and female speakers. In this paper, we investigate the impact of gender and phonetic content on speech enhancement algorithms. We motivate this approach by outlining phoneme- and gender-specific spectral features. Our experiments reveal that while utterance-level differences between genders are minimal, significant variations emerge at the phoneme level. Results show that the tested algorithms better reduce interference with fewer artifacts on female speech, particularly in plosives, fricatives, and vowels. Additionally, they demonstrate greater performance for female speech in terms of perceptual and speech recognition metrics.**

Index Terms—**Multichannel speech enhancement, beamforming, phoneme-level evaluation, gender-level evaluation.**

I. INTRODUCTION

Speech enhancement (SE) aims at retrieving a clean speech signal from a mixture contaminated with noise and/or reverberation. SE finds application in many downstream tasks such as hearing aids [1], speech recognition [2], and audio conferencing [3]. Traditional SE algorithms rely on signal processing techniques, leveraging mathematical assumptions about speech and noise [4], [5]. However, modern approaches are data-driven, and predominantly use deep neural networks (DNNs). A common strategy consists in combining DNNs for estimating spectral parameters (e.g., a correlation matrix or a time-frequency spectrum) with a traditional spatial filter, such as a minimum variance distortionless beamformer (MVDR) or a multichannel Wiener filter [6]–[11]. Alternatively, some algorithms directly estimate enhanced signals or filters via DNNs [12], [13].

SE is typically evaluated at the utterance level using metrics such as signal-to-distortion, artifacts, or interference ratios (SDR, SIR, SAR) [14], [15]. However, Miller et al. [16] emphasize the variability in phoneme noise tolerance, highlighting the importance of a nuanced understanding of how phonemes, particularly consonants and vowels, are affected by noise. Adachi et al. [17] reveal the ways in which different phonemes are impacted by noise for both native and non-native speakers, while Meyer et al. [18], [19] observe confusion among phonemes within both human perception and automatic speech recognition frameworks, indicating that consonants and

vowels are differently affected by the loss of information due to noise exposure. This suggests that evaluating SE using utterance-level metrics may overlook the detailed impacts of noise on different phonemes and the algorithms' processing of these sounds. This has motivated us to evaluate SE algorithms at the phoneme scale in a previous study [20].

However, beyond overall variability in phonemes, significant acoustic variations between male and female voices reveal that there are gender-specific differences in phonemes [21]. Gender perception in voices is mainly related to the fundamental frequency that is due to the length of the vocal tract, which affects formant patterns [22], [23]. Calliope [24] examined gender-based differences in vocalic formants, influencing the performance of data-driven speech processing [25]. These studies highlight the need to consider phonetic gender variations in SE technologies.

In this paper, we extend our previous work [20] by investigating the impact of gender and phonetic content on SE algorithms. We motivate our approach by analyzing spectral disparities in phonemes and gender. We evaluate three state-of-the-art multichannel SE algorithms [6], [11], [26] in a realistic simulated acoustic scenario using various metrics. The results reveal that while overall enhancement performance at the utterance level shows minimal gender differences, a deeper analysis at the phoneme level uncovers distinct trends, with female speech often exhibiting greater interference reduction and perceptual quality improvements, particularly as noise levels decrease.

II. METHODOLOGY

In this section, we describe our methodology for analyzing the impact of SE algorithms on male and female speech at a phoneme level.

A. Evaluation at the phoneme level

SE algorithms are typically evaluated at the utterance level, which provides an overall measure of speech clarity and comprehension. However, Miller et al. [16] suggest that consonants and vowels are impacted to different extents by the loss of information caused by the presence of noise.

To illustrate this, we display in Figure 1 the spectrogram of a clean speech signal and its mixture with a speech-shaped noise (SSN, see Section III-A). We observe that the low-frequency content of the clean speech is masked when mixed with the
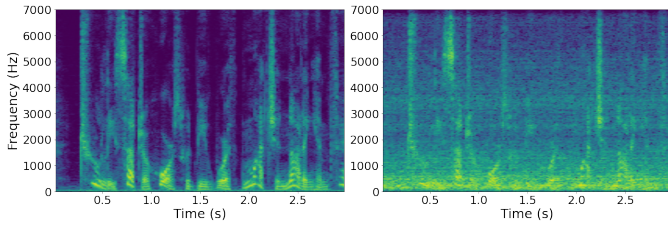
Fig. 1. Spectrograms of a male speech signal: clean (left) and mixed with an SSN at -5 dB SNR (right).

Fig. 2. Spectrum of the noise and clean plosives for male and female speakers at 0 dB SNR, computed from the dry signals.

noise, while only the sharp bursts at high frequencies remain slightly visible. Such high frequency components are indicative of phonemes that are characterized by their wide-band acoustic content, such as plosives or fricatives. This motivates evaluating speech across phonemes rather than solely at the utterance level, as initiated in our previous work [20].

B. Evaluation depending on the gender of the speaker

Beyond phonemic variability, speakers of different genders can introduce differences in speech acoustics. We illustrate this phenomenon by displaying a range of plosive sounds from male and female speakers in Figure 2, and we analyze how male and female plosives respond to SSN, highlighting spectral overlap and differential noise masking effects across genders. This invites a closer analysis of how speech enhancement algorithms might optimally address these gender-specific phonetic characteristics.

Male plosives show strong intensity below 100 Hz, while female plosives dominate at frequencies above 100 Hz. This indicates a shift in spectral emphasis, with male speech contributing more to the low-frequency range and female speech being more prominent in the mid to high frequencies. Both male and female plosives are partially masked by the SSN, but female plosives maintain stronger magnitudes above 100 Hz. These differences also appear in near-close vowels and fricatives at low noise levels, inviting further investigation into male-female speech characteristics to better understand their processing by SE algorithms.

III. EXPERIMENTAL SETUP

In this section we detail our experimental protocol. For a reproducibility purpose, both our code and the pretrained model weights are available online[1].

A. Acoustic scenarios and dataset

We build a dataset from LibriSpeech [27], using the `train-clean-100`, `dev-clean`, and `test-clean` subsets for training, validation, and testing, ensuring balanced male and female durations. For each subset, only 50% of the data is used as clean speech, yielding 50h for training and approximately 2.5h each for validation and testing. The other 50% are used to generate SSN, contributing 30% of total noise. SSN was chosen to provide controlled experimental
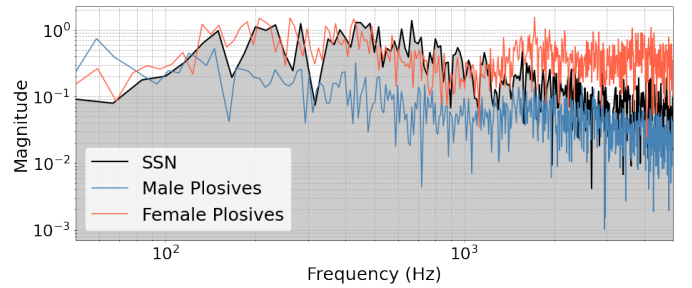
[1] https://github.com/Nasseredd/mcse-phg

conditions while preserving the spectral characteristics of both male and female voices. The remaining 70% comes from ecological sources in Disco-noise [11]. The validation set follows the same process, while the test set uses only SSN. SSN is generated by computing the discrete Fourier transform (DFT) of five male and five female speech signal, randomizing its phase, and applying the inverse DFT to ensure spectral consistency.

We simulate a hearing aid setup with four microphones, two on each ear. For training and validation, room impulse responses (RIRs) are generated using Pyroomacoustics [28] with RT_{60} between 0.15–0.4 s, and room dimensions of 3–8 m (length), 3–5 m (width), and 2.5–3 m (height). Speech and noise sources are randomly positioned, with signal-to-noise ratios (SNRs) from -10 dB to 10 dB, computed on the dry signals. For testing, we use measured RIRs [29] placing the speech source at 0 degrees (directly ahead of the listener) and the noise at 45 degrees to the right of the listener at SNR levels of -5, 0, or 5 dB.

B. Phoneme segmentation

We use the Montreal Forced Aligner (MFA) [30] to segment speech into phonemes, according to the international phonetic alphabet (IPA) chart in MFA. The English MFA dictionary v2.2.1 includes 13 phoneme classes (8 consonants and 5 vowels), and we adopt an extended classification from Monir et al. [20], adding a vowel class for near-close phonemes /[ɪ]/ and /[ʊ]/.

C. Speech enhancement algorithms

We perform SE with three algorithms. Tango [11] is a hybrid algorithm derived from the DANSE algorithm [31]. It employs a convolutional recurrent neural network for estimating time-frequency (TF) masks with binaural cues. FaSNet [26] is an end-to-end time-domain beamformer. It processes time-domain features with a dual-path recurrent network to estimate spectral masks for beamforming. MVDR [6] is a frequency-domain beamformer technique that uses a bidirectional long short-term memory network to predict TF masks that are used to estimate speech and noise covariance matrices. Note that our implementation (including training and evaluation scripts) rely on the Asteroid [32] and ESPnet [33] toolboxes. Full training details (e.g., loss functions, optimizers, etc.) are available in
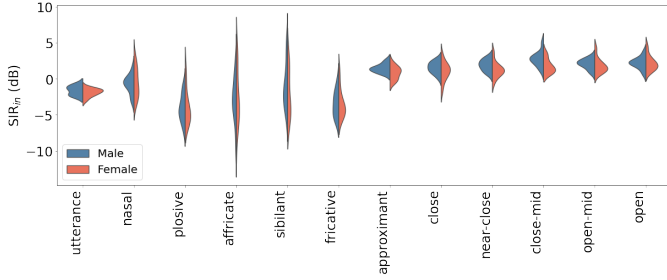
Fig. 3. Input SIR at the utterance level and across phoneme categories and genders at 0 dB SNR. Each violin plot represents the approximate distribution of the data.

Fig. 4. Output SIR at the utterance level and across phoneme categories and genders at 0 dB SNR.

TABLE I
MEAN SIR_{in} ACROSS GENDERS AND SNR LEVELS.

	-5 dB		0 dB		5 dB	
	M	F	M	F	M	F
Consonants	-7.71	-7.87	-1.16	-1.45	4.49	4.07
Vowels	-3.69	-3.79	2.10	1.90	7.46	7.19

Note: M = Male, F = Female.

TABLE II
MEAN SIR_{out} ACROSS GENDERS AND SNR LEVELS.

	-5 dB		0 dB		5 dB	
	M	F	M	F	M	F
Consonants	7.87	8.13	14.38	17.18	20.11	23.35
Vowels	8.50	11.45	14.08	16.97	18.88	21.52

our code. Finally, as this study does not aim to compare algorithms, we average their results to focus on gender and phoneme differences.

D. Evaluation metrics

We evaluate SE using the scale-invariant [15] SIR and SAR, expressed in dB [14]. We report the input and output SIR, respectively denoted SIR_{in} and SIR_{out}. SIR_{in} is the SIR at the ear level, thus it accounts for room acoustics (as opposed to the SNR which is adjusted on dry sources). SIR_{out} measures the residual interference after performing SE. Similarly, the output SAR (denoted SAR_{out}) assesses the overall amount of artifacts after SE. We do not report the input SAR as it is theoretically infinite when no processing has been applied.

Additionally, we report the perceptual evaluation of speech quality (PESQ) [34], short-time objective intelligibility (STOI) [35], and hearing aid speech perception index[2] (HASPI) [36] scores. To assess improvements in perceived quality and intelligibility, we measure PESQ and STOI before ($PESQ_{in}$, $STOI_{in}$) and after enhancement ($PESQ_{out}$, $STOI_{out}$), as well as their difference Δ_{PESQ} and Δ_{STOI}.

Finally, we feed the enhanced speech to five automatic speech recognition models[3] (Wav2vec, Wav2vec-lv60, Conformer-CTC, Conformer-Transducer and Whisper [37]–[39]), selected as proxies for evaluating speech intelligibility. We then compute the average word error rate (WER) over models, which is an overall measure of the impact of SE on speech recognition performance.

To assess statistical significance, we conduct Mann-Whitney U tests, a non-parametric method that is suited for comparing independent samples drawn from two unknown distributions

[2]We use the HASPI-v2 version with the normal-hearing auditory model.
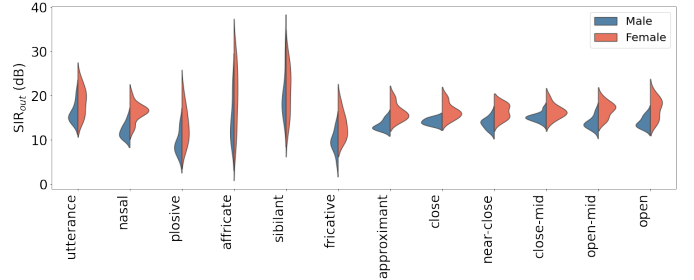[3]Except for Whisper, all models are pre-trained on the gender-balanced LibriSpeech dataset.

of any of the afore-mentioned metrics. The statistical tests primarily compare male and female speech across different levels: at the utterance level, within consonants and vowels separately, and for each phoneme category. The method computes a p-value for a given pair of input distributions, and we consider the difference between categories to be significant when $p < 0.05$. While all statistical analyses were conducted, p-values are not systematically reported to ensure readability and focus on the most relevant findings, which are discussed in the text.

IV. RESULTS AND DISCUSSION

A. Analysis on input signals

First, we analyze the input signals before performing SE. The results in terms of SIR_{in} over phoneme categories and genders at 0 dB SNR[4] are displayed in Figure 3. We observe that interfering noise at the utterance level is similar for males and females, and no significant difference between gender can be observed within each phoneme category.

Table I presents the mean SIR_{in} values across SNR levels for male and female speakers. The results show statistically significant differences between consonants and vowels for both male ($p < 0.001$) and female speakers ($p < 0.002$) across SNR levels. However, no significant differences between genders were found within each phoneme category at any SNR level. This gender similarity extends across phoneme categories, where no significant difference is observed, except for under-represented laterals and taps. This suggests that speech sounds has a greater influence on SIR_{in} than genders.

B. Results after speech enhancement

1) Impact on interference: First, we analyze the impact of SE in terms of noise reduction, as measured by the output

[4]Similar trends can be observed at -5 and 5 dB, but figures are omitted due to space constraints. This also applies to Figures 4 and 5.
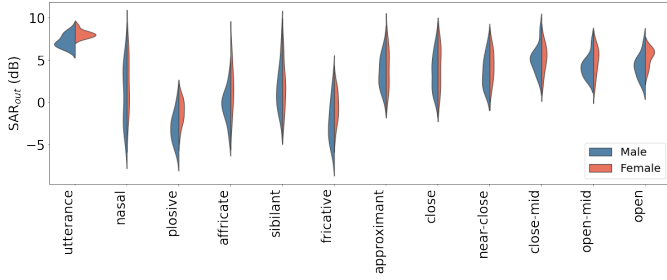
Fig. 5. Output SAR at the utterance level and across phoneme categories and genders at 0 dB SNR.

TABLE III
MEAN SAR_out ACROSS GENDERS AND SNR LEVELS.

	-5 dB		0 dB		5 dB	
	M	F	M	F	M	F
Consonants	-1.33	-1.11	0.89	1.90	2.03	3.99
Vowels	1.21	1.69	4.12	5.01	5.56	7.18

SIR. We display in Figure 4 the output SIR at 0 dB across utterance, phoneme categories and genders, while Table II summarizes the mean SIR_out for consonants and vowels by gender across SNR levels. We observe in Figure 4 that SE algorithms process male and female speech similarly at the utterance level ($p = 0.07$). However, results in Table II show significant performance differences at 5 dB and 0 dB for consonants ($p = 0.025$ and 0.017) and vowels ($p = 0.002$ and 0.003). This suggests that the algorithms process interference differently depending on gender-specific speech characteristics. Nonetheless, at -5 dB, the difference for consonants disappears ($p = 0.064$), likely because the interfering noise masks differences in spectral and temporal cues. Despite vowels also being susceptible to noise masking, we observe gender differences ($p = 0.001$) at the output.

While the overall enhancement performance appears similar across genders when averaged over entire utterances, a closer look at individual phoneme categories in Figure 4 reveals significant differences in nasals, plosives, fricatives, approximants, and most vowel types. This suggests that male and female speakers exhibit distinct acoustic properties in these phonemes, which persist even after interference reduction. On the other hand, affricates and sibilants show no significant gender-based difference ($p = 0.09$ and 0.79, respectively), likely because these phonemes naturally contain turbulent, high-frequency energy, making them harder to distinguish from background noise.

2) Impact on artifacts: Here we analyze the impact of SE in terms of artifacts in the estimated signals. Figure 5 displays the output SAR across phoneme categories and genders at 0 dB. The results suggest that at the utterance level, SE algorithms preserve female speech quality slightly better than male speech ($p = 0.02$). This trend remains at 5 dB, but the difference between genders becomes more pronounced at -5 dB.

Table III presents the mean SAR_out for consonants and vowels by gender across SNR levels. The results show no

TABLE IV
MEAN SPEECH RECOGNITION AND PERCEPTUAL METRICS ACROSS
GENDERS AND SNR LEVELS.

Metrics	-5 dB		0 dB		5 dB	
	M	F	M	F	M	F
WER	70.20	61.37	31.62	27.20	17.90	16.52
STOI_in	0.46	0.47	0.57	0.58	0.69	0.69
STOI_out	0.61	0.62	0.75	0.75	0.82	0.83
Δ_{STOI}	0.15	0.14	0.17	0.17	0.12	0.13
PESQ_in	1.07	1.04	1.10	1.06	1.21	1.12
PESQ_out	1.19	1.19	1.41	1.46	1.67	1.76
Δ_{PESQ}	0.12	0.14	0.31	0.39	0.46	0.64
HASPI	0.38	0.47	0.84	0.81	0.96	0.92

significant gender differences for consonants ($p = 0.24$) and vowels ($p = 0.18$) at 0 dB, or other tested SNR levels. This indicates that, on average, the algorithm does not introduce artifacts in a gender-biased way when considering broad phoneme categories.

At the phoneme level, however, more nuanced differences emerge. Plosives show a significant difference ($p = 0.02$), with female speech having a higher output SAR, which means that male plosives are more affected by artifacts. This suggests that the algorithms show limitations in preserving plosive sounds in male speech, possibly due to their stronger bursts and lower fundamental frequencies. In contrast, nasals, affricates, fricatives, approximants, and vowels do not show statistically significant differences, suggesting that the algorithm affects these sounds similarly across genders.

3) Impact on speech recognition: To evaluate the impact of noise on speech recognition performance across genders, we examine the WER at different SNR levels, which is displayed in the first line of Table IV. As the SNR increases, the WER decreases for both genders, indicating improved speech recognition performance at lower noise levels. At -5 dB and 0 dB, female speech consistently exhibits a lower WER compared to male speech, suggesting that speech recognition models handle female voices slightly better under noisy conditions ($p = 0.012$ and 0.048, respectively). However, at 5 dB, where speech is more dominant over noise, the WER difference between male and female speech is not significant, confirming that gender-related effects diminish as noise interference decreases.

4) Impact on perceptual metrics: Finally, we investigate gender-based variations in terms of perceptual metrics. The mean PESQ, STOI, and HASPI values across SNRs are presented in Table IV. Both males and females show similar patterns in input and output STOI, with a steady increase as SNR improves, and the gap between input and output remains fairly consistent across genders. However, PESQ scores exhibit increasing disparity: female input scores are lower than males' across all SNR levels, but their output scores tend to surpass males' as SNR rises. STOI improvements, in contrast, remain relatively stable for both genders, whereas PESQ shows a more significant difference, with female speech exhibiting greater improvement as SNR increases. Additionally, HASPI scores

indicate that female speech tends to retain slightly higher intelligibility, particularly at lower SNR levels.

Overall, the results across metrics indicate that female speech consistently shows higher perceptual improvements (higher PESQ and HASPI scores) and lower WER at most SNR levels, suggesting a stronger benefit from SE for females. This aligns with trends observed in SIR and SAR, where female speech exhibits greater interference reduction and fewer artifacts, suggesting that the acoustic characteristics of female speech are more effectively enhanced by the SE algorithms.

V. Conclusion

This study highlights the need for a nuanced evaluation of multichannel SE algorithms, considering the distinct acoustic characteristics of phoneme categories and gender differences. No utterance-level SIR differences were found between genders, but most phonemes, except affricates and sibilants, had better interference reduction in female speech, which reveals variations that are overlooked in utterance-level analysis. For artifacts, a difference was found at the utterance level and for plosives, but not for other phoneme categories.

These findings can be exploited in future work, e.g., by integrating filtering algorithms that account for phoneme-specific spectral properties into SE algorithms, or optimizing deep SE algorithms with frequency-weighted / phoneme-informed losses that prioritize spectral regions that are perceptually important. Besides, one can leverage phoneme coarticulation through the use of consonant-vowel and vowel-consonant sequences to design SE algorithms that better capture transitional dynamics and ensure a more natural speech flow.

References

[1] A. J. S. Esra and Y. Sukhi, "Optimized binaural enhancement for digital hearing aids," *Comp. Speech Lang.*, vol. 84, pp. 101554, 2024.

[2] K. Iwamoto, T. Ochiai, M. Delcroix, R. Ikeshita, H. Sato, S. Araki, and S. Katagiri, "Impact of speech enhancement errors on ASR," 2022.

[3] W. Rao, Y. Fu, Y. Hu, X. Xu, Y. Jv, J. Han, Z. Jiang, L. Xie, Y. Wang, S. Watanabe, Z.-H. Tan, H. Bu, T. Yu, and S. Shang, "Conferencingspeech challenge: Far-field multi-channel speech enhancement," in *Proc. ASRU*, 2021, pp. 679–686.

[4] J. Benesty, M. M. Sondhi, and Y. A. Huang, Eds., *Springer Handbook of Speech Processing*, Springer, 2008.

[5] R. C. Hendriks and T. Gerkmann, "Noise correlation matrix estimation for speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 1, pp. 223–233, 2012.

[6] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network spectral mask estimation for beamforming," in *Proc. IEEE ICASSP*, 2016, pp. 196–200.

[7] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 9, pp. 1652–1664, September 2016.

[8] M. Coto-Jimenez, J. Goddard-Close, L. Di Persia, and H. Leonardo Rufiner, "Hybrid speech enhancement with Wiener filters and LSTM autoencoders," in *Proc. IWOBI*, July 2018, pp. 1–8.

[9] Y. Liu, A. Ganguly, K. Kamath, and T. Kristjansson, "Neural MVDR beamforming with time-frequency masks," in *Proc. ICASSP*, April 2018, pp. 6717–6721.

[10] G. Carbajal, R. Serizel, E. Vincent, and E. Humbert, "Joint NN-supported multichannel echo, reverberation, and noise reduction," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2158–2173, July 2020.

[11] N. Furnon, R. Serizel, S. Essid, and I. Illina, "DNN-based mask estimation for distributed speech enhancement," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 2310–2323, June 2021.

[12] Z.-Q. Wang and D. Wang, "All-neural multi-channel speech enhancement," in *Proc. Interspeech*, September 2018, pp. 3234–3238.

[13] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, "Fasnet: Low-latency adaptive beamforming," in *Proc. IEEE ASRU*, December 2019, pp. 260–267.

[14] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, pp. 1462–1469, 2006.

[15] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR: Half-baked or well done?," in *Proc. IEEE ICASSP*, May 2019, pp. 626–630.

[16] G. A. Miller and P. A. Nicely, "Perceptual confusions among english consonants," *J. Acoust. Soc. Am.*, vol. 27, pp. 338–352, 1955.

[17] T. Adachi, R. Akahane-Yamada, and K. Ueda, "Intelligibility of english phonemes in noise for native and non-native speakers," *Acoust. Sci. Technol.*, vol. 27, no. 5, pp. 285–289, 2006.

[18] B. T. Meyer, T. Jürgens, T. Wesker, T. Brand, and B. Kollmeier, "Human phoneme recognition and speech variability," *J. Acoust. Soc. Am.*, vol. 128, no. 5, pp. 3126–3141, 2010.

[19] J. Zaar and T. Dau, "Predicting consonant recognition and confusions in normal hearing," *J. Acoust. Soc. Am.*, vol. 141, no. 2, pp. 1051–1064, 2017.

[20] N.-E. Monir, P. Magron, and R. Serizel, "A phoneme-scale assessment of multichannel speech enhancement algorithms," *Trends in Hearing*, vol. 28, pp. 23312165241292205, 2024.

[21] C. Henton, *Phonetic Sex-Specific Differences Across Languages*, Ph.D. thesis, Univ. Oxford, 1986.

[22] R. O. Coleman, "Voice quality and gender perception," *J. Speech Hear. Res.*, vol. 19, no. 1, pp. 168–180, 1976.

[23] E. Pépiot, "Gender identification by voice in english and french," *Sci. Works*, vol. 49, pp. 418–430, 2011.

[24] L. Calliope and G. Fant, *Speech and Its Automatic Processing*, Masson, 1989.

[25] M. Adda-Decker and L. Lamel, "Speech recognizers and female speakers," in *Proc. Interspeech*, September 2005, pp. 2205–2208.

[26] Yi Luo, Zhuo Chen, Nima Mesgarani, and Takuya Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. IEEE ICASSP*, 2020, pp. 6394–6398.

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: ASR corpus from public audio books," in *Proc. IEEE ICASSP*, April 2015, pp. 5206–5210.

[28] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: Python package for audio room simulation," in *Proc. IEEE ICASSP*, 2018.

[29] L. Delebecque and R. Serizel, "Binaurec: Dataset for binaural speech enhancement with rirs," in *Proc. EUSIPCO*, September 2023, pp. 126–130.

[30] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Text-speech alignment with kaldi," in *Proc. Interspeech*, August 2017.

[31] A. Bertrand and M. Moonen, "Distributed adaptive signal estimation in sensor networks," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5277–5291, 2010.

[32] M. Pariente, S. Cornell, J. Pons, S. B. R., A. Deleforge, and E. Vincent, "Asteroid: PyTorch-based audio source separation toolkit," in *Proc. Interspeech*, 2020, pp. 2637–2641.

[33] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Yalta, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "Espnet: End-to-end speech processing toolkit," 2018.

[34] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)," in *Proc. IEEE ICASSP*, May 2001, pp. 749–752.

[35] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "Intelligibility prediction of noisy speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[36] C. Spille, B. Kollmeier, and B. T. Meyer, "Human vs. automatic speech recognition in acoustic scenes," *Comput. Speech Lang.*, vol. 52, pp. 123–140, 2018.

[37] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: Self-supervised speech representation learning," 2020.

[38] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," 2020.

[39] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.