Noise Be Gone: Does Speech Enhancement Distort Linguistic Nuances?

Iñigo Parra
The University of Alabama[*]
Tuscaloosa, AL
iparra@berkeley.edu

Abstract

This study evaluates the impact of speech enhancement (SE) techniques on linguistic research, focusing on their ability to maintain essential acoustic characteristics in enhanced audio without introducing significant artifacts. Through a sociophonetic analysis of Peninsular and Peruvian Spanish speakers, using both original and enhanced recordings, we demonstrate that SE effectively preserves critical speech nuances such as voicing and vowel quality. This supports the use of SE in improving the quality of speech samples. This study marks an initial effort to assess SE's reliability in language studies and proposes a methodology for enhancing low-quality audio corpora of under-resourced languages.

1 Introduction

Speech is a fundamental mode of human communication, consisting primarily of two components: speech production and speech perception (Deller Jr et al., 1993). Speech production enables individuals to articulate ideas through sound using linguistic structures. Conversely, speech perception involves the decoding of sound waves generated during speech production. These processes can be influenced by external factors such as ambient or background noise, potentially disrupting the communication sequence (Michelsanti et al., 2021).

Humans have evolved mechanisms to filter out these disturbances (Bronkhorst, 2000; Cherry, 1953; Shinn-Cunningham and Best, 2008). However, audio recordings capture both desired and undesired signals indiscriminately. This poses significant challenges for sociophonetic research, which often relies on pre-recorded audio data. Speech enhancement (SE) techniques clean and filter these recordings from external noise, thus enhancing the perceptual quality of the speech (Michelsanti et al., 2021). This presents SE as a useful tool for refining audio corpora.

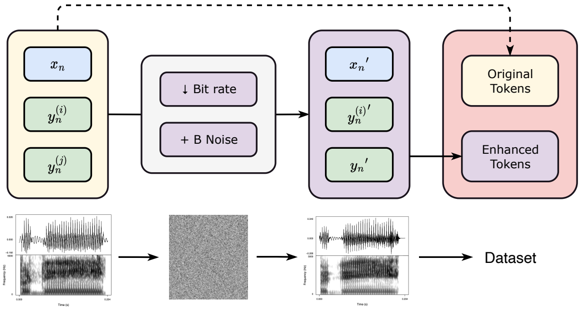

Figure 1: Diagram of the token processing. x_n represent intervocalic voiceless fricative tokens (e.g., /asa/); $y_n^{(i)}$ and $y_n^{(j)}$ represent vocalic /e/ (e.g., /bre/) and /i/ (e.g., /li/) tokens respectively. The original tokens are copied. One version is stored in the final dataset, while the other is processed as explained above to provide the enhanced copies $y_n^{(i)'}$ and $y_n^{(j)'}$. The final dataset includes all tokens, original and enhanced.

The reliability of speech enhancement models in improving the quality of linguistic speech corpora remains an open question. Sociophonetic studies, which explore speech variations among different social groups, provide a robust framework for testing SE models to ensure they maintain essential acoustic characteristics (e.g., vowel quality or voicing). Moreover, these methodologies often focus on subtle speech variations, making them ideal for assessing the ability of SE models to retain these nuances post-enhancement.

This study seeks to evaluate the effects of SE on linguistic corpora by conducting paired sociophonetic studies. We present a case study that examines the voicing and duration of intervocalic voiceless fricatives, as well as vocalic quality variations between Peninsular and Peruvian Spanish speakers. Our findings indicate that the studies using original and enhanced recordings yield comparable results. To our knowledge, this is the first work (1) *address-*

[*] Current affiliation: UC Berkeley, Department of Linguistics, Berkeley, CA.

Proceedings of the 3rd Workshop on NLP Applications to Field Linguistics (Field Matters 2024), pages 52–60
August 16, 2024 ©2024 Association for Computational Linguistics

ing such questions from a linguistic viewpoint and (2) *proposing a novel methodological approach for handling low-quality audio data in linguistic studies*.

2 Previous Work

Although there is ongoing research into the bias introduced by enhanced recordings (Isik et al., 2020), the linguistic community continues to debate the risk of distorting results through potential artifact introduction during enhancement. Previous technologies like WaveNet (Van Den Oord et al., 2016) have shown the ability to replicate speech with particular linguistic and acoustic subtleties (Chen et al., 2018); however, further exploration in this area is limited.

Most of the sociophonetics studies dealing with technology have focused on audio quality. Calder et al. (2022) studied the usability of Zoom as a tool for recording speech data. They found that F1 and F2 values showed significant differences compared to speech recorded with specialized equipment. Rathcke et al. (2017) look at how different normalization methods affect recordings with different degrees of quality, showing that normalization procedures may be relevant to address technical factors in low-quality recordings. Background noise has also been a central topic for perceptual studies, which coincide in that it should be eliminated as much as possible (Thomas, 2002, 2013). To this issue, filtering (Gradoville et al., 2022), especially low-pass, may be useful; however, there is a risk of deleting relevant nuances of speech production. Overall, while some works have used methodologies borrowed from linguistics (Michelsanti et al., 2021), SE has not had much attention in the field.

Avoiding hard filtering is crucial to analyzing high-frequencies (HF) content-heavy speech. Studies have gradually recognized the importance of retaining HF content in speech signals (Best et al., 2005; Yu et al., 2014), particularly when analyzing fricatives (Kharlamov et al., 2023; Jacewicz et al., 2023). Fricatives, which are rich in high-frequency energy, have shown to play a significant role in distinguishing phonetic and phonological features (Jongman et al., 2000). In the context of Peruvian Spanish and Peninsular Spanish, analyzing the voicing of fricatives before and after enhancement is particularly insightful. Chládková et al. (2011) offered a detailed description of Peruvian and Peninsular Spanish and Morrison et al. (2007) compared vocalic sounds in both variations, showing that Peruvian speakers reproduced higher fundamental frequency values.

3 Methodology

3.1 Data

We use two sources of data. The Peruvian Spanish tokens are extracted from a crowd-sourced Latin American Spanish dataset (Guevara-Rukoz et al., 2020), which included recordings of speakers from Lima. The Peninsular Spanish tokens were extracted from an open-source speech corpus from Kaggle (Fonseca, 2023) containing recordings of speakers from Madrid. Both datasets included short recordings (5-10s) of middle-class male and female speakers. We selected eight speakers, divided into two equal groups per variation. We did not consider the education level for this study[1].

From the recording pool of each speaker, we filtered those containing vowels /e/ and /i/, as well as fricative voiceless /s/ in intervocalic contexts. We then filtered out the tokens containing pre-vocalic nasals since they potentially reduce the acoustic power of the sound due to the introduction of antiresonances in the spectrum (Vampola et al., 2020). Sounds /i/ and /e/ have already been studied due to their alternations in Spanish (Brame and Bordelois, 1973). Because they share features (both are front vowels) and diverge in tongue height, any applied enhancement should be able to preserve the unique characteristics of each sound.

The total original tokens for both Spanish variants are described in Table 1. The number of enhanced tokens is the same as the ones described below; therefore, the study analyzed $N = 208$ tokens (for more details, see Appendix B).

Type	Total (n)	/s/ v_v			/i/			/e/		
		Total	M	F	Total	M	F	Total	M	F
Peruvian	68	14	7	7	25	15	10	29	14	15
Peninsular	70	14	7	7	29	14	15	27	13	14

Table 1: Descriptive statistics of the original tokens. With the enhanced tokens, the amount is doubled.

3.2 Token Enhancement

After duplicating the original tokens, we designed a perturbation function that applies additive white

[1] https://github.com/IParraMartin/A2A-ACL24

Model	Coefficient	Estimate	Std. Error	t value	p value	Estimate	Std. Error	t value	p value
Voicing	(Intercept)	4.994	2.057	2.427	**.022**	4.172	1.938	2.152	**.041**
	countrySpain	1.564	2.376	.658	.5163	2.488	2.238	1.112	.276
	genderM	.257	2.376	.108	.914	.015	2.238	.007	.994
Duration	(Intercept)	-4.584	.006	-694.89	**<.01**	-4.584	.006	-714.586	**<.01**
	countrySpain	-.022	.007	-3.00	**<.01**	-.020	.007	-2.796	**<.01**
	genderM	-.005	.007	-.75	0.46	-.006	.007	-.868	.393

Table 2: Results from the generalized linear models (GLM) for voicing and duration using original (left) and speech-enhanced tokens (right).

Gaussian noise (AWGN) to the copies (see Appendix A). We then blended the noise in the background and decreased the bit rate of the sound. To restore the sound, we use Voicefixer (Liu et al., 2021), a neural vocoder-based audio-to-audio model.

3.3 Voicing Experiments: Intervocalic Fricative /s/

In the intervocalic /s/ voicing experiments, we looked for segment voicing variations among the original and enhanced tokens. We fitted multiple statistical models to analyze both versions: ANOVAs, generalized linear models (GLM), robust linear models (RLM), and robust linear mixed-effects models (RLMEM). After analyzing conditions separately, we fit two additional models using condition as a predictor (IV) of voicing and duration (DV) (Appendix C).

The selection of diverse models was motivated by the practices in linguistics literature and the specific characteristics of our data. Although ANOVAs are widely used in linguistic research, we encountered issues related to the robustness of their results with our data specifications. To address these concerns, we tested robust models (RLM and RLMEM) that offer more flexibility in handling data assumptions. Additionally, GLMs were used, providing reliability and reinforcing our findings compared to other methods. This comprehensive approach ensures a robust examination of the variables under study.

3.4 Vocalic Quality Experiments: /i/ vs /e/

To account for the changes in the vocalic quality of /i/ and /e/ tokens, we conducted principal component analyses (PCA) and Procrustes analyses before and after enhancement. We examine the measurements of the first (F1) and second (F2) formant values at 16 evenly spaced intervals throughout the duration of vocalic tokens. These measurements form n-dimensional arrays that we call F-vectors. We compare these F-vectors using PCA and Procrustes tests to assess the statistical significance of the quality changes observed between the original and processed audio tokens.

4 Results

4.1 Voicing of Fricative /s/

Paired Experiments

In Table 2, we provide the results for the models with the best fits during paired experimentation.

In terms of voicing, there was a significant positive effect in the model's intercept using the original tokens ($\beta = 4.994, p = .02$) and the one using enhanced versions ($\beta = 4.172, p = .041$). This indicates that the baseline level of the response variable is significantly different from zero when all other predictors are held constant. However, based on the pseudo-R^2 metrics (ρ), these results show weak effect sizes ($\rho = .01$ and $\rho = .04$ respectively). For voicing, the effects attributed to being Peninsular or being male were not statistically significant. The effect of gender and location was negligible across both models, with high p-values, suggesting that they do not influence voicing in intervocalic fricatives when comparing Peninsular and Peruvian Spanish.

When examining duration, there was a significant negative effect in the intercepts of both models. We also found that the intercepts were identical for the model using the original tokens and the one using their processed versions ($\beta = -4.584, p < .01$). Interestingly, being Peninsular was a significant predictor of duration ($p < .01$), and it was associated with a decrease in the frication ($\beta = -.022$). This result was also reflected in the model using SE tokens ($\beta = -.020, p < .01$). Unlike voicing, the results for duration also showed high effect sizes, $\rho = .28$ and $\rho = .25$ for SE tokens, which are considered to show excellent model fits (McFadden, 1972).

Analyzing the results for voicing and duration in intervocalic /s/ when comparing paired models, we found no evidence suggesting that the enhanced tokens significantly modified or contaminated the original audio samples.

Interaction Experiments

In Table 3, we provide the results of generalized linear models using condition (original or enhanced) as an independent variable.

Model	Coefficient	Estimate	Std. Error	t value	p value
Voicing	(Intercept)	4.977	1.797	**2.769**	**<.01**
	countrySpain	.757	2.273	.333	.74
	genderM	-1.132	2.273	-.498	.62
	conditionOG	.48	1.607	.299	.766
	countrySpain:genderM	2.537	3.215	.789	.433
Duration	(Intercept)	-4.595	.004	**-993.02**	**<.01**
	countrySpain	.001	.005	.244	.808
	genderM	.017	.005	**2.928**	**<.01**
	conditionOG	-.001	.004	-.259	.796
	countrySpain:genderM	-.046	.008	**-5.608**	**<.01**

Table 3: Results from the generalized linear models (GLM) for voicing and duration using condition as independent variable. Underlined results show no significant impact of the condition on voicing and duration of intervocalic fricative (s). OG stands for original.

The generalized linear model for voicing demonstrated a significant intercept ($\beta = 4.977, p < .01$), indicating that the baseline level of voicing is significantly different from zero when all other predictors are controlled. However, the effects of being from Spain ($\beta = .757, p = .740$), being male ($\beta = -1.132, p = .620$), and the condition of original tokens ($\beta = .48, p = .766$) were not statistically significant. The interaction between being from Spain and being male ($\beta = 2.537, p = .433$) also showed no significant impact on voicing. The model accounted for a small portion of the variance in voicing, with a pseudo-R^2 value of $\rho = .043$.

In contrast, the model for duration revealed more significant effects. The intercept was significant and negative ($\beta = -4.595, p < .01$), suggesting a strong baseline effect on duration. The effect of gender was significant, with males exhibiting longer duration ($\beta = .017, p < .01$). The condition of the original tokens did not significantly influence duration ($\beta = -.001, p = .796$). Notably, the interaction term for being a male from Spain indicated a substantial negative impact on duration ($\beta = -.046, p < .01$). The models for duration displayed excellent fit, with pseudo-R^2 values of $\rho = .546$ for both original and enhanced tokens, indicating robust explanatory power.

These results highlight the differing effects of demographic factors and experimental conditions on voicing and duration. While factors such as gender significantly influenced duration, they had minimal effects on voicing. As for condition, the experimental manipulation of audio enhancement did not significantly alter the outcomes, indicating robustness in preserving phonetic characteristics. All these results seem to reflect that the nuanced properties of audio are preserved after SE.

4.2 Vocalic Quality

In this section, we compare the results of the vocalic quality of the Peninsular and Peruvian variants before and after audio enhancement.

/e/ Sound

This section presents the results of the Procrustes analysis performed to compare the principal component analyses (PCA) of the original and enhanced /e/ vocalic sounds across the different demographic groups (Figure 2).

For Peninsular Spanish speakers, the Procrustes analysis revealed distinctive outcomes based on gender. Female speakers demonstrated a Procrustes Sum of Squares (M_{12}) of .121, indicating a moderate degree of shape difference between the original and enhanced datasets. Despite this, a high correlation in a symmetric Procrustes rotation (.937) suggested that the overall structural integrity of the vowel space was largely maintained ($p < .01$). In contrast, male speakers displayed lower Procrustes ($M_{12} = .04$), showing closer alignment between the original and enhanced forms. The correlation coefficient was significantly high (.979), indicating an effective preservation of acoustic characteristics after enhancement. These results were also statistically significant ($p < .01$).

The results for Peruvian Spanish speakers further emphasized the effectiveness of speech enhancement techniques. Female speakers showed an even smaller deviation between the original and enhanced datasets ($M_{12} = .03$). The correlation coefficient (.984) reflected the preservation of vowel characteristics post-enhancement, with a statistically significant value ($p < .01$). Male speakers exhibited $M_{12} = .031$, with a correlation of .984. These results suggest that the speech enhancement process robustly maintained the integrity of the vocalic sounds ($p < .01$).

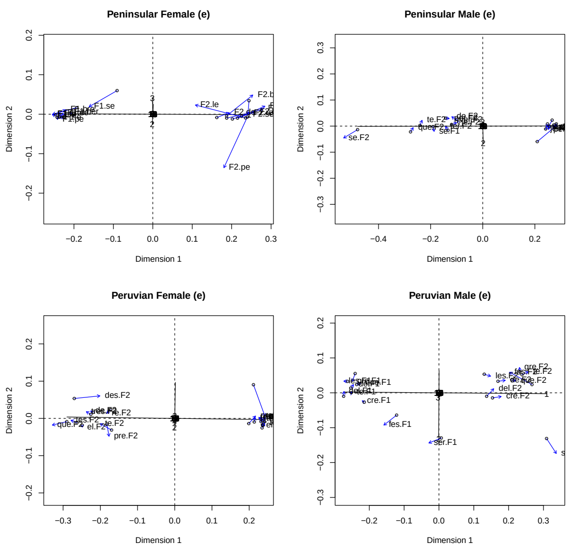

Figure 2: Procrustes plots for /e/ sounds for all groups and genders. Longer arrows display larger displacements between original and enhanced tokens. As seen in the projections, Peruvian vowels tend to be higher.

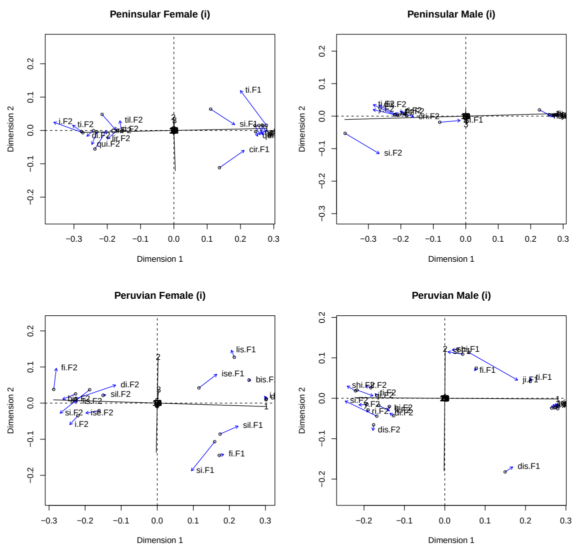

Figure 3: Procrustes plots for /i/ sounds for all groups and genders. Longer arrows display larger displacements between original and enhanced tokens. As seen in the projections, Peruvian vowels tend to be higher.

The analysis confirmed that the SE techniques employed in this study effectively preserve essential acoustic characteristics of /e/ vowel sounds across different Spanish-speaking populations. The high correlations and significant p-values across demographic groups reinforce the reliability of these enhancement methods in linguistic data. The combined results from the Procrustes analysis and the visual representations underscore the effectiveness of SE in retaining the critical acoustic properties and vocalic quality.

/i/ Sound

This section details the outcomes of the Procrustes analysis comparing the principal component analyses of original and enhanced /i/ vocalic sounds (Figure 3).

For Peninsular Spanish speakers, the Procrustes analysis varied between genders. Female speakers showed a $M_{12} = .082$, suggesting a noticeable deviation between the original and enhanced datasets, albeit less significant than for the /e/ sounds. However, the correlation in a symmetric Procrustes rotation was strong (.957), indicating that the speech enhancement preserved much of the vowel space's structural integrity. The significance of these observations was confirmed with a value $p < .01$. Male speakers exhibited $M_{12} = .051$, lower than the previous group, indicating a more faithful preservation of the original vocal characteristics. The high correlation coefficient (.973) further supported the effectiveness of the SE, with results being statistically significant ($p < .01$).

For Peruvian Spanish speakers, the results were similarly instructive. Female speakers recorded $M_{12} = .086$, which was slightly higher than that observed for Peninsular females, indicating a modest shape difference between the original and enhanced versions. The correlation coefficient was .955, reflecting robust maintenance of vowel characteristics despite the enhancements ($p < .01$). Male speakers, on the other hand, showed an even better alignment ($M_{12} = .047$) and a good correlation (.976), highlighting the small impact of the enhancement process in corrupting the acoustic properties of the sound ($p < .01$).

While some deviations were observed, particularly among female speakers, the overall high correlation values indicate that the enhancements largely preserved the essential acoustic characteristics of the /i/ sound. The results and significance were similar to the results for /e/.

5 Conclusion

In this study, we have analyzed the impact of speech enhancement (SE) on the audio properties of fricative and vocalic sounds in Spanish. We use a sociophonetic case study to test whether results are consistent across original and audio-enhanced tokens. We analyzed the results for voicing and duration in intervocalic /s/, comparing paired models fitted on original and enhanced data. We also inspected the impact of condition as an independent variable on voicing and duration.

In the sociophonetic dimension, our analyses show that while demographic factors such as gender and geographic origin influence certain phonetic features like frication duration, they have minimal impact on others such as voicing. Regarding condition, the experimental manipulation of audio enhancement did not significantly alter the outcomes, indicating robustness in preserving phonetic characteristics. We found no evidence suggesting that the enhanced tokens significantly modified or contaminated the statistical results.

Experiments in vocalic quality showed a similar trend. The features captured by the PCA coincide with previous literature on the comparison between Peruvian and Peninsular vowels. We show that SE tokens preserve essential acoustic characteristics of vocalic sounds across different Spanish-speaking populations. The high correlations and significant outputs across all demographic groups reinforce the reliability of the results.

These findings hold the potential to yield advantageous results for languages with limited resources, which usually have lower-quality speech corpora. By demonstrating the robust preservation of acoustic properties and sociophonetic markers, this study supports the effectiveness of speech enhancement for data in which linguistic nuances are critical.

6 Limitations and Future Work

While informative and representative, this study was limited to a relatively small sample size. Future studies may benefit from examining tokens with different amounts of background noise or more realistic artifacts (e.g., inserting noises at intervals, overlaying background conversations, or low-quality recording equipment simulations). We acknowledge that some field work recordings include background conversations that may have sociolinguistic value for the main footage. Those recordings are out of the reach of this study; however, future work may explore how audio separation models may help isolate primary and background sounds. We provide the perturbation functions and hyperparameter configurations for future scholars to investigate feature fidelity thresholds. Similar study cases may reinforce the results obtained in this work and lead to new linguistically grounded methodologies for audio model benchmarking.

7 Ethics Statement

Aligning with ethical and moral standards, we offer a new method to improve the quality of under-researched language corpora. We acknowledge the intricate nature of linguistic variability and its implications on the societal effects of technology. It is crucial for scholars to contribute to the creation of inclusive systems that accurately represent all members of society. The dissemination of these findings paves the way for a transparent and inclusive dialogue within the academic community that upholds respect for linguistic and cultural diversity. In the same way, we also aim to facilitate the progress of multilingual computational tools.

References

Virginia Best, Simon Carlile, Craig Jin, and André van Schaik. 2005. The role of high frequencies in speech localization. *The Journal of the Acoustical Society of America*, 118(1):353–363.

Michael K Brame and Ivonne Bordelois. 1973. Vocalic alternations in spanish. *Linguistic Inquiry*, 4(2):111–168.

Adelbert W Bronkhorst. 2000. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta acustica united with acustica*, 86(1):117–128.

Jeremy Calder, Rebecca Wheeler, Sarah Adams, Daniel Amarelo, Katherine Arnold-Murray, Justin Bai, Meredith Church, Josh Daniels, Sarah Gomez, Jacob Henry, et al. 2022. Is zoom viable for sociophonetic research? a comparison of in-person and online recordings for vocalic analysis. *Linguistics Vanguard*, page 20200148.

Kuan Chen, Bo Chen, Jiahao Lai, and Kai Yu. 2018. High-quality voice conversion using spectrogram-based wavenet vocoder. In *Interspeech*, pages 1993–1997.

E Colin Cherry. 1953. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979.

Kateřina Chládková, Paola Escudero, and Paul Boersma. 2011. Context-specific acoustic differences between peruvian and iberian spanish vowels. *The Journal of the Acoustical Society of America*, 130(1):416–428.

John R Deller Jr, John G Proakis, and John H Hansen. 1993. *Discrete time processing of speech signals*. Prentice Hall PTR.

Carlos Fonseca. 2023. 120h spanish speech.

Michael S Gradoville, Earl Kjar Brown, and Richard J File-Muriel. 2022. The phonetics of sociophonetics: Validating acoustic approaches to spanish/s. *Journal of Phonetics*, 91:101125.

Adriana Guevara-Rukoz, Isin Demirsahin, Fei He, Shan-Hui Cathy Chu, Supheakmungkol Sarin, Knot Pipatsrisawat, Alexander Gutkin, Alena Butryna, and Oddur Kjartansson. 2020. Crowdsourcing latin american spanish for low-resource text-to-speech. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 6504–6513, Marseille, France. European Language Resources Association (ELRA).

Umut Isik, Ritwik Giri, Neerad Phansalkar, Jean-Marc Valin, Karim Helwani, and Arvindh Krishnaswamy. 2020. Poconet: Better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss.

Ewa Jacewicz, Joshua M Alexander, and Robert A Fox. 2023. Introduction to the special issue on perception and production of sounds in the high-frequency range of human speech. *The Journal of the Acoustical Society of America*, 154(5):3168–3172.

Allard Jongman, Ratree Wayland, and Serena Wong. 2000. Acoustic characteristics of english fricatives. *The Journal of the Acoustical Society of America*, 108(3):1252–1263.

Viktor Kharlamov, Daniel Brenner, and Benjamin V Tucker. 2023. Examining the effect of high-frequency information on the classification of conversationally produced english fricatives. *The Journal of the Acoustical Society of America*, 154(3):1896–1902.

Haohe Liu, Qiuqiang Kong, Qiao Tian, Yan Zhao, DeLiang Wang, Chuanzeng Huang, and Yuxuan Wang. 2021. Voicefixer: Toward general speech restoration with neural vocoder. *Preprint*, arXiv:2109.13731.

Daniel McFadden. 1972. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics, Academic Press*, pages 105–142.

Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen. 2021. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1368–1396.

Geoffrey Stewart Morrison, Paola Escudero, et al. 2007. A cross-dialect comparison of peninsular-and peruvian-spanish vowels. In *Proceedings of the 16th international Congress of phonetic sciences*, pages 1505–1508. Citeseer.

Tamara Rathcke, Jane Stuart-Smith, Bernard Torsney, and Jonathan Harrington. 2017. The beauty in a beast: Minimising the effects of diverse recording quality on vowel formant measurements in sociophonetic real-time studies. *Speech Communication*, 86:24–41.

Barbara G Shinn-Cunningham and Virginia Best. 2008. Selective attention in normal and impaired hearing. *Trends in amplification*, 12(4):283–299.

Erik R Thomas. 2002. Sociophonetic applications of speech perception experiments. *American speech*, 77(2):115–147.

Erik R Thomas. 2013. Phonetic analysis in sociolinguistics. *Research methods in sociolinguistics: A practical guide*, pages 119–135.

Tomáš Vampola, Jaromír Horáček, Vojtěch Radolf, Jan G Švec, and Anne-Maria Laukkanen. 2020. Influence of nasal cavities on voice quality: Computer simulations and experiments. *The Journal of the Acoustical Society of America*, 148(5):3218–3231.

Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, et al. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12.

Chengzhu Yu, Kamil K Wójcicki, Philipos C Loizou, John HL Hansen, and Michael T Johnson. 2014. Evaluation of the importance of time-frequency contributions to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 135(5):3007–3016.

A Noise Generation

As mentioned in section 3, we modify the samples using Additive White Gaussian Noise (AWGN) implemented through a Python function. The AWGN implemented in this work is defined by

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \tag{1}$$

where we calculate the root mean square (RMS) of a given signal x_i.

We then use Equation 2 to generate random Gaussian noise z_{noise}. We add parameter λ, which is a scaling factor that allows to blend the noise in the background. For the purposes of this study, we used $\lambda = .1$, but other studies may benefit from experimenting with different parameter settings.

$$z_{\text{noise}} = \mathcal{N}(0, (RMS \cdot \lambda)^2) \tag{2}$$

Finally, we combine the original signal x_i with the Gaussian noise z_{noise} to get the corrupted file $x_i{}'$.

$$x_i{}' = x_i + z_{\text{noise}} \tag{3}$$

B Voicing Data

Original Voicing Measurements															
Peninsular Females				**Peninsular Males**				**Peruvian Females**				**Peruvian Male**			
t (s)	*un (s)*	*v (s)*	*v (%)*	*t (s)*	*un (s)*	*v (s)*	*v (%)*	*t (s)*	*un (s)*	*v (s)*	*v (%)*	*t (s)*	*un (s)*	*v (s)*	*v (%)*
.09	.09	0	0	.08	.08	0	0	.13	.11	.02	15.38	.14	.13	.01	7.14
.13	.12	.01	7.69	.09	.08	.01	11.11	.12	.12	0	0	.14	.14	0	0
.1	.09	.01	10.00	.08	.07	.01	12.50	.1	.1	0	0	.13	.12	.01	7.69
.1	.09	.01	10.00	.06	.06	0	0	.1	.09	.01	10.00	.11	.1	.01	9.09
.13	.11	.02	15.38	.06	.06	0	0	.11	.1	.01	9.09	.12	.11	.01	8.33
.12	.11	.01	8.33	.08	.07	.01	12.50	.09	.08	.01	11.11	.13	.13	0	0
.08	.08	0	0	.09	.07	.02	22.22	.09	.09	0	0	.1	.09	.01	10.00
.107	.099	.009	**7.344**	.077	.070	.007	**8.333**	.106	.099	.007	**6.512**	.124	.117	.007	**6.037**

Table 4: Voicing measurement for original tokens with intervocalic fricative (s) across all speakers. The last row indicates mean values.

Enhanced Voicing Measurements															
Peninsular Females				**Peninsular Males**				**Peruvian Females**				**Peruvian Males**			
t (s)	*un (s)*	*v (s)*	*v (%)*	*t (s)*	*un (s)*	*v (s)*	*v (%)*	*t (s)*	*un (s)*	*v (s)*	*v (%)*	*t (s)*	*un (s)*	*v (s)*	*v (%)*
.1	.1	0	0	.08	.08	0	0	.14	.12	.02	14.29	.14	.14	0	0
.13	.13	0	0	.09	.08	.01	11.11	.11	.11	0	0	.14	.13	.01	7.14
.1	.09	.01	10.00	.08	.07	.01	12.50	.11	.1	.01	9.09	.13	.12	.01	7.69
.11	.09	.02	18.18	.08	.07	.01	12.50	.1	.1	0	0	.11	.11	0	0
.13	.12	.01	7.69	.07	.07	0	0	.11	.1	.01	9.09	.11	.1	.01	9.09
.11	.11	0	0	.08	.07	.01	12.50	.09	.08	.01	11.11	.14	.13	.01	7.14
.08	.07	.01	12.50	.08	.07	.01	12.50	.09	.09	0	0	.09	.09	0	0
.109	.101	.007	**6.911**	.080	.073	.007	**8.730**	.107	.100	.007	**6.226**	.123	.117	.006	**4.438**

Table 5: Voicing measurement for enhanced tokens with intervocalic fricative (s) across all speakers. The last row indicates mean values.

C Models

Model	Coefficient	Df	Sum Sq	Mean Sq	F-value	p-value	Df	Sum Sq	Mean Sq	F-value	p-value
Voicing	Country	1	17.1	17.13	.433	.516	1	43.4	43.35	1.236	.277
	Gender	1	.5	.46	.012	.915	1	.0	.00	.000	.994
	Residuals	25	987.9	39.52			25	876.9	35.07		
Duration	Country	1	.003	.003	9	**<.01**	1	.003	.003	7.819	**<.01**
	Gender	1	0	0	.563	.46	1	0	0	.753	.393
	Residuals	25	.01	0			25	.009	0		

Table 6: Results of the ANOVAs for duration and voicing in original (left) and enhanced tokens (right).

Model	Coefficient	Value	Std.Error	t-value	p-value	Value	Std.Error	t-value	p-value
Voicing	Intercept	5.162	2.014	**2.562**	**.016**	4.158	2.009	**2.069**	**.048**
	countrySpain	1.227	2.326	.527	.602	2.46	2.32	1.060	.299
	genderM	-.079	2.326	-.034	.973	.044	2.32	.019	.984
Duration	Intercept	-4.585	.009	**-497.981**	**<.01**	-4.584	.008	**-541.376**	**<.01**
	countrySpain	-.024	.010	-2.2678	.032	-.021	.009	-2.206	.036
	genderM	-.003	.010	-.3655	.717	-.005	.009	-.578	.568

Table 7: Results of the RLMs for duration and voicing in original (left) and enhanced tokens (right).

Random effects	Name	Variance	Std.Dev.		Variance	Std.Dev.		
id	(Intercept)	0	0		0	0		
	Residual	46.48	6.818		45.55	6.749		
Fixed effects	Estimate	Std. Error	t value	p value	Estimate	Std. Error	t value	p value
(Intercept)	5.169	2.288	**2.259**	**.032**	4.085	2.265	1.803	.083
countrySpain	1.176	2.643	.445	.660	2.4005	2.616	.917	.367
genderM	-.130	2.643	-.05	.960	.1474	2.616	.056	.955

Table 8: Results of the RLMEMs for voicing in original (left) and enhanced tokens (right).

Random Effects	Name	Variance	Std.Dev.	Name	Variance	Std.Dev.		
id	(Intercept)	0	.027	(Intercept)	0	.027		
	Residual	0	.017	Residual	0	.015		
Fixed effects	Estimate	Std. Error	t value	p value	Estimate	Std. Error	t value	p value
(Intercept)	-4.584	.024	**-185.33**	**<.01**	-4.585	.024	**-185.56**	**<.01**
countrySpain	-.022	.028	-.79	.436	-.020	.028	-.71	.484
genderM	-.005	.028	-.2	.843	-.005	.028	-.18	.858

Table 9: Results of the RLMEMs for duration in original (left) and enhanced tokens (right).