INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

The Role of Formant and Excitation Source Features in Perceived Naturalness of Low Resource Tribal Language TTS: An Empirical Study

Ashwini Dasare, Pradyoth Hegde, Supritha Shetty, Deepak K T

Indian Institute of Information Technology, Dharwad, India

ashwini@iiitdwd.ac.in, pradyothhegde@gmail.com, supritha.shetty@iiitdwd.ac.in, deepak@iiitdwd.ac.in

Abstract

Text-to-speech synthesis is a prominent area in the speech-processing domain that has significant use in reading digital content in a given language. In the proposed work, we worked on two tribal languages of India *viz.,* Lambani and Soliga, which are zero-resource languages. The study began with a dataset collection for both tribal languages. Secondly, a Text-To-Speech (TTS) system was built separately based on the transfer learning approach. To validate the voice quality of TTS-generated speech, subjective as well as objective evaluations were performed. As a part of objective analysis, the voice source and vocal tract filter properties of the synthetic speech have been explored. The extensive study on various aspects of speech, such as LP residual, F0 contour, and formants (F1 & F2), have shown interesting results that can correlate to the subjective listening test results. The link to the original and synthetic speech can be found online. [1]

Index Terms: Index Terms: zero-resource, tribal languages of India, transfer learning, end-to-end TTS, voice source, formant analysis.

1. Introduction

Tribal languages aid in understanding the world, our origins, the ancestors from which we evolved, and the human race's capabilities. In India, about 197 spoken languages are classified as endangered by UNESCO [1]. According to the Indian government's 2011 census, any language spoken by less than 10,000 people may also be endangered. Lambani and Soliga are classified as scheduled tribes in India. These two tribes speak their own language. Soliga language belongs to the Dravidian family, and Lambani belongs to the Indo-Aryan family of languages. [2]. With barely 40,000 inhabitants, Soliga is already on the edge of extinction [3].

Preserving and passing ancient knowledge to the next generation is an integral part of the older generation. Preserving a language has a responsibility to preserve the art, culture, and heritage of a community. One such attempt is to build speech synthesis models and preserve the original content, style, and accent of the language. If we do not develop digital platforms for endangered languages, they will likely become extinct in less than a decade[4], which might lead to the loss of the core identical tribal language and its values.

Building Text-To-Speech(TTS) synthesizers for the tribal language could be one of the contributions to preserving the tribal language. It could be used in language learning, creating educational content, health, and voice commerce. Since tribal languages are either low or zero-resource, transfer learning is the best alternative to build a TTS from scratch[5]. Nvidia's tacotron2 [6] is a state-of-the-art model for the transfer learning approach. A few low-resource languages are built using the transfer learning approach [7, 5]. Therefore we have employed a transfer learning technique to build the TTS system using two neural vocoders *viz.,*HiFi-GAN [8] and Waveglow [9].

This paper has adopted a transfer learning technique to build the tribal TTS using Nvidia's tacotron2 model. The model-generated spectrogram is fed to Waveglow and HiFi-GAN vocoders. For Soliga and Lambani languages, both the vocoders have generated good quality audio. To validate the synthetic audio, subjective and objective measures are carried out. It is found that Waveglow performs slightly better than the HiFi-GAN. Further, we tried to compare the perceptually relevant aspects of the synthetic and original speech by analyzing the voice source signal, fundamental frequency, and vocal tract features like formants. Interesting analyses and correlations of perceptual listening and source filter features of the vocal tract were found.

The rest of the paper is organized as follows. The details of the data collection and preprocessing of speech utterances of the Lambani and Soliga languages are given in Section 2. The implementation details of the Lambani and Soliga TTS systems and also the comparative evaluation of the original and synthetic speech versions of the TTS system is presented in Section 3. Glottal and vocal tract analysis, fundamental frequency analysis and formant analysis are given in section 4. Some concluding remarks and scope for future work are presented in Section 5.

2. Dataset Details

The Lambani is a nomadic tribe. Therefore, only the northern Karnataka Lambani dialect is chosen for the TTS data collection. Soliga is spoken in regions near Biligiri Rangana Betta, which is in the southern part of Karnataka state. Both languages have no script. Since Kannada is the state's official language, Kannada script is used for both. A list of 10,000 English sentences was prepared from 'Swadesh' [10], which contains words relevant to the Indian village lifestyle, and used as a source transcript for translation. This set was translated into Kannada first, then to Soliga and Lambani languages by the literate people from the respective tribe. A software tool has been developed to record. The tool has provisions for recording speakers' details like name, gender, age, and educational background. Consent from every speaker is also recorded, where the participant must agree if he or she is willing to give their voice samples. It displays one sentence at a time. For the illiterates, a voice-over of the same is played and provides the speaker with appropriate time to utter the displayed sentence. The utterance can be replayed to verify their correctness and intelligibility. In

[1] https://audio-results-1.vercel.app/

10.21437/Interspeech.2023-1644

case of mistakes, there is also a provision for rerecording.

For the voice recording of both languages, literate female speakers of ages 34 and 38 of the Lambani and Soliga community were chosen, respectively, with good diction and voice for both languages. The speech data collection was recorded in a soundproof studio-quality environment in mono channel with a 44.1 kHz sampling rate and 16-bit depth. The duration of spoken sentences varies from 2 to 10 seconds. A total of 8 hours of Lambani and 6 hours of Soliga voice samples are recorded in the dataset collection process.

3. Tribal Text-to-Speech Synthesis

Nvidia's tacotron2 is the modified architecture of Tacotron2 [11]. It uses dropout instead of zone out for the regularization of LSTM layers and Waveglow for synthesis. It is trained on tensor cores for faster convergence. Any model can be easily fine-tuned on top of a publically released pre-trained model with a lesser time of convergence. To build tribal TTS, apart from Nvidia's tacotron2 and Waveglow, we have also explored HiFi-GAN vocoder for waveform synthesis. We have fine-tuned Nvidia's pre-trained model with 8 hours of Lambani and 6 hours of Soliga speech data for Lmabani and Soliga TTS, respectively. Note: the pre-trained model was originally trained on the publicly available LJ speech data-set [12].

3.1. Evaluation

To evaluate the tribal TTS models, Mean Opinion Score (MOS) and Perceptual Evaluation of Speech Quality (PESQ) estimation have been conducted. The PESQ metric used in the analysis is developed by [13]. This objective measure gives the perceptual quality of synthetic speech in comparison with original speech on a scale of 1 to 5. Any score above 3 is considered a good voice quality. The subjective evaluation included 20 literate speakers of both languages. The speakers were asked to give a score from 1-5 where 1 being the lowest and 5 being the highest. The original and synthetic speech generated by the vocoders were all mixed up, and listeners were asked to rate each voice based on intelligibility, diction, prosody, and accent parameters. The analysis found that Waveglow outperforms HiFi-GAN in both PESQ and MOS evaluations, as shown in table 1. The perceptual parameters for Waveglow were almost near to human quality, as shown in Figure 1 and Figure 2.

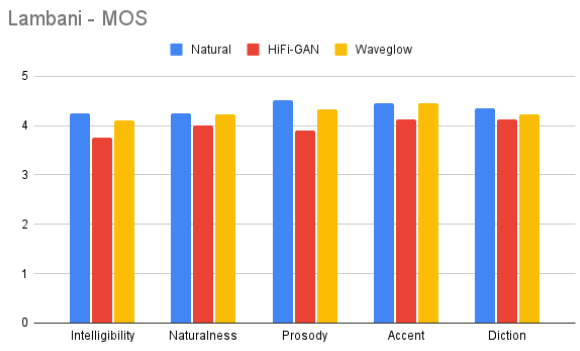

Figure 1: *Mean opinion score of Lambani on a scale of 1 to 5.*

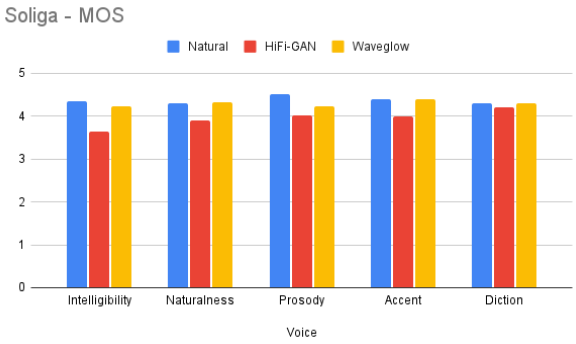

Figure 2: *Mean opinion score of Soliga on a scale of 1 to 5.*

4. Glottal and Vocal tract feature Analysis

In the evaluation of the Lambani and Soliga TTS, the performance of the Waveglow on par with the original speech made us advent to the expedition of analyzing the glottal and vocal tract features. A parametric evaluation of the synthetic speech is conducted considering different aspects of speech *viz.,* voice source signal and filter characteristics. The voice source properties can be analyzed by estimating the speech signal's Linear Prediction Residual (LPR). A similarity distance between original-synthetic signals shows how well the synthetic speech has captured the source properties of the original signal. The dynamic Time Warping (DTW) technique estimates the distance between the original and synthetic speech of two neural vocoders. In addition to distance measure, we also evaluate the significant excitation of the vocal tract *viz.,* glottal closure instant (GCI). Fundamental frequency (F0) is one of the acoustic features that reflects prosodic characteristics in speech. The dynamics of F0 are measured and investigated. In this study, we have considered 500 randomly generated TTS audio files from both languages for testing purposes.

4.1. Voice Source Analysis

Linear Prediction Residual (LPR) signal represents the glottal source waveform of the speech signal. The signal embeds several features that characterize prosody, speaker information, and pathology. The signal can be obtained using the Linear Prediction analysis method [14]. The initial analysis of the LPR of original and synthetic speech is conducted based on distance measure parameters. Dynamic Time Warping (DTW) is applied to each pair of signals to measure distance. Lesser distance value indicates closer to the original signal. On taking the average of distance value of every original-synthetic pair, both the vocoder's output performs equally well. The glottal source signal of the synthetic speech generated by both vocoders is similar.

Apart from LPR signal analysis, we have also looked at the synthetic speech in terms of Glottal Closure Instant (GCI). We have employed Zero Frequency Resonator (ZFR) [15], a state-of-the-art glottal instant estimation algorithm for extracting GCIs from the raw speech signal. The discontinuity in the speech signal due to the impulse-like excitation signal can be observed in all frequencies, including zero. Since the vocal tract filter resonates at much higher frequencies than zero frequency, the output of the resonator with zero frequency should have information of discontinuity. The ZFR signal can be estimated

Table 1: *Objective (PESQ) and subjective (MOS) evaluations for two different vocoders.*

Language	PESQ		MOS		
	HiFi GAN	Waveglow	Original	HiFi GAN	Waveglow
Lambani	3.7	**3.8**	4.3 $\pm$1.2	3.9 $\pm$1.1	**4.2 $\pm$0.9**
Soliga	3.6	**3.7**	4.5 $\pm$1.1	3.7 $\pm$1.0	**4.3 $\pm$0.7**

by passing the speech signal through a cascade of two zero-frequency filters followed by a trend removal operation. The filtered signal shows a significant change at the positive zero crossings, indicating GCI locations.

Metrics such as Identification rate (IDR), miss rate (MR), false alarm rate (FAR), and identification accuracy(IDA) are considered [16] to compare. The GCIs are extracted from the original speech signal and are considered as the reference data. The measures for the synthetic speech are shown in Table 2. The IDR of GCIs in Waveglow model output is slightly higher for both languages as compared to Hi-Fi GAN. The IDR indicates the identification of GCIs in both original and synthetic speech frames.

Table 2: *Performance comparison of GCI estimation using ZFF algorithm for Soliga and Lambani language. Note: IDR, MR, and FAR are expressed in $\%$ while IDA in ms.*

	Model	IDR	MR	FAR	IDA
Soliga	Waveglow	**67.20**	**23.40**	9.38	0.53
	HiFi-GAN	66.20	24.40	9.38	**0.52**
Lambani	Waveglow	**57.41**	**30.06**	12.54	**0.64**
	HiFi-GAN	56.35	31.44	**12.21**	0.65

4.2. Analysis of Fundamental Frequency (F0)

Fundamental frequency (F0) is also an important feature that captures prosodic characteristics in speech. The rise and fall in the F0 pattern contribute to the naturalness of speech. A study in [17, 18] has shown the influence of F0 contour in the perception of speech signals such as Lombard speech. Therefore, in this study, we have used F0 as one of the parameters to examine the synthetically generated speech.

The F0 of the original and synthetic speech signals are extracted using the Praat toolkit [19]. Waveglow and HiFi-GAN model outputs are compared to and evaluated using Gross Pitch Error (GPE) metric. GPE is the ratio of relative error of the predicted value, which is greater than the threshold to the number of voiced frames. The threshold set 20$\%$ for the GPE estimation. Figure 3 and 4 depict F0 contour plots of the original and synthetic signal of both models and languages. The same sequence of plots is followed in both figures. The raw speech of the original and synthetic are represented in plot(a) and plot(b-c), respectively. Similarly, the F0 contour plots of the same are represented in plot(d) and plot(e-f). As per the observation in Figure 3, a rise in the F0 pattern of the original signal at a time between 0.6-0.8 sec in the plot(d) is clearly captured in the Waveglow generated speech as shown in the plot(f). However, the F0 pattern in the HiFi-GAN generated signal is flat (refer figure 3-plot(e). This observation is highlighted in a dashed red line. Also, a similar observation is seen in figure 4. The dip in the F0 value of the original signal in plot(d) is clearly captured by the Waveglow model in plot(f) and this variation seems to be missing in HiFi-GAN model output in plot(e).

In addition, to the above analysis, table 3 shows the over-

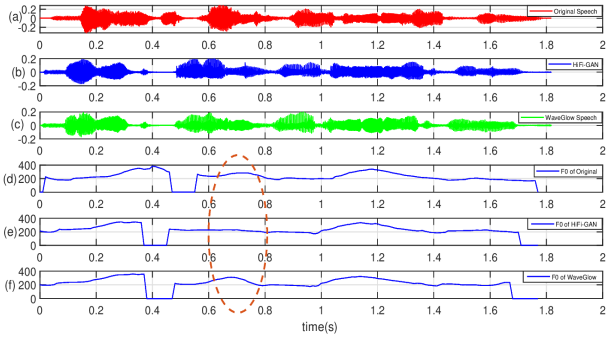

Figure 3: *F0 contour plots of the original Lambani speech (a,d) and its corresponding synthetic speech (HiFi-GAN-b,e)(Waveglow-c,f) from the proposed work.*

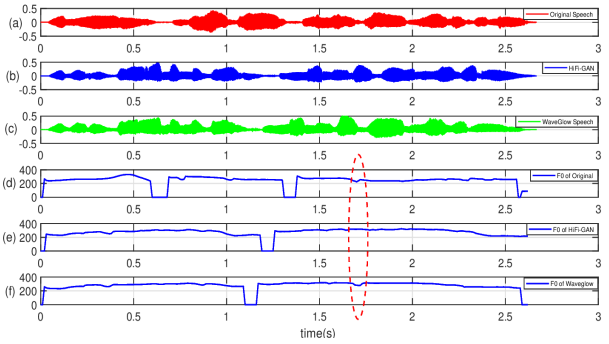

Figure 4: *F0 contour plots of the original Soliga speech (a,d) and its corresponding synthetic speech (HiFi-GAN-b,e)(Waveglow-c,f) from the proposed work.*

all performance of both the neural vocoders. The GPE for the Waveglow model is less by 1-2$\%$ as compared to the HiFi-GAN model for both languages. Owing to the above results, we clearly see the reason for Waveglow generated synthetic speech performing well in subjective listening tests.

Table 3: *Comparision table of Gross pitch error*

	Model	Average GPE ($\%$)
Soliga	Waveglow	**24.86**
	HiFi-GAN	26.72
Lambani	Waveglow	**43.42**
	HiFi-GAN	44.16

4.3. Formant analysis

In speech, the presence of vowels is more than consonants. Manual extraction of 50 vowels each across the original speech, HiFi-GAN, and Waveglow of Lambani and Soliga languages

were performed. Formants are extracted using Praat [19]. The first two formants are taken to analyze the vowels of the original and TTS-produced ones. Though both languages have two unit times of the respective free vowel, the vowel sound is considered for the analysis irrespective of the time. The first formant characterizes the openness of the mouth when producing vowels, and the second formant estimates the tongue's position.

We carried out the formant analysis to see how much the original and synthetically produced data are similar. Table 4 and Table 5 show the mean formant values of the free vowels present in the Lambani and Soliga languages.

The Figure 5 and Figure 6 plots the mean and standard deviation of five vowels of Lambani and Soliga languages, respectively. The standard deviation of the TTS vowels is visibly lesser than that of the original. The mean of the first two formants of synthetic vowels is within the standard deviation of the naturally uttered speech. The natural variation in the original vowel is more than the synthetically produced speech.

In characterizing the speakers of two languages, the Soliga speaking lady has more intonation variation and stretches the duration of the vowels. The style of the pronunciation has shown that the first formant values are generally higher than those of Lambani language speaker. By looking at the formant, it can be seen that the characteristics of the vowels are almost retained by the model and produced by both vocoders. This is one of the possibilities for the synthetic sound to retain the speaker and language characteristics.

Table 4: *First and Second formant mean values of the original, Waveglow & HiFi-GAN produced speech of Lambani language*

Lambani		a	e	i	o	u
F1	Original	745	524	421	512	416
	Waveglow	689	500	389	512	428
	HiFi-GAN	664	527	413	539	432
F2	Original	1606	2315	2547	1265	1281
	Waveglow	1666	2484	2481	1371	1250
	HiFi-GAN	1612	2406	2721	1197	1203

Table 5: *First and Second formant mean values of the original, Waveglow & HiFi-GAN produced speech of Soliga language*

Soliga		a	e	i	o	u
F1	Original	852	548	402	565	443
	Waveglow	834	609	461	601	498
	HiFi-GAN	743	548	388	558	409
F2	Original	1603	2036	2344	1282	1164
	Waveglow	1658	2197	2416	1252	1119
	HiFi-GAN	1594	2198	2263	1212	1290

5. Conclusion

This paper proposes fine-tuning low-resource data with Nvidia's pretrained model to get mel-spectrograms for Lambani and Soliga tribal languages. Two different vocoders, *viz.*, Waveglow and HiFi-GAN are used for converting the mel-spectrogram to the speech wave. From the mean opinion scores, we see that the TTS has produced good quality original-like speech and Waveglow scores were more than the HiFi-GAN vocoder. Further, we tried to study the retention of the speaker and language characteristics in the synthetic speech with two

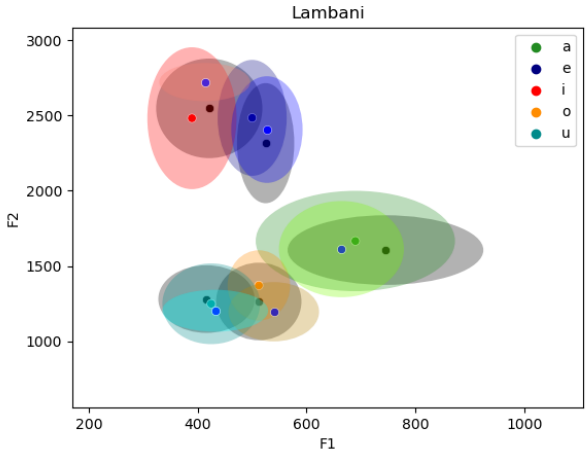

Figure 5: *Mean and standard deviation of the vowels of Lambani language. Black, dark color shade and light color shade represents original, Waveglow & HiFi-GAN respectively.*

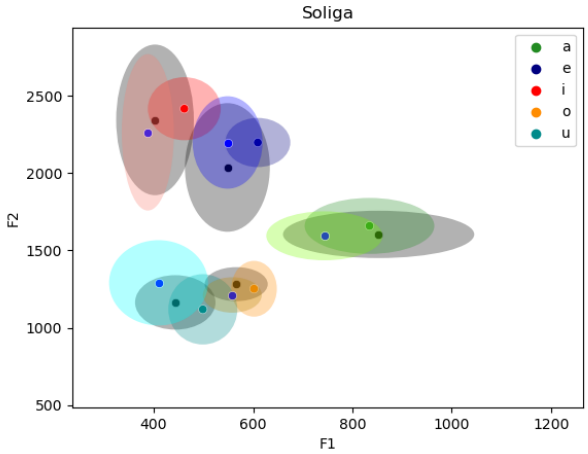

Figure 6: *Mean and standard deviation of the vowels of Soliga language. Black, dark color shade and light color shade represents original, Waveglow & HiFi-GAN respectively.*

vocoders. We have explored some of the relevant aspects of speech such as LP residual, F0 contour, and formants (F1 & F2) to examine. In vocal source analysis, the IDR of the GCI of Waveglow vocoder came out to be better. Similarly, the GPE analysis of the fundamental frequency (F0), was found to be about 2% lesser. i.e. Waveglow model seems closer to the original speech. In formant analysis, the mean of the first and second formants of the Waveglow was well within satisfactory limits. Based on the above study, we were able to relate the performance of both neural vocoders to their corresponding MOS score.

6. Acknowledgment

The authors would like to thank the Ministry of Electronics and Information Technology, Govt. of India and NMICPS TiHAN, IIT Hyderabad, India.

7. References

[1] C. Moseley, *The UNESCO atlas of the world's languages in danger: Context and process*. World Oral Literature Project, 2012.

[2] M. B. Emeneau, "India as a lingustic area," *Language*, vol. 32, no. 1, pp. 3–16, 1956.

[3] C. Chandramouli and R. General, "Census of india 2011," *Provisional Population Totals. New Delhi: Government of India*, pp. 409–413, 2011.

[4] T. Haokip, "Artificial intelligence and endangered languages," *Available at SSRN 4212504*, 2022.

[5] B. Külebi, A. Öktem, À. Peiró Lilja, S. Pascual, and M. Farrús, "Catotron–a neural text-to-speech system in catalan," *Proceedings of Interspeech 2020; 2020 Oct 25-29; Shanghai, China.[Baixas]: ISCA; 2020.*, 2020.

[6] V. Rafael, P. Raul, T. Roman, A. Ivan, S. Sih, S. Taras, B. Caley, and K. Grzegorz, "Nvidia's tacotron2," https://github.com/NVIDIA/tacotron2.git, 2020.

[7] A. Debnath, S. S. Patil, G. Nadiger, and R. A. Ganesan, "Low-resource end-to-end sanskrit tts using tacotron2, waveglow and transfer learning," in *2020 IEEE 17th India Council International Conference (INDICON)*. IEEE, 2020, pp. 1–5.

[8] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[9] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.

[10] M. Swadesh, "Lexico-statistic dating of prehistoric ethnic contacts: with special reference to north american indians and eskimos," *Proceedings of the American philosophical society*, vol. 96, no. 4, pp. 452–463, 1952.

[11] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[12] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[13] R. G. D. Miao Wang, Christoph Boeddeker and ananda seelan, "Pesq (perceptual evaluation of speech quality) wrapper for python users," May 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6549559

[14] J. Makhoul, "Linear prediction: a tutorial review," *Proceesdings of the IEEE*, pp. 2054–2062, 1975.

[15] K. S. R. Murty and B.Yegnanarayana, "Epoch extraction from speech signal," in *IEEE TASLP*, 2008.

[16] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 994–1006, 2011.

[17] B. K. Kumar, P. Gangamohan, and S. V. Gangashetty, "Contribution of f0 contour level, f0 contour shape and durations towards perception of lombard speech," in *Proc. SMM21, Workshop on Speech, Music and Mind*, 2021, pp. 16–20.

[18] J. Carroll, S. Tiaden, and F.-G. Zeng, "Contribution of f0 contour level, f0 contour shape and durations towards perception of lombard speech," *Journal of the Acoustical Society of America*, pp. 2054–2062, 2011.

[19] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, pp. 1–15, 2018.