A spectrotemporal modulation application for distinguishing modal and whistled speech

Benjamin O'brien, Anna Marczyk

A spectrotemporal modulation application for distinguishing modal and whistled speech

Abstract

Machine learning procedures have become efficient at using acoustic information to distinguish modal speech, but their application on other speech modalities remains largely overlooked. Operating under the hypothesis that non-spoken speech classification requires a more robust acoustic analysis in which to train models, this study applies a spectrotemporal modulation (STM) analysis to examine its effectiveness at distinguishing modal and whistled Spanish speech. A linear mixed-effects model revealed that in comparison to modal speech, whistled speech had significantly increased spectrotemporal modulations generally above 1 cycles per octave. Features identified as significantly relevant for distinguishing speech modalities were extracted, and an automatic speech modality classification task was developed. STM- and MFCC-feature models performed similarly, boasting weighted accuracy performances of 92% and 94%, respectively. Although an STM analytical approach had not previously been applied to whistled speech, our results support existing evidence on its acoustic characteristics. Moreover, our findings have key implications for using STM-features for automatic speech modality classification tasks, as it reduced the feature space at little cost to performance.

Keywords: automatic speech modality classification, whistled speech, non-speech vocalisations, spectrotemporal modulation

1 Introduction

Machine learning procedures have become efficient at using acoustic information to distinguish modal speech, but their application on other speech modalities remains largely overlooked. Fortunately, recent work has applied automatic speaker verification methods to non-typical speech, e.g., pathological speech [1, 2] and non-spoken speech, such as whispered [3] and shouted speech [4]. Nevertheless, despite the productive and perceptual differences between spoken and non-typical speech, oftentimes the models are trained on traditional acoustic features, i.e., cepstral coefficients, associated with modal speech. It is plausible that models designed to distinguish different speech modalities might enhance performance when trained on acoustic characteristics associated with non-typical speech. However, this depends on

the extent to which these characteristics provide distinct and reliable patterns for distinguishing speech modalities.

One example of a non-typical speech is non-spoken speech, which, in general, can be characterized as a vocalization used to convey meaning through a transformed speech signal. Whistled speech is an example that, unlike its spoken modality, does not require vocal folds. Instead it relies on the flow of compressed air trapped in the anterior cavity by manipulating lips, incisors, or fingers for its production (see [5] for a comprehensive review of whistled languages). Whistled speech is based on a spoken language, however, it expressed through a different modality [6]. Unlike whispering or shouting, the transformation of the speech signal is much more extreme, causing phonetic details to be smoothed or erased entirely [7]. The anterior cavity is the resonator that determines the intensity and high frequencies that are characteristic of whistling. Based on these differences with its spoken modality, whistling speech a good candidate for further study, however, additional work is needed to identify which acoustic domains and dimensions are useful for automatic speech processing.

The purpose of the current study was to identify acoustic features that were capable of distinguishing modal and whistled speech modalities and efficient for modeling. To do so, two experiments were designed. The first experiment relied on statistical analyses to examine a feature space that we hypothesized would best capture the acoustic characteristics of whistled speech. Based on any significant differences between the speech modalities, the second experiment selected the most relevant acoustic features to train machine learning models and tested them on a separate dataset. By reporting our findings, our goal was to add to the discussion surrounding the acoustic characteristics of whistled speech and how derived features played a role on automatic speech modality classification performance.

2 Related work

There are approximately 80 world populations that have adapted their local spoken language to a whistled modality [5]. In general, whistled speech can be described as having a strong, clear signal that is used to communicate over long-distances [8]. By not being forced to vibrate, the vocal folds act as a conduit for air flow, which amplifies the speech signal, but also limits the presence of phonemes, thus reducing the frequency space, i.e., higher order formants. More specifically, whistling transposes speech signal frequencies to within 1-4 kHz [5], a range humans are sensitive to and associate with speech [9]. Signal intensity can vary between 75 dB and 120 dB, which, again, can help propogate speech signals over large distances with minimal distortion, e.g., natural obstacles, noise. While there are tonal whistled languages, the current study focused on the non-tonal Silbo Gomero whistled form of Spanish spoken on La Gomera in the Canaries. This decision was based on the volume of available recordings and their use in a previous work involving automatic speech recognition [10].

One approach to improving our understanding of the differences between modal and whistled speech is to examine the dynamic temporal and spectral modulations underpinning the signals. Numerous studies have focused on unpacking modulation

representations of speech signals in the context of assessing intelligibility [11, 12] and voice pathology [13, 14], shouted speech [15], and more recently laughter [16, 17]. Unlike time-frequency representations, e.g., spectrograms, spectrotemporal modulation (STM) representations unpack signals in terms of joint spectral and temporal modulations, thus providing a combined and more comprehensive characterization of the sound. Moreover, previous work has applied STM analysis to topics such as birdsong [18] and communication in reverberant environments [19], suggesting its usefulness for further study, given whistled speech is generally used for long distance communication in natural settings. To the best of our knowledge, a STM analytical method has not yet been used to evaluate the acoustic characteristics of whistled speech, although its implementation could offer considerable advantages. Given the characteristics of whistled speech as outlined above, our goal was to examine whether STM-features were sensitive to differences between these speech modalities.

Although limited, previous work has relied on STM-features for automatic speech processing tasks, including speaker verification in the contexts of reverberant environments [19] and voice-spoofing [20]. Typically, models are trained on Mel-Frequency Cepstral Coefficients (MFCC), which scale the frequency components in audio recordings to the Mel-scale, a domain that is better suited to model the human auditory perception. While they are effective at transforming the spectral information encoded in an audio recording into compressed, decorrelated vector, it is plausible that nuances associated with non-typical speech signals are not represented in the vector. As [10] and others suggest, models are then trained on constrained information and, in turn, limiting their effectiveness. As evidence has shown machines and humans model speech differently [21], our goal was to examine which STM-features distinguished modal and whistled speech and whether this information was effective when training automatic speech modality classification models.

3 Methods

3.1 Corpus

Table 1 describes the speech stimuli used in the experiments. The corpus was divided into Train and Test datasets, each containing modal and whistled speech recordings with no overlap between them. As described in further detail below, the Train dataset was used in Experiments 1 and 2, whereas the Test dataset was used solely to test models developed in Experiment 2.

Table 1: Description of the corpus

	Train		Test	
	Modal	Whistle	Modal	Whistle
Samples (#)	436	463	180	44
Duration (s)	3.8 ± 0.9	7.2 ± 2.9	5.36 ± 1.59	1.89 ± 0.76
Corpus	Common Voice	[10]	Common Voice	[7, 22, 23]

The Train dataset contained 463 whistled speech recordings produced by four male whistlers (see [10] for corpus details). The Test dataset contained 44 whistled speech recordings produced by a different group of 4 male speakers (see the Supplementary Materials of [7, 22, 23] for more information). Following similar procedures described in [10], the Train and Test modal speech recordings were obtained from the Mozilla Common Voice's Spanish Common Voice Corpus 4[1] data set. Although Common Voice provides speaker sex information, speaker identifications were anonymised. Modal Spanish speech recordings in the Train dataset (N = 436) were the same as those used in [10], while the Test dataset (N = 180) contained randomly selected Spanish speech recordings made by male speakers. Although the Modal:Whistle speech ratio in the Train dataset was approximately 1:1, the decision to have a 4:1 ratio for the Test dataset was based on our interest in developing a dataset that reflected the natural unbalance between speech modalities.

Prior to data processing, all stimuli were down-sampled to 8 kHz and normalized such that the maximal amplitude of each recording was adjusted to a target of 100% of the signal dynamic. Finally, to ensure MPS representations had the same dimensions, all recordings were zero-padded (see [24]).

3.2 Spectrotemporal modulation features

The spectrotemporal modulation domain can be characterized in terms of the Modulation Power Spectrum (MPS). Previous work [11, 25, 26] has defined MPS as a two-dimensional Fourier transform of the time-frequency representation of an audio signal. Equation 1 provides the formal definition of the MPS, where s and r are spectral and temporal modulations, respectively, and $Y(t, f)$ is the amplitude extracted from the Fourier transform:

$$\mathrm{MPS}(s,r) = \int\int \mathrm{Y(t,f)} e^{-2\pi i s f} e^{-2\pi i r t} df dt \tag{1}$$

Similar procedures described in [14] and [16] were used to obtain MPS representations of the speech recordings. All processing was done in MATLAB 2021a (MathWorks Inc, USA) and based on adaptations to scripts described in [12]. Time-frequency representations were obtained using a gammatone filter bank summation method (128 full-width half-maximum Gaussians with center frequencies logarithmically spanning the frequency domain). Hilbert transforms were then used to extract the analytical amplitudes from these filter outputs. The *fft2* MATLAB function transformed the time-frequency representations into the modulation domain, where temporal and spectral modulations were measured in increments of 0.05 Hz and 0.22 cycles per octave (c/o), respectively. The amplitudes of positive and negative temporal and spectral modulations were averaged (A) and transformed into decibels, i.e., $20 * log_{10}(A/Ref)$ where Ref was set to the peak value of 16-bit audio (32,678).

Temporal and spectral modulation boundaries ranged from 0 to 20 Hz and 0 to 3 (c/o), respectively. The former was selected due to findings reported in [15], which associated < 20 Hz temporal modulations with communicative meaning, while the latter was identified as speech perception threshold in [11, 12]. Thus, for each speech

[1] https://commonvoice.mozilla.org/

recording we obtained a 280-feature MPS representation, i.e., a 2-D matrix of amplitudes associated with temporal modulations (N = 20) by spectral modulation (N = 14).

4 Experiments

4.1 Experiment 1

4.1.1 Experimental setup

The goal of Experiment 1 was exploratory, aiming to identify and validate whether spectrotemporal modulation (STM) features were useful when distinguishing modal and whistled speech modalities. Developed as a two-step experiment, the Train dataset was split into two parts, where 80% (N = 718) was used to extract STM-features and model speech modalities, henceforth the *exploratory* dataset, and the remaining 20% (N = 181) was used for validation, henceforth *validation* dataset.

First, a linear mixed-effects model (LMEM) was applied to the exploratory dataset (*lm* from the *lme*4 R-package). Model 2 set the spectrotemporal modulation amplitude (Amp) as the dependent variable and Modality (modal, whistled) as a fixed effect. To examine any differences between speech modalities and STM-features, our model included the interaction between Modality and Rate, which expressed a coordinate on the 2-D spectrotemporal modulation feature space, e.g., coordinate (2, 2) represented 1 Hz temporal modulations and 0.44 c/o spectral modulations. As our model only included Rate in interaction with Modality, it was treated as a nested factor. Following [15], Bonferroni correction was applied to correct for multiple comparisons (N = 50,679), where α = 0.001. Estimated marginal means (*emmeans* R-package) were used to carry out pairwise comparisons.

$$\text{lm}(\text{Amp} \sim \text{Type} + \text{Type:Rate}, \text{data} = data) \tag{2}$$

Based on these results (see Section 4.1.2), adjacent STM-features that were identified as significantly relevant for distinguishing speech modalities were bounded into separate regions and averaged. Thus, for each recording in the validation dataset, a mean amplitude was calculated for each region. A linear mixed-effects model (Model 3) was designed to evaluate any differences between regions across speech modalities, where the Mean amplitude of the region was set as a dependent variable and Modality (modal, whistled) was set as a fixed factor. To examine any differences between speech modalities across regions, our model included the interaction between Modality and Region, which was treated as a nested factor. Similarly, post-hoc Bonferroni-adjusted t-tests were carried out using the *emmeans* R-package.

$$\text{lm}(\text{Mean} \sim \text{ Modality} + \text{Modality:Region}, \text{data} = data) \tag{3}$$

4.1.2 Results

Model 2 was statistically significant $F_{559,200480}$ = 3995, SE = 2.8, $p < 0.001$ (Intercept: β = 28.3, SE = 0.15, t = 191.96, $p < 0.001$) with a strong fit to the data (adjusted R^2 = 0.92). We observed a significant fixed effect on Modality β = 3.4, SE = 0.21, t = 16.56, $p < 0.001$. Posthoc pairwise tests showed Whistled speech (-16.74 $\pm$ 0.01 dB) had a significantly increased spectrotemporal modulation energy in comparison to Modal speech (-18.91 $\pm$ 0.01 dB), $p < 0.001$.

Figure 1 illustrates the t-values based on pairwise comparisons between Modal and Whistled speech across spectrotemporal modulations. In comparison to Modal speech, Whistled speech had a much more expansive, but nuanced, spectrotemporal modulation profile, largely operating at spectral modulations above 1 cyc/oct and across the temporal modulation domain (0-20 Hz, red in Fig. 1), henceforth *Region 1*. Alternatively, Modal speech exhibited a relatively uniform increase in temporal modulations ranging from 4-20 Hz and spectral modulations ranging from 0 to 0.22-0.44 cyc/oct (blue in Fig. 1), henceforth *Region 2*. Region boundaries were set at $t > 3$ (Region 1) and $t < -3$, respectively. Overall, the 280 STM-feature space was reduced to 248 features when combining Region 1 (N = 217) and Region 2 (N = 31).

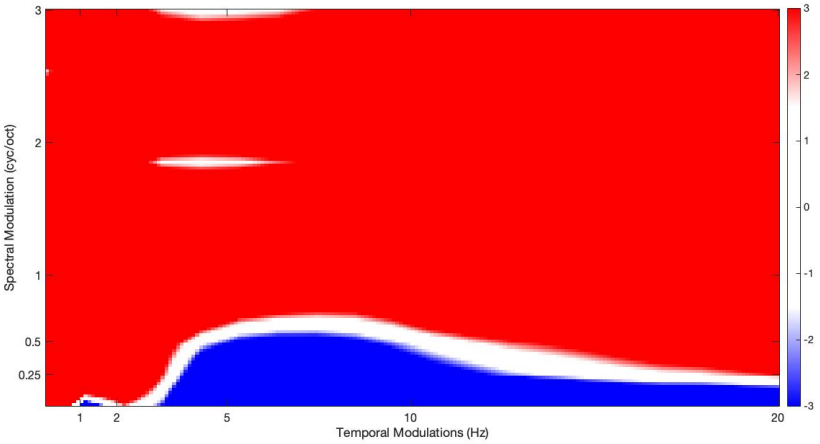

Fig. 1: Color map representing the posthoc pairwise comparisons (t-value) between modal and whistled speech across the spectrotemporal modulation domain (temporal modulations: 0-20 Hz; spectral modulations: 0-3 cyc/oct). Region 1 (red) and Region 2 (blue) represent all spectrotemporal modulations where t-values were $t > 3$ and $t < -3$, respectively.

Model 3 was statistically significant $F_{3,358}$ = 583, SE = 2.28, $p < 0.001$ (Intercept: β = -20.84, SE = 0.24, t = -85.66, $p < 0.001$) with a strong fit to the data (adjusted R^2 = 0.83). Our statistical analysis revealed a significant fixed effect on Modality β = 2.33, SE = 0.34, t = 6.87, $p < 0.001$. Post-hoc pairwise comparisons showed

differences between Modal and Whistled speech for Region 1 (Modal: -20.84 $\pm$ 0.24 dB; Whistled: -18.51 $\pm$ 0.24 dB), and Region 2 (Modal: -8.21 $\pm$ 0.24 dB; Whistled: -11.85 $\pm$ 0.24 dB), $p < 0.001$. Figure 2 illustrates the distribution and means across Modality and Region factors.

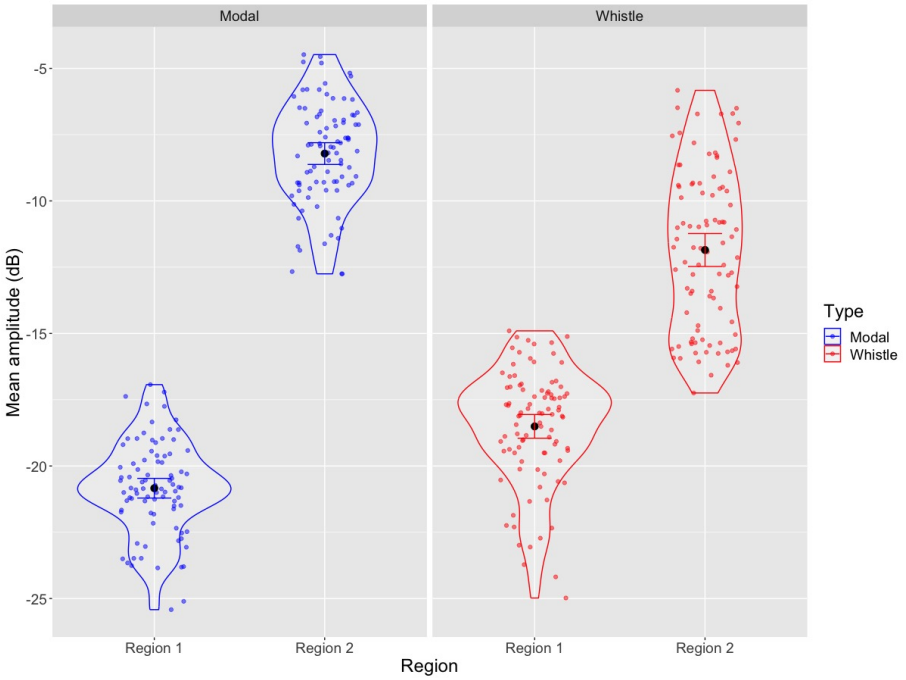

Fig. 2: Mean amplitude (dB) for regions across speech modalities. Significant differences were observed between speech modalities for Region 1 and Region 2, $p < 0.001$.

4.2 Experiment 2

4.2.1 Experimental procedure

The goal of Experiment 2 was to examine whether STM-features were effecient for modeling modal and whistled speech. To do so, an automatic speech modality classification task was developed and performed by two separate models. The STM model used features identified in Experiment 1. To compare its performance, a second model was developed that was trained on traditional Mel-Frequency Cepstral coefficients (MFCC). To obtain the latter, we extracted 13 MFCCs and the first and second derivatives resulting in 39-feature vector (window: 25 ms; step-size: 10 s) using the *librosa* Python package. Each model was fitted using the Python scikit-learn library. A 4-fold cross-validation design was performed on the Train dataset, whereupon the models were tested on the Test dataset. A Support Vector Machines (SVM) model

was selected, as our goal was to identify the best margin between modal and whistled speech modalities. Using hyperparameterization, both models were set with 128 coefficients and linear kernels.

4.2.2 Results

Table 2 illustrates the results of our STM and MFCC-feature models. Both models achieved high performances when distinguishing speech modalities, however, the MFCC-feature model enhanced performance slightly (weightd average: 94%) in comparison to the STM-feature model (92%).

Table 2: Results of Spectrotemporal Modulation (STM) and Mel-Frequency Cepstral Coefficients (MFCC) feature models. Precision, Recall, and F1-scores are reported across speech modalities and weighted averages (WA).

Model	Modality	Precision	Recall	F1
STM	Modal	0.98	0.92	0.95
	Whistle	0.74	0.91	0.82
	WA	0.93	0.92	0.92
MFCC	Modal	0.98	0.94	0.96
	Whistle	0.80	0.91	0.85
	WA	0.94	0.94	0.94

5 Discussion

The overall goal of these works was to examine whether spectrotemporal modulation (STM) features were capable and efficient at distinguishing modal and whistled speech. Our Experiment 1 findings revealed two regions of the STM-feature space as characterizing differences between speech modalities. Based on these observations, Experiment 2 was designed to compare automatic speech modality classification performance by models trained on these specific STM-features and on traditional cepstral coefficients. In general our findings suggest STM-features are effective at distinguishing speech modalities. Moreover, the results of Experiment 2 suggest that models trained on STM-features perform similarly to those trained on traditional acoustic metrics, despite the limited number of features required.

Our Experiment 1 results revealed two regions of spectrotemporal modulation (STM) features where modal and whistled speech differed. On one hand, whistled speech exhibited an increase in modulations in Region 1 in comparison to modal speech. In general, this region can be characterized as spectrotemporal modulations above 1 cyc/oct. As described in [11], higher spectral modulations are associated with pitch, whereas lower spectral modulations represent formants and higher order harmonics. This observation is consistent with previous acoustic characterizations of whistling, where the absence of vibrating vocal folds reduces the production of

phonemes and, in turn, the frequency space [5]. It is important to note that this difference between speech modalities occurs mainly in the spectral modulation range associated with formants, where modal speech may be richer in comparison to whistled speech. On the other hand, in Region 2 (temporal modulations: ranging from 4 to 20 Hz; spectral modulations: ranging from 0 to approximately 0.5 cyc/oct), we observed that modal speech exhibited an increase in energy in comparison to whistled speech. It is in this region that phonemes exist, as slow spectral modulations with temporal modulations in the 4-8 Hz and 8-16 Hz ranges have been associated with syllabic rhythm and articulation, respectively (see [27]). Numerous studies have attributed this region as critical for speech comprehensibility (see [11, 12], among others). These findings support previous work suggesting certain phonetic details may be lost during whistled speech production [7].

The results of Experiment 2 suggest both STM- and MFCC-feature models performed similarly. Although the MFCC-feature model slightly outperformed the STM-feature model, two major takeaways were observed. First, the size of the feature spaces were not the same. While the dimensionality of the STM-feature vector was 248, the average number of MFCC-features was 21801 (vector dimensions: 39 by 559.1 $\pm$ 228.0). Thus, despite being nearly 0.01 the scale of the MFCC-feature vectors, the STM-feature vectors were quite sensitive to these differences in speech modalities. Although STM-feature vector were more compact in size, it is difficult to attribute the efficiency of their use, given the pre-processing required in comparison to the ease in which MFCCs can be extracted from audio recordings. Second, although the spectrotemporal modulation feature space is generally complex and difficult to unpack, our findings are consistent with previous acoustic characterizations of whistled speech [5, 7], making it easier to interpret in the context of automatic speech processing. Alternatively, MFCCs provide a broadband description of the spectral information encoded in audio recordings, making them useful for automatic speech recognition tasks, i.e., speaker or speech emotion recognition, but not necessarily when connecting model efficiency with the acoustic information used for training.

Although our findings suggest spectrotemporal modulation metrics are sensitive to differences between modal and whistled speech, the evaluation of automatic speech modality classification performance could be improved in a number of ways. First we relied on Support Vector Machines, when other models, e.g., Decision Trees, Random Forests, may yield different results. Second, given how uncommon whistled speech is and volume of available recordings, we were unable to validate our models on a large dataset. A dataset of diverse speakers performing both modal and whistled speech is required. Of particular interest would be to obtain a dataset where each speaker performs the same work, phrase, or sentence in both modal and whistled speech. Finally it is difficult to evaluate the robustness of our method on another or mixed-language dataset, however, we hypothesize that similar STM-feature regions might emerge, given their acoustic characteristics match those previously reported in whistled speech [5]. Our future work aims to create such a dataset in order to perform not only phonetic analysis, but also examine performance of different automatic speech processes, e.g., speaker and language recognition.

6 Conclusion

This study investigated whether spectrotemporal modulation metrics were sensitive to distinguishing modal and whistled speech. The following lists take-away messages from the study: (1) statistical analysis was used to identify spectrotemporal modulation features that distinguished modal and whistled speech; (2) these features formed two separate regions and significant differences between speech modalities were observed; and (3) STM-feature and MFCC-feature models both exhibited high automatic speech modality classification performance. The findings support previously reported acoustic characterizations of whistled speech and suggest spectrotemporal modulation features are useful for distinguishing whistled and modal speech in automatic speech processing applications.

Funding. The authors did not receive support from any organization for the submitted work.

References

[1] Mohammed, M., Abdulkareem, K., Mostafa, S., Maashi, M., Zapirain, B., Oleagordia, I., Al-Dhief, F., Alhakami, H.: Voice pathology detection and classification using convolutional neural network model. Applied Sciences **10**, 3723 (2020). https://doi.org/10.3390/app10113723

[2] Alhussein, M., Muhammad, G.: Voice pathology detection using deep learning on mobile healthcare framework. IEEE Access **6**, 41034–41041 (2018). https://doi.org/10.1109/ACCESS.2018.2856238

[3] Vestman, V., Gowda, D., Sahidullah, M., Alku, P., Kinnunen, T.: Speaker recognition from whispered speech: A tutorial survey and an application of time-varying linear prediction. Speech Communication **99**, 62–79 (2018). https://doi.org/10.1016/j.specom.2018.02.009

[4] Pohjalainen, J., Raitio, T., Yrttiaho, S., Alku, P.: Detection of shouted speech in noise: Human and machine. The Journal of the Acoustical Society of America **133**(4), 2377–2389 (2013). https://doi.org/10.1121/1.4794394

[5] Meyer, J.: Whistled Languages: A Worldwide Inquiry on Human Whistled Speech. Springer (2015). https://doi.org/10.1007/978-3-662-45837-2

[6] Busnel, R.-G.: Recherches experimentales sur la langue sifflee de kuskoy (experimental research on the whistling language of kuskoy). Revue de phonétique appliquée (1970)

[7] Meyer, J., Magnasco, M.O., Reiss, D.: The relevance of human whistled languages for the analysis and decoding of dolphin communication. Frontiers in Psychology **12** (2021). https://doi.org/10.3389/fpsyg.2021.689501

[8] Classe, A.: The whistled language of la gomera. Scientific American **196**(4), 111–120 (1957). https://doi.org/10.1038/scientificamerican0457-111

[9] Moore, B., Glasberg, B.: A revision of zwicker's loudness model. Acta Acustica united with Acustica **82**, 335–345 (1996)

[10] Jakubiak, A.: Whistle-to-text: Automatic recognition of the Silbo Gomero whistled language. In: Proc. INTERSPEECH 2023, pp. 3402–3406 (2023). https://doi.org/10.21437/Interspeech.2023-989

[11] Elliott, T.M., Theunissen, F.E.: The modulation transfer function for speech intelligibility. PLOS Computational Biology **5**(3), 1–14 (2009). https://doi.org/10.1371/journal.pcbi.1000302

[12] Flinker, A., Doyle, W., Mehta, A., Devinsky, O., Poeppel, D.: Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries. Nature Human Behaviour **3** (2019). https://doi.org/10.1038/s41562-019-0548-z

[13] Moro-Velázquez, L., Gómez-García, J., Godino Llorente, J., Andrade-Miranda, G.: Modulation spectra morphological parameters: A new method to assess voice pathologies according to the grbas scale. BioMed Research International **2015** (2015). https://doi.org/10.1155/2015/259239

[14] Marczyk, A., O'Brien, B., Tremblay, P., Woisard, V., Ghio, A.: Correlates of vowel clarity in the spectrotemporal modulation domain: Application to speech impairment evaluation. The Journal of the Acoustical Society of America **152**, 2675–2691 (2022). https://doi.org/10.1121/10.0015024

[15] Arnal, L., Flinker, A., Kleinschmidt, A., Giraud, A.-L., Poeppel, D.: Human screams occupy a privileged niche in the communication soundscape. Current biology : CB **25** (2015). https://doi.org/10.1016/j.cub.2015.06.043

[16] Mazzocconi, C., O'Brien, B., Bodur, K., Fourtassi, A.: Do children laugh like their parents? conversational laughter mimicry occurrence and acoustic alignment in middle-childhood. Journal of Nonverbal Behavior (2025). https://doi.org/10.1007/s10919-025-00478-z

[17] Ludusan, B., Wagner, P.: An Evaluation of Manual and Semi-Automatic Laughter Annotation. In: Proc. Interspeech 2020, pp. 621–625 (2020). https://doi.org/10.21437/Interspeech.2020-2521

[18] Woolley, S.M.N., Fremouw, T.E., Hsu, A., Theunissen, F.E.: Tuning for spectrotemporal modulations as a mechanism for auditory discrimination of natural sounds. Nature Neuroscience **8**(10), 1371–1379 (2005). https://doi.org/10.1038/nn1536

[19] Avila, A., Fraga, F., Sarria-Paja, M., Falk, T.: Investigating the use of modulation spectral features within an i-vector framework for far-field automatic speaker verification, pp. 1–5 (2014). https://doi.org/10.1109/ITS.2014.6948012

[20] Suthokumar, G., Sethu, V., Wijenayake, C., Ambikairajah, E.: Modulation dynamic features for the detection of replay attacks, pp. 691–695 (2018). https://doi.org/10.21437/Interspeech.2018-1846

[21] Park, S.J., Yeung, G., Vesselinova, N., Kreiman, J., Keating, P., Alwan, A.: Towards understanding speaker discrimination abilities in humans and machines for text-independent short utterances of different speech styles. The Journal of the Acoustical Society of America **144**, 375–386 (2018). https://doi.org/10.1121/1.5045323

[22] Meyer, J.: Whistled Speech as a tool for shepherds (demonstration in Whistled Spanish at the 'House of shepherds', Ecrins National Park, France). The World Whistles/Maison Du Berger/ Adaparle Project (UGA). https://vimeo.com/327303614

[23] Meyer, J., Ridouane, R.: Parler en sifflant dans les montagnes de l'Atlas : dialogue sifflé à longue distance en langue berbère tachelhit. (Speaking while whistling in the Atlas : whistled dialogs in Berber Tashlhiyt). A co-production CNRS - GIPSA-Lab/CNRS – LPP. https://vimeo.com/361557210

[24] Venezia, J.H., Hickok, G., Richards, V.M.: Auditory "bubbles": Efficient classification of the spectrotemporal modulations essential for speech intelligibility. The Journal of the Acoustical Society of America **140**(2), 1072–1088 (2016). https://doi.org/10.1121/1.4960544

[25] Singh, N.C., Theunissen, F.E.: Modulation spectra of natural sounds and ethological theories of auditory processing. The Journal of the Acoustical Society of America **114**(6), 3394–3411 (2003). https://doi.org/10.1121/1.1624067

[26] Thoret, E., Caramiaux, B., Depalle, P., Mcadams, S.: Learning metrics on spectrotemporal modulations reveals the perception of musical instrument timbre. Nature Human Behaviour **5** (2021). https://doi.org/10.1038/s41562-020-00987-5

[27] Giraud, A.-L., Poeppel, D.: Cortical oscillations and speech processing: Emerging computational principles and operations. Nature neuroscience **15**, 511–7 (2012). https://doi.org/10.1038/nn.3063