INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2024, VOL. 70, NO. 4, PP. 815-821
Manuscript received August 19, 2024; revised October 2024. doi: 10.24425/ijet.2024.152065

Decoding soundscape stimuli and their impact on ASMR studies

Tomasz Piernicki, Sahar Seifzadeh, and Bozena Kostek

***Abstract*—This paper focuses on extracting and understanding the acoustical features embedded in the soundscape used in ASMR (Autonomous Sensory Meridian Response) studies. To this aim, a dataset of the most common sound effects employed in ASMR studies is gathered, containing whispering stimuli but also sound effects such as tapping and scratching. Further, a comparative analytical survey is performed based on various acoustical features and two-dimensional representations in the form of mel spectrogram. A special interest is in whispering sounds uttered in different languages. That is why whispering sounds are compared in the language context, and the characteristics of speaking and whispering are investigated within languages. The results of the 2D analyses are shown in the form of similarity measures, such as Normalized Root Mean Squared Error (NRMSE), PSNR (peak signal-to-noise ratio), and SSIM (structural similarity index measure). The summary is produced, showing that the analytical aspect of the inherently experiential nature of ASMR is highly affected by the subjective, personal experience, so the evidence behind triggering certain brain waves cannot be unambiguous.**

***Keywords*—ASMR acoustic stimuli; soundscape; acoustic features; 2-dimensional signal representation; speech processing**

I. Introduction

BY definition, soundscapes are an integral part of creating an immersive storytelling environment. They encompass a wide range of auditory components—from ambient noises and sound effects to musical patterns—all carefully curated to enhance the narrative experience. In the context of Autonomous Sensory Meridian Response (ASMR), the idea of soundscapes translates into a unique therapeutic role. ASMR involves specific auditory stimuli, such as whispering, tapping, scratching, or other soothing sounds, designed to evoke a sensation of relaxation and well-being. These sounds can be thought of as creating a "narrative" of calm and comfort.

ASMR is characterized by the onset of tingling sensations in response to specific auditory and visual stimuli. ASMR triggers are a variety of sensory inputs commonly referred to as ASMR triggers or stimuli. As already said, whispers, tapping sounds, and personal attention gestures are examples of such triggers [1]. Hence, for the purpose of our paper, we call them 'soundscape.'

Individual susceptibility to ASMR stimuli varies significantly, with acoustic triggers evoking distinctive responses according to personal preferences. Depending on the individual, some may find comfort in soft-spoken voices, while others may find comfort in tapping and scratching sounds [2]. The diversity of ASMR experiences underscores the multifaceted nature of ASMR experiences and emphasizes the subjective nature of trigger preferences within the ASMR community [3].

A cross-correlation analysis revealed a strong relationship between estimates of ASMR and specific acoustic characteristics of auditory stimuli, such as volume, spectral centroid, and spectral width [4]. It appears that low-frequency sounds with a deeper tonal quality are more effective at inducing ASMR. Intriguingly, the peak ASMR experience occurred approximately two seconds after the change in these acoustic traits, indicating a delay consistent with the integration of multiple senses. In addition, they found that ASMR susceptibility appears to be closely related to an individual's emotional state, particularly anxiety feelings, rather than personality characteristics.

According to Pablo et al. [5], they developed a framework for processing large quantities of whispered speech data within ASMR recordings. The framework effectively separated whispered signals from other audio elements typical of ASMR content by utilizing acoustic features identified as valuable in Whispered Audio Detection (WAD). As a result of the integration of recurrent neural networks (RNNs), WAD's capabilities were enhanced, particularly in the modeling of temporal dependencies. Edyson [6], an application for semi-automatic audio data labeling, was employed to optimize processing efficiency for vast datasets. Data augmentation techniques were also used to refine a clean whisper speech detector (CWAD) specific to ASMR speech style and acoustic triggers. This approach was presented as a generalizable method applicable to similar data scenarios, promising advancements in ASMR content analysis and interpretation.

The purpose of this study is to extract and understand the acoustical features embedded in ASMR triggers. Understanding the acoustical features embedded in the 'soundscape' used in ASMR is important for several reasons.

Firstly, it allows for a deeper understanding of the mechanisms involved in ASMR experiences. Researchers can gain insight into how specific sounds interact with the brain and sensory pathways in order to evoke ASMR by dissecting the auditory components that elicit tingling sensations and relaxation responses.

A second benefit of knowing acoustical features is the ability to create and curate effective ASMR content. In order to maximize their potential to induce ASMR experiences in viewers, content creators may leverage this understanding to

Tomasz Piernicki and Sahar Seifzadeh are with the Gdansk University of Technology, Faculty of Electronics, Telecommunications and Informatics, Multimedia Systems Department, Poland (e-mail:

Bożena Kostek is with Gdansk University of Technology, Faculty of Electronics, Telecommunications and Informatics, Audio Acoustics Laboratory, Poland (e-mail: bozenka@sound.eti.pg.gda.pl, corresponding author).

© The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0, https://creativecommons.org/licenses/by/4.0/), which permits use, distribution, and reproduction in any medium, provided that the Article is properly cited.

tailor their videos. Creators can enhance the effectiveness and appeal of their ASMR content by optimizing the auditory elements, such as whispers, tapping, and scratching sounds.

In addition, understanding acoustical features can assist in the development of ASMR-related technologies and applications. Using machine learning algorithms, researchers and developers may be able to identify and classify ASMR stimuli within audiovisual content automatically. The creation of ASMR-specific search engines and recommendation systems could facilitate the discovery of personalized ASMR experiences. All the above form the objectives of our study. Therefore, the objectives of this research paper are twofold. First, by aligning soft-spoken speech with whispering, we aim to identify the commonalities and differences between whispered speech and normal speech. Second, we seek to extract and discern similarities between these speech features across different genders and languages. Additionally, we aim to extract common features between speech and other ASMR acoustic triggers.

The paper includes the method, presents the material, and describes the acoustic features employed in the study. Further on, measures for the assessment of similarities and differences between acoustic stimuli, as well as tools used in ASMR stimuli recordings, are included. This is followed by analyses of soundscapes that are typically used in ASMR studies, focusing on commonalities between acoustic features. Next, measures for image quality assessment are used to compare ASMR soundscapes. Image quality measures quantify the overall difference between the image under test and a corresponding baseline. This Section also contains a discussion of the results obtained. Finally, concluding remarks are listed.

II. METHODS

A. Material

Several notable ASMR-related datasets have been introduced for research purposes:

1) YouTube-ASMR-300K Dataset [7]

This dataset comprises approximately 300,000 10-second ASMR video clips with spatial audio sourced from YouTube. It also includes a curated subset of 30,000 clips from 30 ASMR channels featuring more spatially moving sound events. Introduced in a CVPR 2020 paper, it is available on the project's companion website.

2) A-SIREN: GAN-synthesized ASMR audio clips [8]

This dataset consists of recorded and GAN-synthesized ASMR audio clips, along with corresponding psychological survey results. It was introduced as part of a research project on ASMR audio synthesis.

3) ASMR-WS (Autonomous Sensory Meridian Response Whispered-Speech) Database [9]

This is a novel database containing 38 ASMR-related whispered speech audio clips in seven different languages (Chinese, English, French, Italian, Japanese, Korean, and Spanish). It was created to facilitate the development of ASMR-specific unvoiced language identification systems.

Despite the availability of those datasets, multiple problems were identified. Our primary objective was to compare the efficacy of whispers versus soft-spoken ASMR content across different languages. We encountered challenges in sourcing suitable content from platforms like YouTube due to the scarcity of ASMRtists who consistently produce content in multiple languages while maintaining identical themes and stimuli. For this reason, we created the dataset to meet our research needs.

B. Recording process

The sound studio recording process and setup followed strict standards to ensure accurate data acquisition. Before recording, speakers relaxed for 15 minutes to stay calm and consistent. The studio environment was controlled to reduce any outside noise. Professional-grade microphones and recording devices were used, like condenser microphones and digital recorders, which captured sound at a 48 kHz sampling rate. For all speakers and recordings, the recording environment in the sense of soundscape was the same. The microphone was set to different sensitivities for normal speech and whispering to catch all vocal nuances. This change in sensitivity led to more background noise in whispered recordings. This can be observed in the analysis of parameters like the harmonics-to-noise ratio. Despite this challenge, our studio setup ensured we got a high-quality and consistent recording [10].

C. Detailed dataset description

In our methodology, we enlisted nine participants, comprising six males and three females, to individually record both whispering and soft-spoken speech in their native language as well as in English. Each participant provided recordings in their native language and then repeated the process in English, resulting in a total of 35 soft-spoken recordings and 32 whispering recordings across the following 15 languages, i.e., Arabic, English, Spanish, Persian, French, Gujarati, Hindi, Indonesian, Italian, Polish, Portuguese, Russian, Telugu, Turkish. We utilized a sample text of ASMR and relaxation content comprising 67 words and 374 characters. Such texts are commonly employed for the purpose of inducing relaxation or promoting a sense of calmness.

This database serves as a valuable tool for exploring the nuances of ASMR triggers across languages and cultures. By examining interlanguage similarities and differences in whispered speech patterns, we aim to gain insights into the underlying mechanisms of ASMR experiences and their cultural variations.

D. Acoustic Features

By applying pure mathematic transformations and algorithms, numerical metrics are calculated. Those metrics, called acoustic features, represent various subjective characteristics of speech. The simplest example of such a metric is the average energy of the signal that corresponds to the loudness of the signal. Calculation and statistical analyses of such metrics for a rich set of ASMR examples is a way of trying to understand the nature of such signals. This is one of the approaches to understanding the commonalities and differences between whispered speech and normal speech.

Statistical methods may be applied to discern similarities between speech features across different genders and languages. Those scalar values may be compared with the results of psychological questionnaires regarding the patient's well-being. Such analyses may be crucial to understanding the impact of ASMR on the subject's state of mind.

1) Shimmer and Jitter [11]

Shimmer measures the variation in amplitude between consecutive periods of a signal, indicating the instability or roughness of the voice. It is commonly used in voice quality assessment, especially in diagnosing voice disorders. Higher shimmer values suggest increased voice instability or hoarseness.

Jitter quantifies the variation in pitch period of consecutive cycles of a waveform, indicating the irregularity in pitch or frequency modulation. It is employed in voice analysis to assess vocal fold vibratory patterns. Increased jitter values are associated with voice disorders and vocal fatigue. It may be influenced by factors such as speaking rate and phonetic content.

2) Harmonics-to-noise ratio (HNR)

HNR measures the ratio of harmonics to noise in a signal, reflecting the clarity of the voice and the level of background noise [12]. It's commonly used in voice analysis to assess voice quality and signal-to-noise ratio. Higher HNR values indicate clearer voice production, while lower values suggest increased noise or breathiness. HNR may be affected by speech intensity and microphone characteristics.

3) Spectral Tilt

Spectral tilt describes the slope of a signal's spectral envelope [13]. The axis of the oscillation of the audio waveform is not required to be flat. For example, a trumpet signal registered by a microphone has a positive spectral tilt. This is a reflection of a strong airflow present in the acoustic wave. It indicates the distribution of energy across frequencies and is often related to the perceived brightness or warmth of the sound. Spectral tilt is used in speech and audio processing to characterize a sound's timbre. A flat spectral tilt indicates a balanced energy distribution, while a tilted spectrum may suggest variations in voice quality.

4) Zero-crossing rate

The zero-crossing rate counts the number of times a signal crosses the zero amplitude line, providing information about the rate of changes in the waveform. This information is useful for detecting fricatives and silence [11]. It is also used for feature extraction and segmentation. Higher zero-crossing rates indicate more rapid changes in the signal, such as in speech fricatives, while lower rates suggest longer periods of silence. The zero-crossing rate is sensitive to noise.

5) Formant Dispersion

Formant dispersion measures the spacing between harmonic frequencies present in the speech signal. Their presence and density are determined by the characteristics of the speaker's vocal tract. It is useful for identifying different phonemes and speech sounds [14],[15]. It can also provide insights into articulatory differences between speakers or speech sounds. Formant dispersion is sensitive to speech rate and dialectical variations.

6) Spectrogram

Spectrograms are visual representations of the frequency and amplitude components of a sound signal over time. They display how the intensity of different frequencies changes over time, providing insights into the characteristics of the sound. Spectrograms are being constructed from a waveform using a fast Fourier transform (FFT) algorithm. As an outcome, a 2D representation of the registered sound is obtained. The x-axis corresponds to the time, while the energy in a particular frequency band is shown on the y-axis. The value of the point (or intensity of the pixel) is the energy level at a given frequency and time.

7) Mel spectrogram

A mel spectrogram is a spectrogram with frequency bands scaled according to the human perception of pitch. It provides a more perceptually relevant representation of a signal's frequency content. Mel spectrograms are commonly used for feature extraction and analysis in speech and audio processing. They capture important spectral characteristics of speech and music.

E. Similarity Measures

1) Normalized Root Mean Squared Error (NRMSE)

RMSE quantifies the average difference between the corresponding values of two signals, indicating the overall deviation between them [16]. It's widely used in signal processing and regression analysis. Lower RMSE values indicate better agreement between signals but may not fully capture perceptual differences.

2) Peak Signal to Noise Ratio (PSNR)

PSNR measures the ratio between the maximum possible power of a signal and the power of noise present in the signal, providing a measure of signal fidelity [17]. Higher PSNR values indicate better signal quality but may not correlate perfectly with perceived quality.

3) Structural similarity index measure (SSIM)

SSIM assesses the structural similarity between two signals, capturing both pixel-wise and structural differences and providing a perceptually relevant measure of image similarity [16],[18]. Higher SSIM values indicate better structural similarity. It does not capture all aspects of perceptual quality and is sensitive to image content and distortion types.

F. Tools

To facilitate the analysis, we developed a Python toolbox capable of extracting acoustic features in bulk from files in widely used audio formats such as *.wav and *.mp3. We employed libraries like librosa [19] for audio manipulation, feature extraction, and signal processing, enabling tasks like spectrogram generation and mel-frequency cepstral coefficients (MFCC) extraction. Additionally, we utilized python-praat software to extract advanced phonetic features, particularly for speech analysis [20],[21]. To enhance visualization and gain insights into the underlying characteristics of sound, we utilized matplotlib. Similarity measures were computed using the scikit-image python package [22]. This toolbox allows for bulk analyses from directories based on folders and a metadata.csv file. Any type of metadata in tabular format is supported. The toolbox is available on the MIT license on GitHub.

III. Analyses and Results

This analysis involved analyzing speech signals. Both time-domain and 2D time-frequency domain features were explored.

A. Gender-centric acoustic features analysis

Various acoustic features that are characterized by an interesting behavior are depicted in Fig. 1. Except for the spectral tilt, all features show significant differences when normal speech and whispering are compared. Shimmer and Jitter, depicted respectively in Fig. 1a and Fig. 1b, show an increase in whispering when compared to normal speech. These qualities are perceived as roughness, breathiness, or hoarseness in a speaker's voice. Every natural speech includes some degree of jitter and shimmer, but quantifying them is a common method for identifying voice disorders. Personal behaviors like smoking or alcohol consumption may elevate the levels of jitter and shimmer in the voice. This is intuitive for both described features based on the general behavior of jitter and shimmer for normal speech. This may imply that whispered speech may be treated as a dysfunctional speech. This suggests that the measures and methods applied to dysfunctional speech can be applied with success to the shepherd speech signal.

The average harmonics-to-noise ratio depicted in Fig. 1c for whispering differs both between the whisper and normal speech and between the genders. In general, the average harmonics-to-noise ratio for males is lower than for females by approximately 2.5 dB, which is a perceptually significant value. The acoustic environment for each recording session was identical. Higher HNR for female speakers when compared to men shows that female voice signals carry more harmonic (modal) information than men. The difference in HNR between the whisper recordings and normal speech recordings can be attributed to the microphone sensitivity settings, as any minor disturbance in the acoustic pressure of the recording room, the soundscape, may be reflected in the spoken measure.

In Fig. 1d, the spectral tilt for both males and females is negative. This is an expected behavior as the speakers were asked to relax before recording both normal speech and whisper. High spectral tilt is often assigned to high emotional states like joy or anger. None of the speakers were in the states described in the experiment.

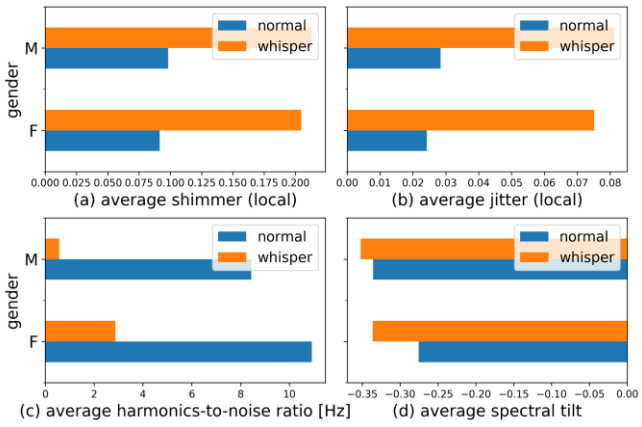

Fig. 1. Averaged acoustic features for normal speech and whispering in relation to the speaker's gender. a) shimmer; b) jitter (local); c) Harmonics-to-noise ratio; d) spectral tilt

B. Language-centric acoustic features analysis

In-depth analyses of the whispering speech acoustic features for individual languages were performed. In Fig. 2a, the spectral tilt of the voice signal is shown. The relation of spectral tilt between normal speech and whispered speech is not consistent between languages. While the harmonics-to-noise ratio, depicted in Fig. 2b, is significantly lower for all of the analyzed languages, no such behavior can be observed for the spectral tilt. HNR is sourced from both the quality and "musicality" of the speaker's voice and the recording environment. In contrast, spectral tilt correlates with changes in the speaker's sound pressure level. This underlines the importance of the recording methodology selection.

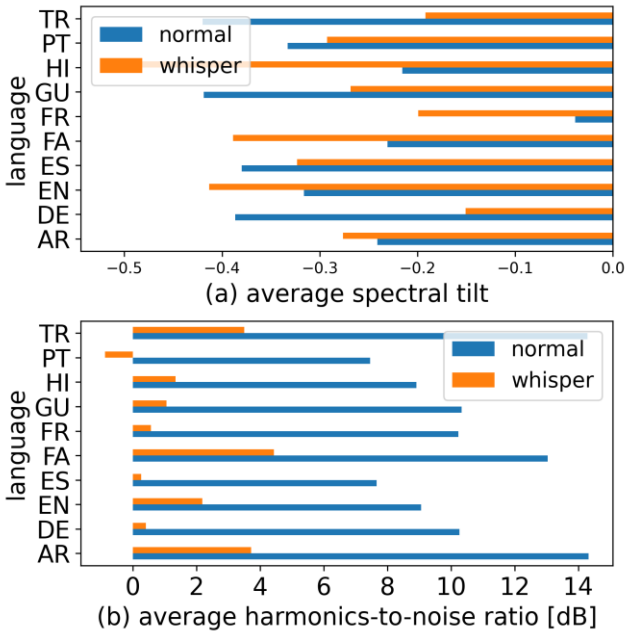

Fig. 2. a) Average spectral tilt and b) average harmonics-to-noise ratio acoustic features with respect to speaker's language (Turkish (TR), Portuguese (PT), Hindi (HI), Gujarati (GU), French (FR), Persian-Farsi (FA), Spanish (ES), English (EN), German (DE), Arabic (AR))

The negative spectral tilt can be attributed to the soft nature of spoken language. Notably, there is no consistency between the ratio of whispered speech and normal speech within language groups. As an example, German and English both belong to the Indo-European Germanic group. For German, firstly, the spectral tilt for normal speech is lower than that for whispering. Secondly, the delta between German whispering and normal speech is significantly higher when compared to the English language. Similar behavior may be observed for other languages from the Indo-European family. For example, Portuguese, French, and Spanish are all Romance languages. While Portuguese and Spanish share some similarities in terms of the average spectral tilt value, the value of spectral tilt for French is, in general, more zero-centric. Moreover, the relation between the spectral tilt of normal speech and whisper speech for French has properties opposite to those of spoken features for Spanish and Portuguese. Again, a lack of consistency in spectral tilt acoustic feature is observed in the Hindi and Gujarati languages. The fact that two of those languages were recorded

by the same speaker in identical acoustic soundscapes really puts some light on the fact that whispered speech can be characterized by changes in the properties of the acoustic information carried by the recorded signal. Those characteristics vary not only when the gender of the speaker is considered but also with the differentiation of the speaker's language.

An average delta of approximately 8 dB is observed for each language when considering the HNR ratio. Portuguese is the only language with observed negative HNR for the whisper speech signal. Romance languages from the Indo-European family are all characterized by the value of HNR around zero.

Fig. 3 depicts deltas of four acoustic features. They help to understand the difference between normal speech and whispered speech normalized by "normal speech values." A negative delta value means that the given acoustic feature's observed value for normal speech was smaller than the corresponding value for whispering speech. Normalization is applied to prevent the features' values from exploding.

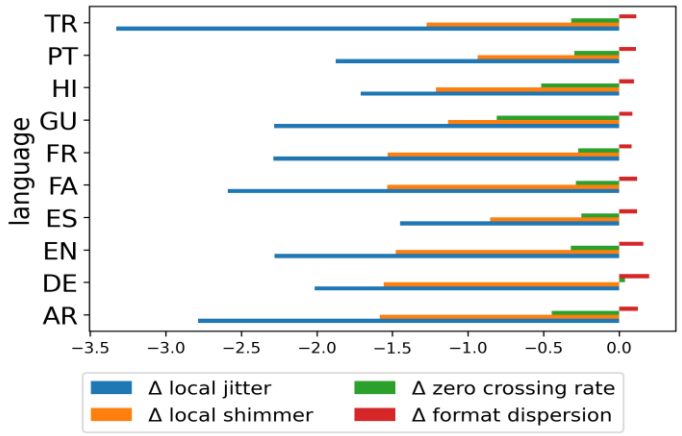

Fig. 3. Differences in chosen acoustic features normalized by the values for normal speech features

Maximum local jitter differences were observed for the Turkish and Arabic languages. In general, the normalized value for local jitter oscillates around -2.0. This difference is not language or language-family-dependent. For example, the local jitter for French, Spanish, and Portuguese, languages in the same Indo-European language family, is dispersed from the upper to the minimum part of the distribution.

As depicted in Fig. 3, the change of the local shimmer shared similar properties to the jitter described in the previous paragraph. The values for Shimmer have a smaller standard deviation, focused around the normalized delta value of -1. Jitter and shimmer can both be used to measure the quality of the speaker's voice. Jitter and Shimmer are considered to resemble the condition of a speaker's vocal cords, with higher values mapped to the "unhealthy" cognition of those. This analysis suggests that this pair of features may be used to distinguish the whispered speech from normal speech.

Zero crossing rate may be used to identify the voiced and unvoiced parts of the speech signal. In speech analysis, the distinction between voiced and unvoiced speech is significant. Voiced speech sounds, such as vowels, have a low zero crossing rate due to the regular vibration of the vocal cords, resulting in a smooth waveform with fewer zero crossings. In contrast, unvoiced speech sounds, like fricatives and plosives, exhibit a higher zero crossing rate. This is because unvoiced sounds are characterized by turbulent airflow, causing rapid signal amplitude changes and leading to more frequent zero crossings. Negative zero crossing rate data for languages from all languages except for German suggest that, in general, the whispered speech contains more voiced parts. Higher saturation of voiced parts in the speech signal may be one of the causes of the relaxing effect that ASMR speech has on the listener.

Formant dispersion in speech acoustics refers to how the frequencies of formants change relative to each other across different speech sounds. Higher formant dispersion indicates greater variation in formant frequencies, often reflecting tongue position and articulation differences. This variation helps distinguish between different vowels and consonants, contributing to the perception of speech sounds. All analyzed languages share a similar normalized delta formant frequency behavior, where the feature value for noise is higher when compared to whisper. This implies that for all analyzed languages, the speaker's tongue has less positional whispering variance compared to normal speech. This may result in a softer, more static articulation.

C. 2D time-frequency representations of sound comparison

For the purpose of this analysis, metrics typically used for image comparison are used to identify differences between whispering and normal speech time-frequency representations. These representations, like MFCC, spectrograms, or mel spectrograms, may be treated as images, with time and frequency being treated as image dimensions and the power and phase components as color channels.

To ensure fair and correct comparison, corresponding speech recordings were aligned. Dynamic Time Warping (DTW) was used to determine each speaker's alignment vector between whispering and normal speech recordings. As every speaker recorded the same passage, this alignment approach was considered valid. Next, based on the alignment vector, time stretching was applied. Depending on the vector value, the original signal was either locally stretched or compressed. An example of the described operation is shown in Fig. 4.

In this study, we employ three key similarity metrics—Mean Structural Similarity Index (MSSIM), Normalized Root Mean Squared Error (NRMSE), and Peak Signal-to-Noise Ratio (PSNR)—to compare mel spectrograms of normal and whispered speech. Mel spectrograms serve as visual representations of the frequency content of speech signals over time. Differences between various speech modes can be observed using these representations. A summary of averaged distance measures for normal (reference) and whispered (test) speech is presented in Fig. 5.

Whispered speech, characterized by reduced voicing and increased aspiration, presents distinct acoustic features compared to normal speech. By utilizing MSSIM, we assess the structural similarity between the corresponding spectrograms. Low MSSIM indicates that there is a substantial structural difference between modes of speech for all analyzed languages. NRMSE allows us to quantify the

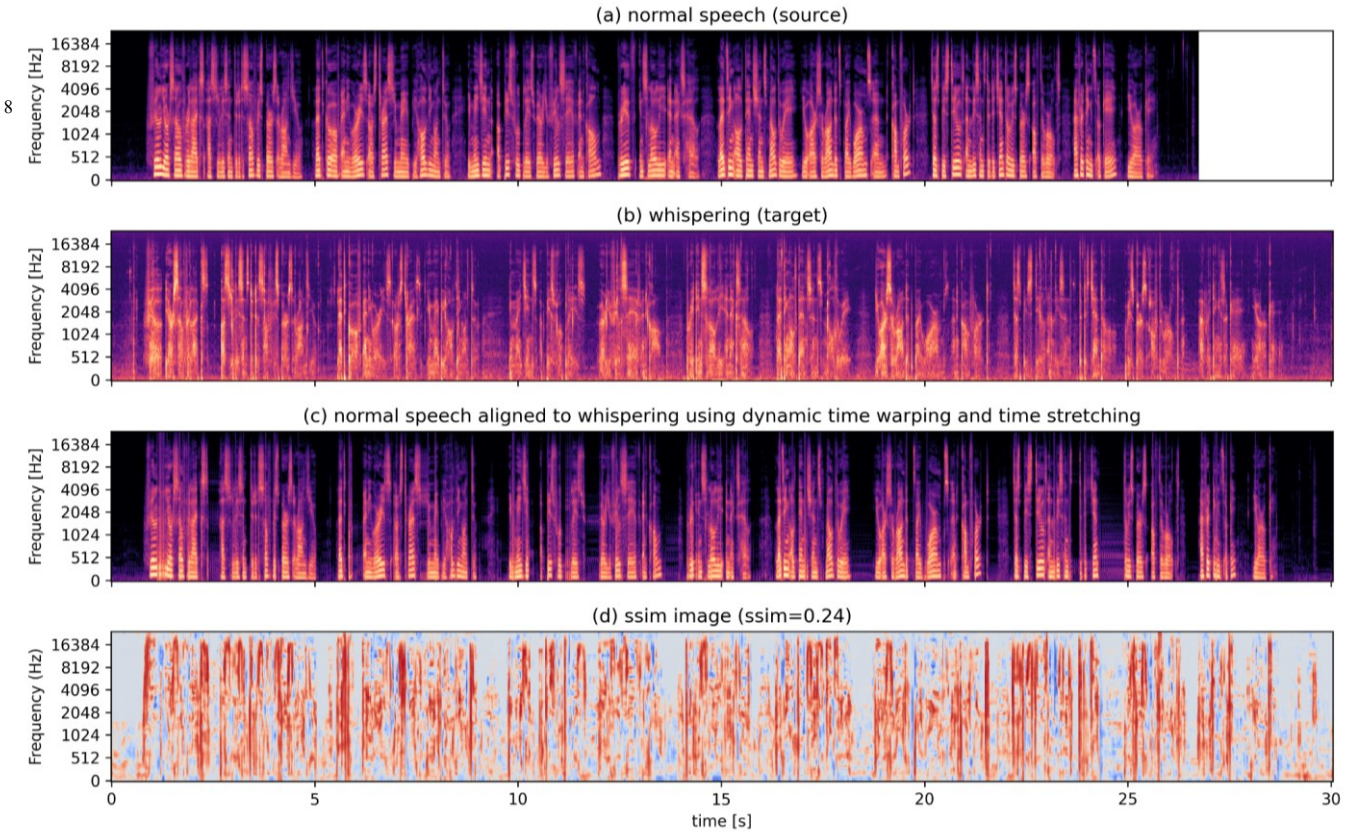

Fig. 4. Comparison of whispered and normal speech of a male speaker in the form of mel spectrograms: a) mel spectrogram of normal speech; b) mel spectrogram of whispered speech; c) mel spectrogram of time stretched normal speech aligned to whispering; d) SSIM image

average difference in pixel values between the spectrograms. Based on the observed values for NRMSE, the distance between individual pixel values is significant for all. NRMSE, in comparison to MSSIM, has greater interlanguage variance. For the Arabic language, the "pixel-wise" similarity is highest compared to other languages while maintaining the same structural similarity on a level comparable to other languages. PSNR is a measure of the signal-to-noise ratio, helping evaluate the quality of spectrograms by considering both the presence of signal and the level of noise. In the task of image compression, PSNR below 20 dB is considered to indicate a low-quality image. With PSNR being below 20 dB for all analyzed audio samples, whispered speech can be considered a poorly compressed audio signal.

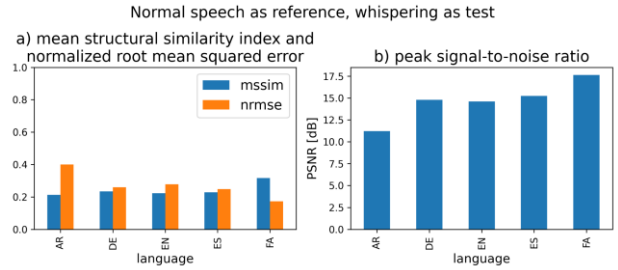

Fig. 5. Averaged distance metrics between normal (reference) and whispered (test) speech where a) shows MSSIM and NRMSE and b) shows PSNR with respect to speaker's language (Arabic (AR), German (DE), English (EN), Spanish (ES), Persian-Farsi (FA))

IV. Concluding Remarks

The techniques used to analyze and modify dysfunctional speech signals proved to be effective when applied to whispered speech. This suggests that the underlying methods for handling speech signals may be broadly applicable, regardless of the vocal effort or style. Moreover, the observation that the same speaker recorded multiple languages under identical environmental conditions highlights the robustness of whispered speech characteristics. This consistency supports the reliability of studies focusing on whispered speech and ensures that findings are attributable to the speech signal itself rather than external variables.

In the study, whispered speech properties change not only with the speaker's gender but also with the language spoken. This implies that both intrinsic (e.g., anatomical features of the speaker) and extrinsic (e.g., linguistic characteristics of the language) factors influence the acoustic properties of whispered speech.

Given the variations in acoustic properties based on gender and language, further research might be needed to fully understand the mechanisms behind these variations and explore how these findings could be generalized across different populations and conditions. In particular, subjective tests should be undertaken to assess individual reactions to various ASMR triggers. These tests can provide insights into personal preferences and sensitivities, helping to elucidate how different people experience ASMR. By incorporating subjective measures, researchers can better capture the nuances of ASMR responses and develop a more comprehensive understanding of how acoustic properties influence the ASMR experience across diverse demographic groups. Such studies could ultimately inform the creation of more effective and universally appealing ASMR content.

However, at this stage, these findings may have the potential to create advanced and, to some extent, standardized diagnostic tools. This nuanced understanding may have a significant impact on ASMRtists who create content aimed at reducing anxiety. Recognizing the importance of subjective experience in ASMR responses highlights the need for ASMRtists to consider individual preferences and sensitivities when producing their material. Future studies incorporating both subjective and objective analyses can

provide valuable guidelines for ASMRtists, enabling them to create more personalized and effective content for their audience. Ultimately, this approach could lead to better outcomes for individuals seeking anxiety relief through ASMR. For example, the ability to characterize whispered speech by changes in its acoustic properties opens the door for developing more refined diagnostic tools and therapeutic approaches for speech disorders, which could be tailored to individual characteristics such as gender and language.

Finally, we should also address the limitations of our study, in which we focused on the acoustic properties of whispering sounds in different languages and the characteristics of speaking and whispering within these languages. As already said, objective analyses performed confirmed that the ASMR experience is highly influenced by subjective, personal factors. Although no subjective tests were conducted, the results underscore the inherently experiential nature of ASMR, suggesting that individual differences play a significant role in how certain sounds trigger specific brain waves. However, future research should incorporate subjective assessments to complement the objective findings, providing a more holistic understanding of the ASMR phenomenon and its variability across different populations and contexts. Such comprehensive studies will help to elucidate the intricate mechanisms behind ASMR and guide the development of more universally effective ASMR content.

REFERENCES

[1] E. L. Barratt, C. Spence, and N. J. Davis, "Sensory determinants of the autonomous sensory meridian response (ASMR): Understanding the triggers," PeerJ, vol. 5, e3846, 2017. https://doi.org/10.7717/peerj.3846

[2] E. L. Barratt, and N. J. Davis, "Autonomous Sensory Meridian Response (ASMR): A flow-like mental state," PeerJ, vol. 3, e851, 2015. https://doi.org/10.7717/peerj.851

[3] G. L. Poerio, E. Blakey, T. J. Hostler, and T. Veltri, "More than a feeling: Autonomous sensory meridian response (ASMR) is characterized by reliable changes in affect and physiology," PLOS ONE, vol. 13(6), e0196645, 2018. https://doi.org/10.1371/journal.pone.0196645

[4] T. Koumura, M. Nakatani, H.-I. Liao, and H. M. Kondo, "Deep, soft, and dark sounds induce autonomous sensory meridian response," 2019. https://doi.org/10.1101/2019.12.28.889907

[5] P. P. Zarazaga, G. Eje Henter, and Z. Malisz, "A Processing Framework to Access Large Quantities of Whispered Speech Found in ASMR," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5, 2023. https://doi.org/10.1109/ICASSP49357.2023.10095965

[6] P. Fallgren, Z. Malisz, and J. Edlund, "How to Annotate 100 Hours in 45 Minutes," Interspeech 2019, pp. 341–345, 2019. https://doi.org/10.21437/Interspeech.2019-1648

[7] K. Yang, B. Russell, and J. Salamon, "Telling left from right: Learning spatial correspondence of sight and sound," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9932-9941, 2020.

[8] Bin Han, "A-SIREN: GAN-synthesized ASMR audio clips," IEEE Dataport, November 1, 2022. https://doi.org/10.21227/d5zd-6n82

[9] M. Song, Z. Yang, E. Parada-Cabaleiro, X. Jing, Y. Yamamoto, and B. Schuller, "Identifying languages in a novel dataset: ASMR-whispered speech," Frontiers in Neuroscience, vol. 17, 1120311, 2023. https://doi.org/10.3389/fnins.2023.1120311

[10] F. A. Everest, and K. C. Pohlmann, "Master Handbook of Acoustics," Sixth Edition, McGraw-Hill Education, 2015.

[11] T. Bäckström, O. Räsänen, A. Zewoudie, P. Pérez Zarazaga, L. Koivusalo, S. Das, E. Gómez Mellado, M. Bouafif Mansali, D. Ramos, S. Kadiri, and P. Alku, "Introduction to Speech Processing," 2nd Edition, 2022. https://speechprocessingbook.aalto.fi, https://doi.org/10.5281/zenodo.6821775.

[12] J. Fernandes, F. Teixeira, V. Guedes, A. Junior, J.P. Teixeira, "Harmonic to Noise Ratio Measurement - Selection of Window and Length," Procedia Computer Science, Volume 138, Pages 280-285, ISSN 1877-0509, 2018. https://doi.org/10.1016/j.procs.2018.10.040.

[13] Y. Lu, and M. Cooke, "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise," Speech Communication, vol. 51(12), pp. 1253-1262, 2009.

[14] S. Sapir, L.O. Ramig, J.L. Spielman, and C. Fox, "Formant centralization ratio: a proposal for a new acoustic measure of dysarthric speech," J Speech Lang Hear Res, vol. 53(1), pp. 114-25, Feb. 2010. https://doi.org/10.1044/1092-4388(2009/08-0184). Epub 2009 Nov 30. PMID: 19948755; PMCID: PMC2821466.

[15] A. De Cheveigné, and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," The Journal of the Acoustical Society of America, vol. 111(4), pp. 1917-1930, 2002.

[16] Z. Wang, and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," IEEE Signal Processing Magazine, vol. 26(1), pp. 98-117, 2009.

[17] J. Korhonen, and J. You, "Peak signal-to-noise ratio revisited: Is simple beautiful?," in 2012 Fourth International Workshop on Quality of Multimedia Experience, pp. 37-38, July 2012. IEEE.

[18] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13(4), pp. 600-612, 2004.

[19] B. McFee, C. Raffel, D. Liang, D.P.W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in Proceedings of the 14th Python in Science Conference, pp. 18-25, 2015. https://librosa.org/doc/latest/index.html

[20] Y. Jadoul, "Parselmouth – Praat in Python, the Pythonic way," 2022. https://parselmouth.readthedocs.io/en/stable/

[21] N.H. De Jong, J. Pacilly, and W. Heeren, "PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically," Assessment in Education: Principles, Policy and Practice, vol. 28(4), pp. 456-476, 2021. https://doi.org/10.1080/0969594X.2021.1951162

[22] S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, & T. Yu, "scikit-image: image processing in Python," PeerJ, vol. 2, e453, 2014.