Contrastive Learning Approach for Assessment of Phonological Precision in Patients with Tongue Cancer Using MRI Data

Tomás Arias-Vergara[1,2], *Paula Andrea Pérez-Toro*[1,2], *Xiaofeng Liu*[3], *Fangxu Xing*[2], *Maureen Stone*[4], *Jiachen Zhuo*[4], *Jerry L. Prince*[5], *Maria Schuster*[6], *Elmar Nöth*[1], *Jonghye Woo*[2], *Andreas Maier*[1]

[1]Pattern Recognition Lab. Friedrich-Alexander University, Erlangen, Germany,
[2]Massachusetts General Hospital - Harvard Medical School, Boston, MA, USA,
[3]Department of Radiology and Biomedical Imaging, Yale University, New Haven, CT, USA
[4]University of Maryland, Baltimore, MD, USA,
[5]Johns Hopkins University, Baltimore, MD, USA,
[6] Department of Otorhinolaryngology, Ludwig-Maximilians University, Munich, Germany

tomas.arias@fau.de

Abstract

Magnetic Resonance Imaging (MRI) allows analyzing speech production by capturing high-resolution images of the dynamic processes in the vocal tract. In clinical applications, combining MRI with synchronized speech recordings leads to improved patient outcomes, especially if a phonological-based approach is used for assessment. However, when audio signals are unavailable, the recognition accuracy of sounds is decreased when using only MRI data. We propose a contrastive learning approach to improve the detection of phonological classes from MRI data when acoustic signals are not available at inference time. We demonstrate that frame-wise recognition of phonological classes improves from an f1 of 0.74 to 0.85 when the contrastive loss approach is implemented. Furthermore, we show the utility of our approach in the clinical application of using such phonological classes to assess speech disorders in patients with tongue cancer, yielding promising results in the recognition task.

Index Terms: speech processing, magnetic resonance imaging, contrastive learning, pathological speech processing, phonological analysis, tongue cancer, deep learning.

1. Introduction

The analysis of human speech with Magnetic Resonance Imaging (MRI) provides essential information on the dynamic processes involved in speech production, allowing unobtrusive monitoring of the complete vocal tract during speech production. In clinical applications, personalized monitoring and increased speed of speech rehabilitation can be achieved through targeted phonological therapy i.e., by breaking down spoken words into their linguistic units [1, 2]. For example, by looking at the position of the speech articulators during spoken language production. For example, Figure 1 shows two MRI video frames displaying the tip and body of the tongue changing the position when going from one sound to another.

In the case of tongue cancer patients, speech articulation often faces significant challenges due to the physical and functional changes caused by the tumor and its treatment. The tongue's mobility, strength, and coordination can be severely affected, leading to difficulties in producing clear speech sounds. Articulation problems may manifest as slurred or unclear speech, with particular trouble in producing consonants that require precise tongue movements [3, 4].

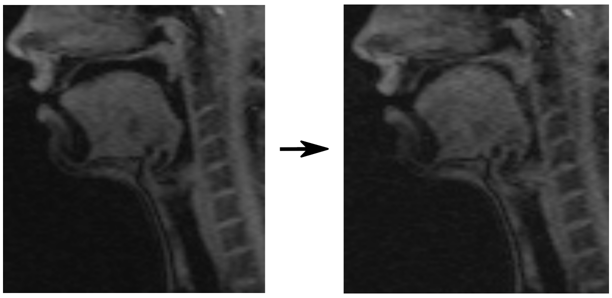

Figure 1: *Example of two MRI video frames displaying the tip and body of the tongue changing the position when going from one sound to another.*

Surgical interventions for tongue cancer involve removing the tumor along with nearby tissues through a procedure called glossectomy. The extent of the glossectomy is determined by the tumor's size and involves various adjacent structures within the tongue. Typically, when a substantial amount of tissue is removed, reconstructive surgery is necessary. In this reconstruction phase, a flap may be transplanted onto the residual tongue tissue to restore the tongue's original volume and shape, aiming to match the patient's pre-surgical appearance as closely as possible [5].

While MRI provides detailed visualization of the anatomical structures involved in speech production, it lacks high temporal resolution. In addition, the acoustic information provided by audio signals helps to identify certain sounds that are otherwise more difficult to recognize from the MRI data alone due to the surrounding structures during coarticulation of sounds [6]. However, having access to *reliable* acoustic information is not always possible, mainly because the high noise levels produced by the machine result in poor speech recording quality; thus, a specialized non-magnetic microphone (e.g., fiber-optic) is necessary for simultaneous speech-MRI recording or the signal must go through a denoising process. In this paper, we propose the use of contrastive learning for the automatic detection of vowels and consonants grouped into nine phonological classes from MRI data, aiming to analyze speech disorders in TC patients. Further details about the proposed approach and main contributions are provided in the following sections.

10.21437/Interspeech.2024-2236

1.1. Related work

To the best of our knowledge, there are no prior works in the literature that have focused on evaluating speech disorders using a phonologically targeted approach with MRI data. However, there are works in the literature that have considered phoneme-level automatic speech recognition from MRI data. For example, Pandey et al. [7] proposed to use a combination of 3D convolutional layers, bidirectional recurrent networks with gated recurrent units, and connectionist temporal classification to generate text from articulatory motions captured from MRI data, achieving a phoneme error rate of 40.6% at sentence-level. Van Leeuwen et al. [8] proposed a deep learning model to classify 27 different phonemes using midsagittal MRI of the vocal tract. In that work, a convolutional neural network (CNN) was trained to classify vowels (13 classes), consonants (14 classes), and all phonemes (27 classes) across 17 subjects, yielding accuracies of up to 57 % (top-1 accuracy). Saha et al. [9] proposed using Long-term Recurrent Convolutional Networks models, to identify different VCV sequences from dynamic shaping of the vocal tract, where an accuracy of 42% was reported in the prediction of 51 different VCV combinations.

1.2. Contributions of this work

Although MRI has been used for clinical applications previously and different studies have addressed the automatic phoneme recognition from MRI data, there are no papers addressing a phonological-inspired approach to MRI to analyze speech disorders. The approach consists of computing phonological posterior probabilities from MRI frames, considering vowels and consonants grouped according to the place of articulation. Although this has been used previously in speech signals, there are no works in the literature addressing the same approach using medical imaging techniques. Furthermore, we propose to use a contrastive learning-based approach during training to improve the performance of the phonological class recognition from MRI, when synchronized acoustic speech recordings are available.

2. Datasets

In this work, we considered two MRI datasets to perform our experiments. The first dataset is the *USC 75-Speaker Speech MRI*, which we used to train our model for frame-wise phonological class recognition. The second dataset is an in-house dataset (*cine MRI Tongue Cancer Data*), which comprises cine MRI sequences of TC patients and healthy controls [10]. We used this dataset to demonstrate the clinical application of our proposed approach. Both datasets include MRI videos and speech recordings, however, the speech recordings of the TC dataset are heavily affected by noise; thus, they are not used in this work. The image sequences were re-sampled at 26 frames per second while the speech recordings were at 16 kHz.

2.1. USC 75-Speaker Speech MRI

This is an open-source dataset [11] containing 2D sagittal-view real-time MRI videos and synchronized speech recordings of 75 subjects performing 21 speech tasks. The data was collected using a commercial 1.5T MRI scanner with a custom 8-channel upper airway receiver coil array, with four elements on each side of the subject's cheeks for signal reception.

2.2. Cine MRI Tongue Cancer Data

This is an in-house dataset consisting of 23 controls and 15 TC patients [12]. The dataset includes speech recordings and cine MRI sequences captured while the participants were asked to speak the words "a geese" and "a souk". The data were acquired using a Siemens 3.0T TIM Trio system, with a 12-channel head coil and a 4-channel neck coil utilized for a segmented gradient echo sequence.

2.3. Phonological Groups

Instead of classifying phonemes individually, we opted to group vowels and consonants according to the place of articulation. For this purpose, we used the stimuli related to the VCV triplets [11], as shown in Table 1, which summarizes the phonological classes considered for automatic recognition and the corresponding stimuli. The words Front, Central, and Back, refer to the vowel that is produced in the corresponding VCV triplet. The frame-wise phonological labels were obtained using a forced alignment system on the speech recordings of the first dataset. Then, an estimation of the frames of interest was initially performed, and we corrected the generated labels manually when necessary. The tool used for creating the initial labels was the BAS CLARIN web service [13] which provides a forced alignment tool in a variety of languages. The acoustic speech recordings (from the USC 75 dataset) are uploaded with their corresponding orthographic transcription to obtain the time stamps of the phonemes represented in SAMPA format.

Table 1: *Phonological classes considered in this study.*

Class	Stimuli
Labial Front	ipi, ibi, imi
Labial Central	apa, aba, ama
Labial Back	upu, ubu, umu
Alveolar Front	iti, idi, ini, isi, ishi
Alveolar Central	ata, ada, ana, asa, eese, asha
Alveolar Back	utu, udu, unu, usu, ushu
Velar Front	iki, igi
Velar Central	aka, aga, agee
Velar Back	uku, ugu

3. Methods

Figure 2 summarizes the training procedure of the proposed contrastive learning approach for the automatic prediction of phonological classes from MRI data. During training, a single video frame and its corresponding speech signal are processed by two different encoders. The image encoder is fine-tuned, while the parameters of the speech encoder remain unchanged. Before passing the MRI feature embedding for classification, we maximize the similarity between the image and speech embeddings. For this, we must first project the MRI embedding into the same dimensions as the audio encoder. Then, a linear multilayer perceptron with a softmax activation function is used for the classification. Besides the contrastive loss, we also compute the cross-entropy loss and add it together to improve accuracy. During inference, only the MRI is used to predict the phonological classes.

Table 2: *Performance of the phonological class recognizer for the different groups.* ***Base ViT:*** *Baseline ViT model without fine-tuning.* ***Fine-tuned ViT:*** *Fine-tuned ViT model with USC MRI data.* ***Contrastive ViT:*** *Fine-tuned ViT model and frozen Wav2Vec model.*

	Base ViT			**Fine-tuned ViT**			**Contrastive ViT**		
Class	**Precision**	**Recall**	**f1-score**	**Precision**	**Recall**	**f1-score**	**Precision**	**Recall**	**f1-score**
Labial Front	0.82	0.38	0.52	1.00	0.67	0.80	0.96	0.88	0.92
Labial Central	0.69	0.67	0.68	0.81	0.84	0.83	0.91	1.00	0.95
Labial Back	0.42	0.28	0.34	0.63	0.54	0.58	0.78	0.78	0.78
Alveolar Front	0.41	0.19	0.26	0.70	0.56	0.62	0.72	0.74	0.73
Alveolar Central	0.71	0.46	0.56	0.94	0.92	0.93	0.92	0.94	0.93
Alveolar Back	0.18	0.66	0.28	0.69	0.80	0.74	0.98	0.78	0.87
Velar Front	0.82	0.09	0.16	0.68	0.70	0.69	0.78	0.77	0.77
Velar Central	0.40	1.00	0.57	0.69	0.94	0.80	0.89	0.97	0.93
Velar Back	0.00	0.00	0.00	0.70	0.72	0.71	0.81	0.83	0.82
Average	0.50	0.42	0.38	0.76	0.74	0.74	0.86	0.85	0.85

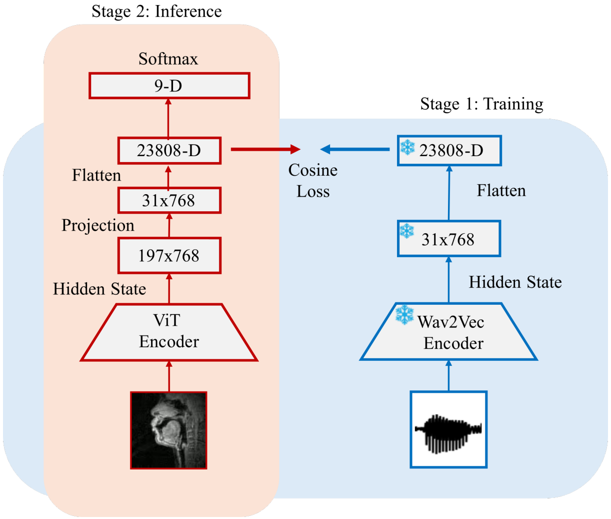

Figure 2: *Overview of our proposed network for our phonological class recognizer using a contrastive learning approach*

3.1. MRI Encoder Model

The MRI data is processed using a Visual Transformer (ViT) architecture. Particularly, we used a ViT checkpoint pre-trained on ∼14 million images and 21k classes (ImageNet-21k). During training, we fine-tune all parameters of ViT, leading to better results. The ViT model processes the input data, by splitting an MRI frame into fixed-size patches (16x16), with each patch representing a local region of the image. Then, positional encodings are added to the patch embeddings; thus, providing spatial information about the patches' locations within the MRI frame. Next, the patch embeddings, along with the positional encodings, serve as input to a transformer encoder with multiple layers of self-attention mechanisms (learning contextual information from the entire MRI image) and feed-forward neural networks. The original model has a classification head for image identification. In this work, we used the output of the last hidden state as the feature representation of the MRI data.

3.2. Speech Encoder Model

Acoustic embeddings are obtained from raw speech signals using Facebook's Wav2Vec2 base model, which was pre-trained and fine-tuned on 960 hours of Librispeech on 16kHz sampled speech audio. This model consists of three main components: feature extraction, a context network, and a linear projection to the output. Temporal convolutions are used in the feature extraction part to convert speech information S into a latent space representation $z_1, \ldots, z_T$, which, in this work, is later used for computing the contrastive loss with the MRI embedding representation from ViT. The audio segments are masked and quantized for self-supervised training, and a contextualized representation is obtained through a Transformer-based approach. We only use the last hidden state as acoustic embeddings, while **freezing all parameters** of this model i.e., no fine-tuning is performed on Wav2Vec2.

3.3. Inference & Phonological Features

During inference, only the MRI images are necessary to predict the phonological class of the processed frame. The predictions are based on posterior probabilities obtained from the softmax activation function. These probabilities have been used previously with speech signals to measure phonological precision in pathological speech [14, 15] and as a measure for second language learners [16]. We assume that the phonological class recognizer is trained with "good spoken" English; thus, the posterior probability indicates how well the system "understands" certain sounds, i.e., the closer the phonological class probability to one, the better a speaker pronounces it.

4. Experiments & Results

4.1. Experiment 1: Phonological Class Recognition

We divided the speakers of the USC dataset into training (50 speakers), validation (15 speakers), and test (10 speakers) sets. The test set was used only during inference. The models were trained on an NVIDIA RTX A6000 with 48GB during 30 epochs, a learning rate of 10^{-4}, a weight decay of 10^{-3}, and a batch size of 32. The criterion used for optimization was the AdamW algorithm. Then, we performed three experiments using the ViT model as a reference:

1. Baseline: The default checkpoint of the ViT model is taken as it is, without further training.
2. Fine-tuned ViT: The classification is performed while all parameters of the model are free to be trained.
3. Contrastive learning: The ViT is fine-tuned and the Wav2Vec model is used to compute the cosine embedding loss.

The classification results (in the test set) for each one of the phonological classes are reported in Table 2. Including the speech and contrastive loss during the training significantly improved the class recognition on all metrics. We verified these results by calculating the confidence interval (bootstrapping method) for the best-performing model:

- Contrastive vs baseline: p-value<0.05; $0.568 < \mu < 0.641$
- Contrastive vs fine-tuned: p-value<0.05; $0.046 < \mu < 0.098$

The results also show that the class with the lowest recognition accuracy (considering only the contrastive learning approach) is the alveolar front (e.g. /iti/). By looking at the MRI frames in Figure 3 we can observe that the position of the tongue to produce both consonants is similar, except that the place of articulation for /t/ uses the tongue tip and the /k/ uses the body. Such a "small" difference might be imperceptible for the model; thus, resulting in lower recognition accuracy.

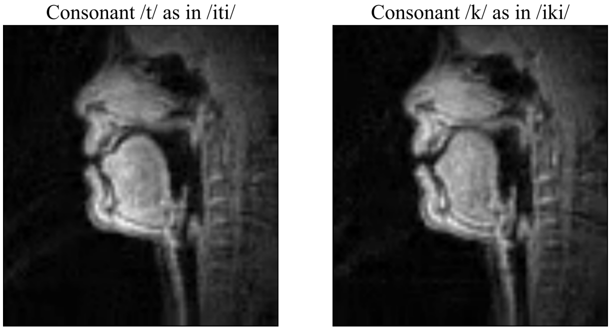

Figure 3: *Example of a person producing the alveolar consonant /t/ and velar /k/*

4.2. Experiment 2: Analysis of Phonological Features in Tongue Cancer Patients

As described in Section 3.3, the output of the phonological class recognizer (after the softmax) is used to measure the model's confidence that an MRI frame belongs to one of nine phonological classes. The higher the number (maximum one) the "better" is the production of the consonant. We computed the phonological posterior probability for the TC patients and controls described in Section 2 and visualized their phonological class posteriors in Figure 4. From the radar plot, it can be observed that the controls had a higher phonological precision than the TC patients in uttering the velar central and velar back consonants compared to the TC patients. In the radar plot, we can also observe a tendency from the patient to have a "higher precision" to produce labial sounds, which does not make sense since there are no labial sounds in the "a geese" and "a souk" tasks. This is an artifact that appears to reflect the lip shapes of the vowels. /i/ uses spread lips and /u/ uses rounded lips. Since the lip position is not specified for the consonants, the lip shape from the vowel spreads to the neighboring consonants.

5. Discussion and Conclusions

In this paper, we proposed to use a contrastive learning approach to enhance the recognition of phonological classes in speech production, especially in the context of clinical applications for patients with tongue cancer. The study is motivated by the challenge of accurately recognizing phonological classes (or phonemes in general), when only MRI data are available, due to the poor quality or absence of synchronized acoustic speech recordings caused by, for instance, the noisy environment of MRI procedures. The results showed that using a contrastive learning-based approach significantly improved the accuracy of phonological class recognition from MRI data, achieving a frame-wise F1-score of up to 0.85. This approach not only enhances the accuracy of phonological class recognition but also demonstrates potential in clinical settings for assessing speech disorders in patients with tongue cancer, offering a promising tool for speech therapy and rehabilitation. The approach involves using two MRI datasets: the USC 75-Speaker Speech MRI dataset for training the model on frame-wise phonological class recognition, and an in-house Cine MRI Tongue Cancer Data set to validate the clinical application of the proposed approach.

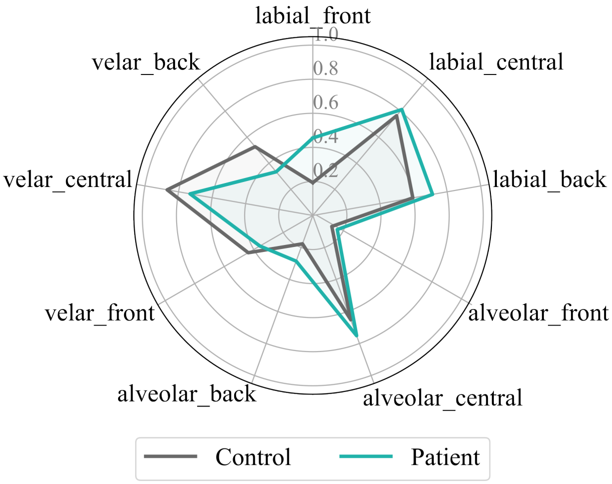

Figure 4: *Phonological articulatory precision (average posterior probabilities) measured in the TC patients and controls*

In this paper, we presented two main experiments: In the first one, we performed phonological class recognition (which improved through contrastive learning), and, in the second experiment, we computed phonological features in tongue cancer patients, using the developed model. The results from both experiments validate the potential of using contrastive learning for improving phonological class recognition from MRI data and its application in clinical settings for assessing speech disorders in patients, such as those with tongue cancer. The method's ability to accurately predict phonological classes even without the need for synchronized speech recordings at inference time could be particularly beneficial in clinical environments where such recordings may not be available.

Overall, we achieved the goals proposed for this study; however, there are different points where we can improve our research. For instance, we relied on the ViT to process the MRI data without performing any pre-segmentation or detection of the region of interest, so the model could focus on the specific motion patterns necessary to recognize the different phonological classes more accurately. In future work we will also investigate the influence of using different MRI modalities on the obtained results i.e., the data used to train the phonological recognizer was captured with a completely different setting than the data used to test the clinical approach. Although the overall results seem to support our claims, we need more data to strengthen our results.

6. Acknowledgements

This work was funded by the EVUK programme ("Next-generation Al for Integrated Diagnostics") of the Free State of Bavaria. The authors also thank the National Institutes of Health for the support through the grants R01DC014717, R01DC018511, and R01CA133015.

7. References

[1] M. H. Franciscatto, M. D. Del Fabro, J. C. D. Lima, C. Trois, A. Moro, V. Maran, and M. Keske-Soares, "Towards a speech therapy support system based on phonological processes early detection," *Computer speech & language*, vol. 65, p. 101130, 2021.

[2] C. Tessier, A. Weill-Chounlamountry, N. Michelot, and P. Pradat-Diehl, "Rehabilitation of word deafness due to auditory analysis disorder," *Brain Injury*, vol. 21, no. 11, pp. 1165–1174, 2007.

[3] C. L. Furia, L. P. Kowalski, M. R. Latorre, E. C. Angelis, N. M. Martins, A. P. Barros, and K. C. Ribeiro, "Speech intelligibility after glossectomy and speech rehabilitation," *Archives of Otolaryngology–Head & Neck Surgery*, vol. 127, no. 7, pp. 877–883, 2001.

[4] G. Saravanan, V. Ranganathan, A. Gandhi, and V. Jaya, "Speech outcome in oral cancer patients–pre-and post-operative evaluation: A cross-sectional study," *Indian Journal of Palliative Care*, vol. 22, no. 4, p. 499, 2016.

[5] A. Acher, P. Perrier, C. Savariaux, and C. Fougeron, "Speech production after glossectomy: Methodological aspects," *Clinical linguistics & phonetics*, vol. 28, no. 4, pp. 241–256, 2014.

[6] P. S. Aleksic, G. Potamianos, and A. K. Katsaggelos, "Audiovisual speech processing," in *The Essential Guide to Video Processing*. Elsevier Inc, 2009, pp. 689–737.

[7] L. Pandey and A. Sabbir Arif, "Silent speech and emotion recognition from vocal tract shape dynamics in real-time mri," in *ACM SIGGRAPH 2021 Posters*, 2021, pp. 1–2.

[8] K. Van Leeuwen, P. Bos, S. Trebeschi, M. J. van Alphen, L. Voskuilen, L. E. Smeele, F. van der Heijden, R. van Son *et al.*, "CNN-Based Phoneme Classifier from Vocal Tract MRI Learns Embedding Consistent with Articulatory Topology." in *Interspeech*, 2019, pp. 909–913.

[9] P. Saha, P. Srungarapu, and S. Fels, "Towards automatic speech identification from vocal tract shape dynamics in real-time MRI," *arXiv preprint arXiv:1807.11089*, 2018.

[10] V. Parthasarathy, J. L. Prince, M. Stone, E. Z. Murano, and M. NessAiver, "Measuring tongue motion from tagged cine-mri using harmonic phase (harp) processing," *The Journal of the Acoustical Society of America*, vol. 121, no. 1, pp. 491–504, 2007.

[11] Y. Lim, A. Toutios, Y. Bliesener, Y. Tian, S. G. Lingala, C. Vaz, T. Sorensen, M. Oh, S. Harper, W. Chen *et al.*, "A multispeaker dataset of raw and reconstructed speech production real-time MRI video and 3D volumetric images," *Scientific data*, vol. 8, no. 1, p. 187, 2021.

[12] J. Lee, J. Woo, F. Xing, E. Z. Murano, M. Stone, and J. L. Prince, "Semi-automatic segmentation of the tongue for 3d motion analysis with dynamic mri," in *2013 IEEE 10th International Symposium on Biomedical Imaging*. IEEE, 2013, pp. 1465–1468.

[13] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[14] T. Arias-Vergara, *Analysis of Pathological Speech Signals*. Germany: Logos Verlag Berlin, 2022.

[15] T. Arias-Vergara, E. Londono-Mora, P. Pérez-Toro, M. Schuster, E. Nöth, J. Orozco-Arroyave, and A. Maier, "Measuring Phonological Precision in Children with Cleft Lip and Palate," *In Proceedings of the 24th Interspeech*, pp. 4638–4642, 2023.

[16] H. Hung, P. A. Pérez-Toro, T. Arias-Vergara, A. Maier, and E. Nöth, "Speaking clearly, understanding better: Predicting the l2 narrative comprehension of chinese bilingual kindergarten children based on speech intelligibility using a machine learning approach," *In Proceedings of the 24th Interspeech*, pp. 4623–4627, 2023.