Supplemental Document
AudioViewer: Learning to Visualize Sounds

This document supplements the main paper by providing more analysis with the related works, more implementation details, content and style disentanglement, an additional ablation study regarding the translation network T, an additional analysis on the human study, and training details. We also prepared an additional supplemental video that contains audio and video snippets that can not be played in a conventional PDF. These give additional qualitative results for spoken sentences and environment sounds.

Licenses. The used audio datesets TIMIT dataset [2] and ESC-50 dataset [5] in our experiments are public. TIMIT dataset [2] is under the terms of LDC User Agreement for Non-Members license and ESC-50 dataset [5] is under the terms of the Creative Commons Attribution Non-Commercial license.

Risk mitigation and scope. Sensory substitution bears non-negligible risks. Our comparative study design is approved by our IRB to have a low risk. Whether an entire language can be learned will require psychophysical studies controlled by domain experts to mitigate the risk of side effects on long-term participants.

1. Relationship Between AudioViewer and the Audio-to-Scene, Audio-to-Text Methods

We would like to highlight that we do not attempt to compete with but to complement existing high-level audio translation methods. Ours addresses a scenario that they cannot handle. The goal of mapping sound to a scene, such as an airplaine or car for their respective engine noise, is to generate the possible corresponding environment image related to the sound information. This does not address the direct translation of sound signals that is desired for learning to speak and other tasks requiring low-level feedback. In a similar vein, mapping speech to text could be applied to adult who want to understand a conversation, but not to children who learn to read only at age 6-8, and not for learning pronounciation in general as no tonal feedback is given. Learning to speak requires a low-level mapping, like the frequency visualization currently applied and the method we developed. We believe it is particularly important as early childhood learning is hampered by hearing deficits. The related work section highlights the advantages and disadvantages. We cannot use these methods as baselines as they do not apply to our setting, e.g., to distinguish individual phones. The goal of our approach is to learn a phoneme-level mapping between sound and visual signals. So we choose the most closely related low-level mapping methods including audio to lip motion and audio to Mel Spectrogram methods as our baselines.

2. Implementation Details

The architecture of audio VAE is shown in Figure 1. We train the audio modules for 300 epochs with batch size 128 and initial learning rate 10^{-3}. We train the low resolution CelebA visual models for 38 epochs and MNIST visual models for 24 epochs with batch size 144 and fixed learning rate 0.005. Figure 2 shows its architecture. For the visual model with high resolution, we use the pre-trained Soft-IntroVAE model [1] provided by the authors. We use two different strategies to link the audio and visual latent spaces. When mapping the audio inputs to low resolution facial and digital images, we fix the audio model and fine tune the visual model on audio and image examples for 10 audio epochs. When linking the audio signals to high resolution facial images, we train the translation module T instead of the entire visual model, whose architecture is shown in Figure 3, with fixed audio model and visual model on audio examples for 10 audio epochs. For optimization, we use Adam [4] with parameters $\beta_1 = 0.95$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$.

3. Results on Environment Sounds

With the style disentangling training of our AudioViewer method, the audio model is more specific to human speech than general sounds. In the following, we test the generalization capability of our method (trained on the speach TIMIT dataset [2]) on the ESC-50 environment sound dataset [5].

Table 1 shows the reconstruction accuracy when going via the audio and video VAEs (see Information Throughput section in the main document). The SNR for reconstructed Mel spectrum is generally lower than the speech dataset, which is expected since it was trained on the latter. The analysis of the reconstruction ability of the audio VAE in isolation (without

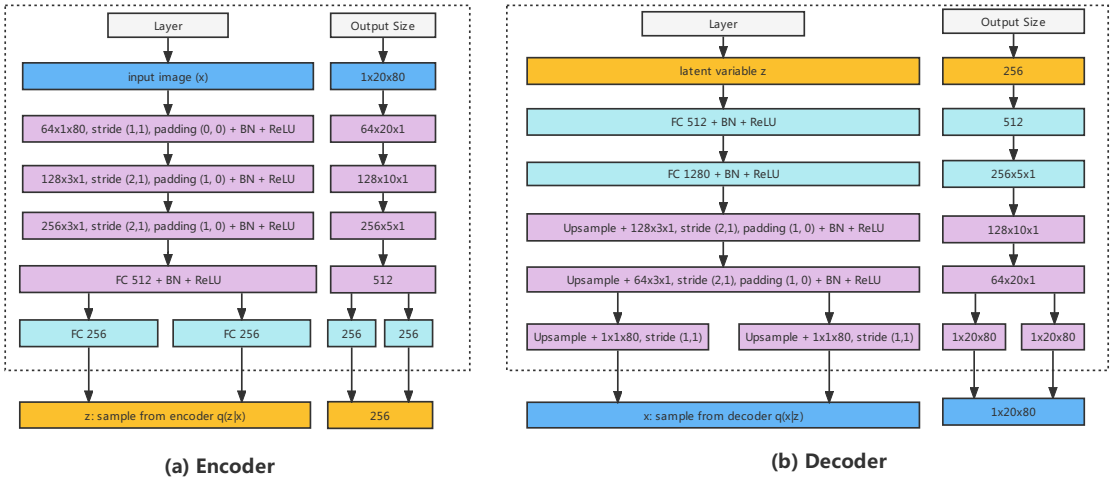

Figure 1. **AudioVAE Framework.** (a) and (b) illustrate, respectively, the encoder and decoder parts of the audio model.

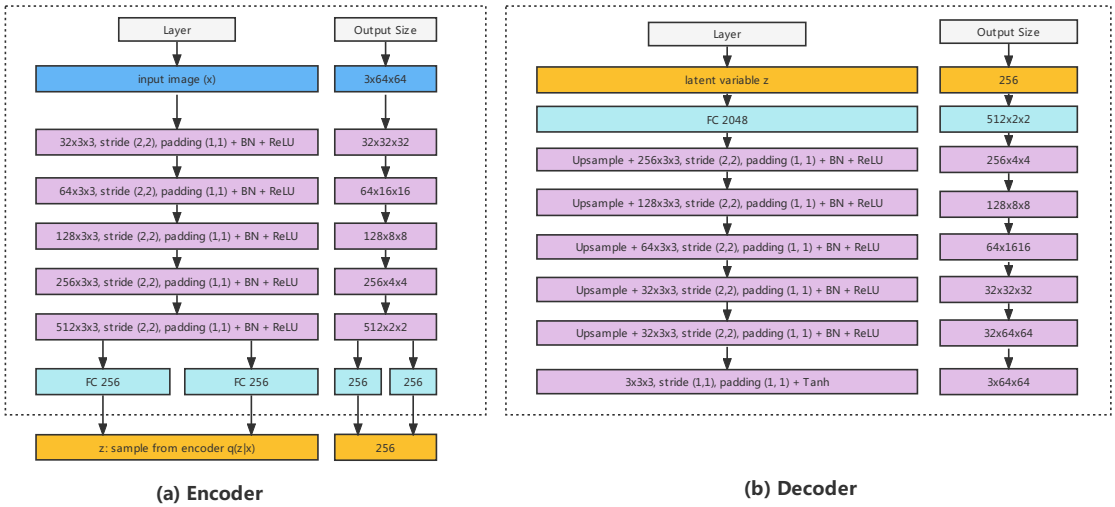

Figure 2. **VisualVAE Framework.** (a) and (b) illustrate, respectively, the encoder and decoder parts of the video model.

going through the video VAE) reported in Table 2 shows that a large fraction of this loss of accuracy stems from the learning of speech specific features of the audio VAE. Moreover, with a recombined reconstruction loss term on the human speech dataset, the model was fitted to speech features and tended to loss high pitch information. Still, according to the face visualization of the content encoding as we showed in the supplemental video, our AudioViewer can generate consistent visualization to given environment sounds.

4. Disentangling content and style

We construct a SpeechVAE that disentangles the style (speaker identity) content (phonemes) in the latent encodings, i.e., the latent encoding $\mathbf{z} = [z_1, \cdots, z_d]^T \in \mathcal{R}^d$ can be separated as a style part $\mathbf{z}_s = [z_1, \cdots, z_m]^T$ and a content part $\mathbf{z}_c = [z_{m+1}, \cdots, z_d]^T$, where d is the whole audio latent space dimension and m in the audio style latent space dimension.

We use an audio dataset with phone and speaker ID annotation. However, this still requires to disentangle the audio signal into style and content codes, which we obtain similarly to [6] by mixing embeddings from different speakers. Fig-

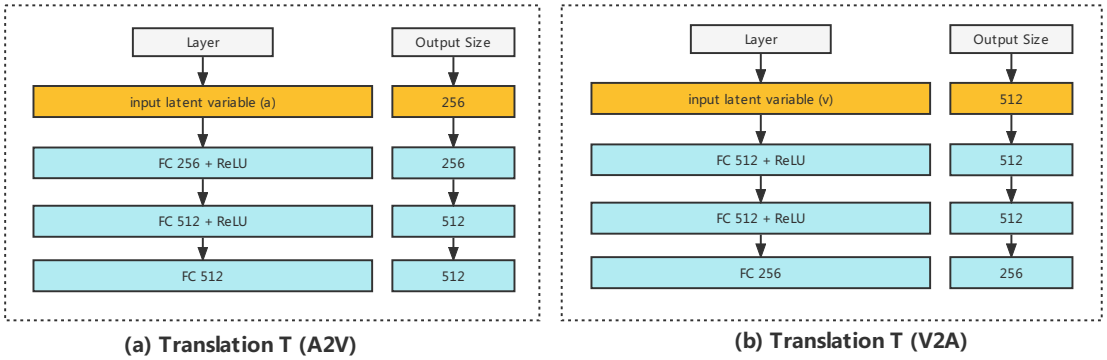

Figure 3. **Translation Networks.** (a) and (b) illustrate the translation networks mapping audio signal to visual signal, and vice verse.

ure 4 gives an overview. At training time, we feed triplets of mel spectrogram segments $\mathbf{T}_{a,b,i,j} = \{\mathbf{M}_{a,i}, \mathbf{M}_{b,i}, \mathbf{M}_{a,j}\}$, where $\mathbf{M}_{a,i}$ and $\mathbf{M}_{b,i}$ are the same phoneme sequence p_i spoken by different speakers s_a and s_b respectively, and $\mathbf{M}_{a,j}$ shares the speaker s_a with the first segment but a different phoneme sequence. Each element of the input triplet is encoded individually by E_A, forming latent triplet $\{\mathbf{z}_{a,i}, \mathbf{z}_{b,i}, \mathbf{z}_{a,j}\} = \{[\mathbf{z}_{s_a}, \mathbf{z}_{c_i}]^T, [\mathbf{z}_{s_b}, \mathbf{z}'_{c_i}]^T, [\mathbf{z}'_{s_a}, \mathbf{z}_{c_j}]^T\}$, instead of reconstructing the inputs from the corresponding latent encodings in an autoencoder, we reconstructed the

Table 1. **Information throughput on environment sounds,** showing that the reconstruction error increases when evaluating on a test set that contains sounds vastly different from the training set (speech vs. environment sounds).

Audio models	Visual models	SNR(dB)
Audio PCA	Visual PCA	17.23
SpeechVAE	DFC-VAE on CelebA	1.03
	DFC-VAE on MNIST	2.22
	DFC-VAE on CelebA (refined w/ $\mathcal{L}_{cycle}$)	**2.83**
	DFC-VAE on MNIST (refiend w/ $\mathcal{L}_{cycle}$)	0.76
SpeechVAE w/ $\mathcal{L}_{p,\log MSE}$, $\mathcal{L}_{rr}$, dim=256	DFC-VAE on CelebA	0.46
	DFC-VAE on MNIST	0.46
	DFC-VAE on CelebA (refined w/ $\mathcal{L}_{cycle}$)	1.26
	DFC-VAE on MNIST (refiend w/ $\mathcal{L}_{cycle}$)	**1.68**

Table 2. **Audio VAE mel spectrum reconstruction.** The average SNR of autoencoding and decoding mel spectrograms on ESC-50 shows a significant reconstruction loss. The average speed and acceleration between the latent vector (dim = 128) of neighbouring frames ($\Delta t = 0.04$s) confirms the experiments in the main document, that smoothness comes at the cost of lower reconstruction accuracy.

Audio models	SNR (dB)	Velocity (s^{-1})	Acc.(s^{-2})
Audio PCA	**17.23**	**170.13**	**6960.80**
SpeechVAE[3]	10.17	172.57	7331.77
SpeechVAE w/ $\mathcal{L}_{p,\log MSE}$	**8.92**	58.78	1859.95
SpeechVAE w/ $\mathcal{L}_{rr}$	2.98	40.54	1580.86
SpeechVAE w/ $\mathcal{L}_{p,\log MSE}$, $\mathcal{L}_{rr}$	2.19	**30.66**	**909.39**
SpeechVAE w/ $\mathcal{L}_{p,\log MSE}$, $\mathcal{L}_{rr}$, dim=256	2.51	33.88	1037.97

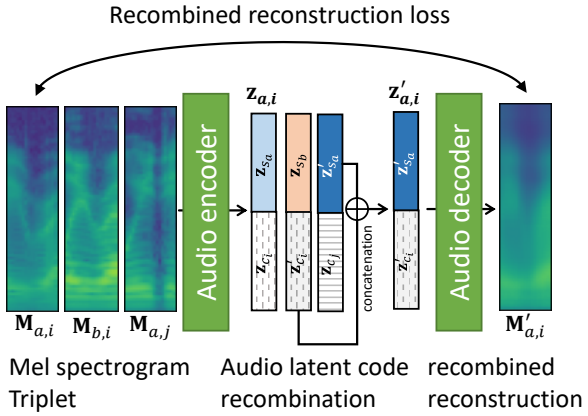

Figure 4. **Disentanglement**, by mixing content encodings from two speakers saying the same phones ($\mathbf{z}_{c_i}$ with $\mathbf{z}'_{c_i}$) and style encodings from the same speaker saying different words ($\mathbf{z}_{s_a}$ with $\mathbf{z}'_{s_a}$).

first sample $\mathbf{M}_{a,i}$ from a recombined latent encoding of the other two, $\mathbf{z}'_{a,i} = [\mathbf{z}'_{s_a}, \mathbf{z}'_{c_i}]^T$. Formally, we replaced the reconstruction loss term in the VAE objective by a recombined reconstruction loss term,

$$\mathcal{L}_{rr}(\mathbf{T}_{a,b,i,j}) = \mathbb{E}_{q_\phi(\mathbf{z}'_{\mathbf{a},\mathbf{i}}|\mathbf{M}_{b,i},\mathbf{M}_{a,j})}\big(\log p_\theta(\mathbf{M}_{a,i}|\mathbf{z}'_{a,i})\big). \quad (1)$$

This setup forces the model to learn separate encodings for the style and phoneme information while not requiring additional loss terms.

Note that we could alternatively enforce $\mathbf{z}_{a,i}$ to be close to $\mathbf{z}'_{a,i}$ without decoding (the unused $\mathbf{z}_{a,i}$ in Figure 4). However, an additional L2 loss on the latent space led to a bias towards zero and lower reconstruction scores than the proposed mixing strategy that works with the original VAE objective.

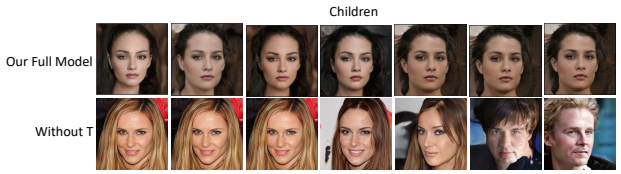

Figure 5. **The functionality of the translation network T.** Here we simply repeat the audio latent codes to match different latent space dimensions and support. Compared with the results of our full model, the smoothness of the generated image sequences is poor.

5. Ablation study on translation network T

The translation network T is used to build the bridge between audio-vision latent spaces with different dimensions and support. To evaluate the function of T, we tried in preliminary experiments to replace T by simply repeating the audio latent code to match latent space dimension. Figure 5 shows that the resulting image sequences have poor smoothness as their support is not matched. This is confirmed with a low throughput score of -0.22, showing that the sound information is lost by mapping to regions that the image decoder does not support.

6. Human study I - Discriminating Sounds

The study was conducted with 22, 14, 15, 12, 14, 14, 14 participants for the CelebA-HQ-content, CelebA-content, CelebA-style, CelebA-combined, MNIST-content, MakeItTalk [7] and mel spectrogram (MEL) questionnaires, respectively. Each version of the questionnaire asked the same set of 29 questions with randomized ordering of answers within each question. It took participants between 10-15 minutes to complete the each questionnaire. The questionnaire asked participant to perform two possible tasks: matching and grouping visualizations. The format of the questionnaire is outlined in Table 3. The questions tested for two factors: sound content, sounds that share the same phoneme sequences, and sound style, sounds produced by speakers of the same sex or speaker dialect. In total, we tested 100 different sounds and words. This purely visual comparison allows us analyze different aspects of the translation task individually.

Matching questions Matching questions asked the participants to choose which of two possible visualizations which is most visually similar to a given reference visual. Figure 6 shows examples of matching questions. Matching questions were used to assess the viability for users to distinguish between the same sounds produced by speakers possessing different speaker traits as well as determining whether structural similarities in the underlying audio translated into similarities in the visualization. In particular, the

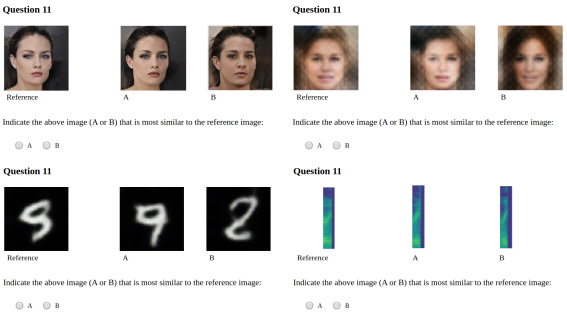

Figure 6. **Matching Example.** Examples from the CelebA-HQ-content, CelebA-combined, MNIST-content, MEL questionnaire of a matching question asked to participants in the user study.

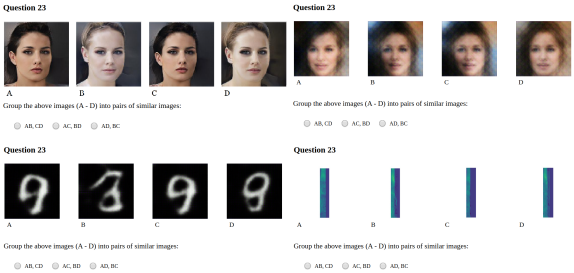

Figure 7. **Grouping Example.** Examples from the CelebA-HQ-content, CelebA-combined, MNIST-content and MEL questionnaire of a grouping question asked to participants in the user study.

Table 3. **A breakdown of the questionnaire format** by sound type and tested factors, listing their frequency of occurrence.

Question type	Sound type	Tested factor	Questions
Matching questions	Phone-pairs	Content	3
		Style (sex)	3
		Style (dialect)	2
	Words	Content	3
		Style (sex)	3
		Style (dialect)	2
Grouping questions	Phone-pairs	Content + style (dialect)	2
		Content + style (dialect + sex)	2
	Words	Content	3
		Content + style (dialect)	2
		Content + style (sex)	1
		Content + style (dialect + sex)	1
		Content (similar sounding words)	2
Total			29

questionnaire contained 6 questions for evaluating the ability to distinguish between sound content, which compared visualizations of sounds of different phoneme sequences (3 for phoneme-pairs and 3 for words). Phone-pairs are short in length and therefore the corresponding visualisation was a single frame image, whereas visualisations of words were videos. In order to evaluate the ability to distinguish between sound style, 6 questions compared visualizations of the same phoneme sequence between male and female speakers and 4

Table 4. **User study results.** Values indicate mean accuracy and standard error for distinguishing between visualizations of the tested factor across participants as a percentage. The disentangled representation clearly outperforms the combined baseline.

Tested Factor	PCA-Baseline	CelebA-disentangled	CelebA-combined
Content	38.4	**85.0** $\pm$ **1.8**	72.8 ± 2.9
Style (dialect)	50.0	**56.7** $\pm$ **5.7**	39.6 ± 7.2
Style (sex)	43.3	**78.0** $\pm$ **2.8**	43.3 ± 2.6

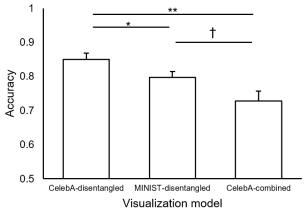

Figure 8. **Model Comparison on content questions.** The mean user accuracy and standard error for distinguishing phoneme sequences is compared between the models, annotated with the significance level († for $0.05 < p < 0.1$, * for $p < 0.05$, ** for $p < 0.01$). The mean accuracy for random guessing is 0.384.

questions for distinguishing between speakers of different dialects. In total there were 16 matching questions. Since each question has two options, the expected mean accuracy for random guessing is 50%.

Grouping questions. Grouping questions asked the participants to group 4 visualizations into two pairs of similar visualizations. Figure 7 shows examples of grouping questions. Grouping questions were used to assess the degree to which visualizations of different words are distinguishable and visualizations of the same word are similar. In particular, the study required users to group visualizations of two pairs of sounds, whereby different pairs are sound clips with shared factors of the same sound content or same sound style. In total, the human study consisted of 4 grouping questions based on phone-pairs and 9 grouping questions based on words. Since there are three possible options, the expected mean accuracy for random guessing is 33.3%.

Results. For each of the models, we tested for sound content: phoneme sequences, and sound style: speaker dialect and speaker sound. We generated the mean accuracy and standard deviation for each tested factor and each question sub type. Table 4 extends the results shown in the main document by comparing entangled and disentangled representations. The results of the disentangled models with CelebA visualizations (CelebA-disentangled) is aggregated by taking the results of CelebA-content on the questions which tested for sound content and the results of CelebA-style on questions which tested for sound style.

We analyze the significance of our improvements by reporting accuracy, standard error, and using the student-t test.

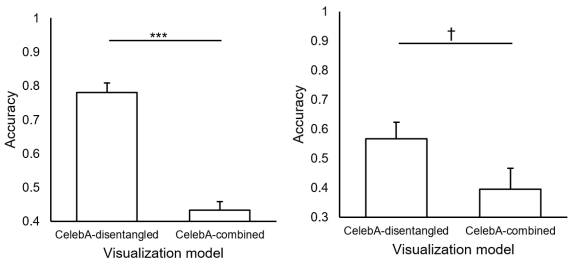

Figure 9. **Model Comparison on style questions.** The mean user accuracy and standard error for distinguishing speaker sex (left) and dialect(right) is compared between the models, annotated with the significance level († for $0.05 < p < 0.1$, *** for $p < 0.001$). The mean accuracies for random guessing are 0.43 and 0.5 respectively.

For MEL the accuracy is 66.2 ± 1.3 with ($p = 7\text{e}{-12}$) and for MakeItTalk [7] the accuracy is is 47.7 ± 2.7 with ($p = 6\text{e}{-12}$). Figure 8 shows the significance of our other models, it illustrates that users achieve the overall accuracy on the CelebA-disentangled model with $85.0 \pm 1.8\%$ (significant with $p < 0.05$) for distinguishing between visualizations of different content. The MNIST-content model has the highest accuracy for distinguishing between different phone pairs with $91.8 \pm 2.5\%$, although not significantly higher than the CelebA-disentangled one with ($p > 0.05$), but has a much lower accuracy for distinguishing between different words, suggesting that the MNIST visualizations may be better suited for representing shorter sounds. The CelebA-disentangled model outperforms the CelebA-combined model for distinguishing between speakers of different sex with $78.0 \pm 2.8\%$ (significant with $p < 0.001$) and between speakers of different dialects with $56.7 \pm 5.7\%$ (marginally significant with $0.05 < p < 0.10$) as shown in Figure 9. The task of distinguishing between different speakers of different dialects is much more difficult than distinguishing between phoneme sequences since there are 8 categories of dialects in the dataset and differences in dialects are much more subtle and can contain often contain overlaps. Significance comparing model means were calculated using a two-sample two-tailed t-test with unequal variance and without any outlier rejection.

7. Human study II - Learning Sounds

In the second study, we evaluate whether participants can learn to recognize sounds from our visualizations. Conceivable large scale studies would require a large amount of time to train users, which would take months and and be very expensive when scaling to a representative group size. It is hence not effective for comparing our algorithm variations and unsuitable for establishing a benchmark for future methods. To more effectively evaluate methods, we set human studies in a simple but representative environment that can quickly be repeated with a new cohort of participants to

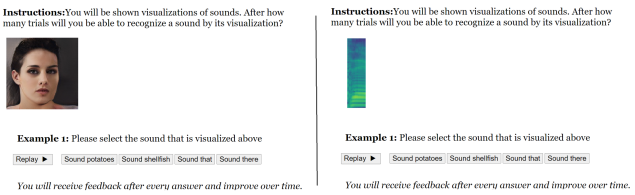

Figure 10. **Learning study examples.** A video generated by the disentangled content model (left)/MEL (right) is shown to evaluate the participants' capability in learning to recognize sounds from visual contexts. Each video corresponds to one variant of four word labels.

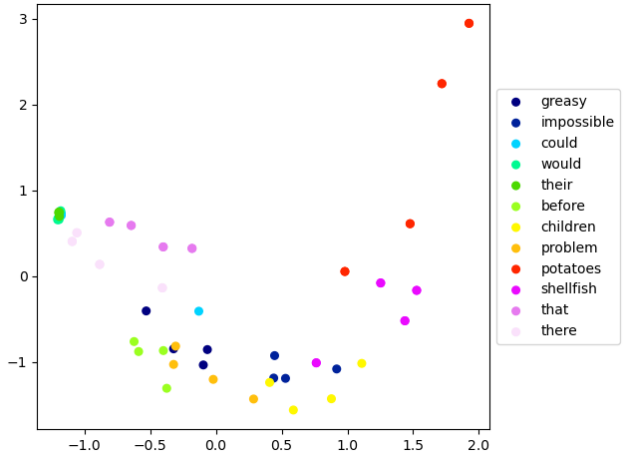

Figure 11. **Sound embedding distance** visualized by the latent embedding of each word's speech projected to 2D. The same color points correspond to different sounds of the same word.

compare the latest methods. Specifically, as described in the main paper, we design a small speech dataset containing 16 words, 4 words with different meanings spoken by different people. Among them, 'that' and 'there' are pronounced with similar length and phonemes (phone editing distance of one or two, depending on the pronunciation), while 'potatoes' and 'shellfish' are pronounced in similar length but contain completely different phonemes (editing distance 8 or more). Note that we chose word pairs that have a similar length as words of different length would make it easy to distinguish simply by the length of the generated video.

To further analyze the distance between the chosen words, we visualize the latent distance between sounds in Figure 11. We randomly select 12 words, each word is spoken by 4 different people. We first encode them to the latent space. Then we expand the latent codes of all words to the same dimension by filling in one for shorter ones. Finally, we embed them into two dimensions using PCA to preserve distances in the original high-dimensional space. This embedding shows that 'there' and 'that' is very close in the embedding while 'patatoes' and 'shellfish' are separated by other words (light green). Please note that this is a non-linear embedding from high-dimensional space that best approximates distances but contains some deformation. Moreover, due to the concatenation and one padding, a slight differences in duration or speed would lead to quite different word embeddings. However, we did not see another way of visualizing such word embedding in a 2D space.

To showcase the robustness to different speakers, we included words spoken by people with two different dialects. To ensure a fair comparison to using the MEL specturm baseline, we decided to only test on words spoken by the same gender (male speakers). Otherwise, the MEL spectrum of female speakers would look very different from male speakers due to their higher pitch. This would lead to slower learning on MEL and results would hence only support the better gender normalization of our method (as validated in Study I) instead of validating better learnability. Words are given in random order to participants in order to avoid any bias towards the visualizations that appear first.

We considered testing the learning of sentences or phones (cf. Study I on distinguishing but not learning). However, entire sentences are longer and contain more information than a single word. Hence, it would be easier to distinguish them. Moreover, single phones map only to a single frame and hence would not test the continuous translation into a video. Since our approach is a low-level sound visualization method, it is more challenging and suitable to select words than either sentences or single phones for evaluating learnability.

Examples of user learning are shown in Figure 10. The learnability of our method is evaluated by comparing the tracking of the changes in the accuracy curves between our model and MEL.

Results. We recruited 9 participants for human study II. It took participants between 10-15 minutes to complete each variant. During the learning period, samples are shown in a random order to avoid bias to the ordering. The learning curves are reported in the main document. The accuracy after 16 rounds of learning is for Ours 87.0% vs. MEL 57.8%, a significant improvement with ($p = 0.016$). We conclude that compared to spectogram representations, our mapping to images of faces or digits is more natural and easier for people to distinguish and match. Whether remaining ambiguities could be overcome by longer learning sessions and how the learning can be further facilitated remains an open question for future work. The main paper discusses the results in more detail.

References

[1] Tal Daniel and Aviv Tamar. Soft-introvae: Analyzing and improving the introspective variational autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4391–4400, 2021.

[2] John S Garofolo. Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium, 1993*, 1993.

[3] Wei-Ning Hsu, Yu Zhang, and James Glass. Learning latent representations for speech generation and transformation. In *Interspeech*, pages 1273–1277, 2017.

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[5] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015.

[6] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, pages 5210–5219. PMLR, 2019.

[7] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: Speaker-aware talking-head animation. *ACM Trans. Graph.*, 39(6), Nov. 2020.