Article

Data-driven Stimulus Continuum Generation with Variational Autoencoder

Zhu Li [1], Yuqing Zhang [1], Dengfeng Ke [1], Binghuai Lin [1] and Yanlu Xie [1,*]

[1] Beijing Language and Culture University, School of Information Science, 100083, Beijing; lzblcu19@gmail.com; yuqingelsa@gmail.com; dengfeng.ke@blcu.edu.cn; binghuailin@tencent.com

* Correspondence: xieyanlu@blcu.edu.cn

Featured Application: The VAE-based stimulus continuum generation approach can be used in speech perception studies to generate smoother and more gradual transitions between two endpoint reference stimuli.

Abstract: Creating stimuli for studies on the categorical perception of speech sounds involves manual manipulation of acoustic parameters (e.g., pitch contours for lexical tone perception, formant frequencies for /r, l/ perception) extracted from spoken words. Difficulties arise when manipulated parameters need to be gradual and smooth transitions between two reference conditions. Furthermore, manually interpolating between endpoint parameter values may lead to unnatural sounding re-synthesized stimuli. Recent studies have demonstrated the effectiveness of deep probabilistic generative models for generating meaningful samples based on embeddings created by performing linear interpolation in latent space. Our work bridges stimulus continuum generation and state-of-the-art deep learning (DL) techniques. We propose a data-driven approach to stimulus continuum generation based on Variational Autoencoders (VAEs). The unsupervised neural network maps the high-level acoustic features into low-dimensional representations that follow a normal distribution. This allows to traverse between two known locations in latent space and produce desired perceptual characteristics. We illustrated this approach in two case studies on syntheses of tone continuum and /ʐ/-/l/ continuum in Mandarin Chinese. Analyses of reconstruction error and subjective evaluations (i.e., identification test and mean opinion score (MOS)) show that our proposed method slightly improves the naturalness of stimulus samples.

Keywords: speech synthesis; variational autoencoders; fundamental frequency; acoustic parameters; continuum

Citation: Li, Z.; Zhang, Y.; Ke, D.; Lin, B.; Xie, Y. Data-driven Stimulus Continuum Generation with Variational Autoencoder. *Journal Not Specified* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

1. Introduction

In speech perception studies, stimulus continua (i.e., several sets of artificially generated stimuli varying along a specific dimension between two given categories) are often used as experimental materials to probe human speech perception mechanisms. The quality of a synthetic stimulus continuum has a particularly significant impact on the result of perceptual experiments. A common approach is to manually modify the key acoustic parameters of natural speech sounds, which is time-consuming and laborious. For example, in perception experiments of lexical tones [1,2], phonemic categories (e.g., the perception of English liquid consonants /r, l/ by Japanese learners [3,4]), and physiological characteristics (e.g., voice gender perception [5]), the synthesis of perceptual stimuli usually includes three steps: (i) extract relevant acoustic parameters from spoken utterances; (ii) perform interpolation operations between the relevant acoustic parameters based on mathematical formulas; (iii) use a vocoder to convert parameter sequences obtained by interpolation back to the speech signal. Although this method has been widely used in perceptual experiments and proven effective, it has several crucial limitations. First, it is difficult to achieve a global and smooth transition between two endpoint stimuli by directly operating on key acoustic cues [6]. Chances are that the resulting stimulus set sounds unnatural when the reference conditions differ in several acoustic dimensions (e.g., /r/ and /l/ differ in first, second and third formant frequencies [3]). Second, since acoustic parameters are continuous physical variables, directly performing interpolation by hand for key acoustic features may mask subtle but important dynamic variations that are used as discriminative clues by listeners, and none of these clues will appear in the ensuing perception experiment [7].

Recently, generative modeling has demonstrated the potential to become an important tool for exploring the parallels between perceptual, physical, and physiological representations in fields such as psychology, linguistics, and neuroscience [8–10]. In this study, we propose a new method for creating a series of stimuli for categorical perception experiments. This is a data-driven approach based on VAEs [11] to model the generative process of the key acoustic feature of original signals. The VAE is a generative model based on a regularized version of the standard autoencoder (AE). The AE is an unsupervised modeling approach that compresses the data (original space) into low dimensional variables (latent space) while attempting to preserve as much information as possible. This means that given a set of acoustic features, like f_0 contours, one can obtain a compact description of variations in the whole curve set in the latent space. In addition, VAE puts a constraint on the latent space, so that the original data is not encoded by a single point, but a standard normal distribution over the latent space. The advantage of this method is that the learned model has the ability to generate new samples, which may not exist in the original data. Figure 1 illustrates the idea behind our approach intuitively: Input A and Input B are the two samples in the original data (in our case studies, they correspond to f_0 contours or vocal tract parameters extracted from monosyllabic words), and the two gradient circles below (Distribution A and Distribution B) are the normal distributions encoded in the latent space. The intuition is that when the point we sampled is closer to the center of the distribution, the sample reconstructed is more similar to the original data.

Interpretable representation learning in the latent space has been extensively investigated for a variety of tasks. [12] has examined the effects of latent space inter-class sampling data augmentation on image classification. [13,14] have demonstrated how to do image transformation via latent space interpolation. Latent space interpolation has also been successfully used in music applications [15–17]. Moreover, [18,19] have successfully applied the VAE to the task of modeling and transforming frame-wise spectral envelopes and spectrograms via sampling from the latent space. However, despite considerable attention devoted to modeling natural speech and interpreting learned representations from the latent space, relatively few studies have attempted to introduce these advanced DL models to address questions of interest in the field of speech perception.

This study proposes a data-driven approach to stimulus continuum generation based on VAEs. There are three major contributions in this paper. First, our work bridges stimulus continuum generation and state-of-the-art DL techniques by applying a data-driven approach (VAEs) to stimulus continuum generation. Second, instead of directly performing manipulation on key acoustic cues, our proposed approach performs resampling after learning the distributions of key acoustic features, which avoids possible information loss and problems of unnaturalness caused by manual interpolation. Third, we conduct two case studies on tone continuum generation and /ʐ/-/l/ continuum generation and the results prove the effectiveness of our method.

2. Related work

2.1. Variational Autoencoders

A VAE [11] is a generative model based on a regularized version of the standard autoencoder (AE). An AE is a form of the deep neural network built to learn a bottleneck for data that ensures only the main structured part of the information can go through

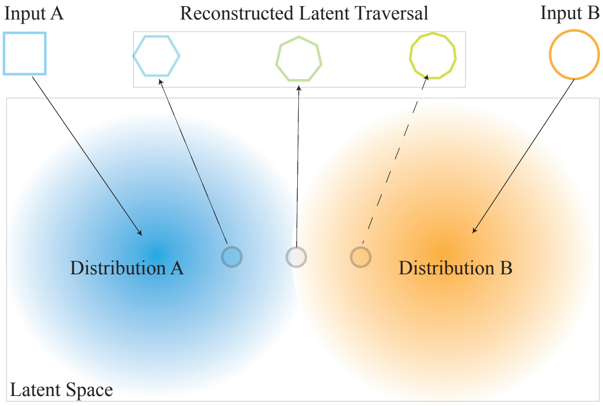

Figure 1. A schematic illustration of the latent space of VAE. Solid lines represent encoding and dashed lines represent decoding.

and be reconstructed. The generic AE architecture comprises an encoder that receives the input signal and transforms it through a bottleneck layer to a latent low-dimensional representation (i.e., the latent code) and a decoder that regenerates the input signal from the latent representation.

However, AEs are not generative models [20] since they do not model the joint probability of the observable and target variables. In order to enable AEs to have the generative ability in the latent space, [11] proceeded to a slight modification of the encoding-decoding process: instead of encoding an input as a single point, they encoded it as a distribution over the latent space by variational inference. The architecture of a VAE model is shown in Figure 2(a). A variational encoder maps an input vector x into a latent space representation z using an encoder neural network with parameters ϕ that outputs $q_\phi(z|x)$, i.e., a probability distribution of the hidden representation conditioned on the input. In fact, $q_\phi(z|x)$ is an approximation of the intractable true posterior $p_\theta(z|x)$, which takes a multivariate Gaussian form with a diagonal covariance matrix, i.e., for a given input data point x:

$$q_\phi(z|x) = \mathcal{N}(z; \mu_x, \sigma_x) \tag{1}$$

Thus the output of the encoder network, for a given input x is a vector of N means and N variances, where N is the chosen dimension of the latent space representation z. We can then sample the posterior distribution using the reparametrisation trick:

$$z = \mu_x + \sigma_x \circ \epsilon \;, \quad \text{where} \epsilon \sim \mathcal{N}(0, 1). \tag{2}$$

The obtained sample can then be passed through the decoder neural network with parameter θ, which models $p_\theta(x|z)$, and outputs an approximation of the original input vector x. The parameters of the encoder and decoder networks ϕ and θ are trained using backpropagation and gradient descent so that the VAE reproduces its input as close as possible. As a by-product of this process, the VAE learns the $q_\phi(z|x)$, structuring the latent space representation.

2.2. Sequential modeling with gated CNN

Gated CNN [21] is a non-recurrent approach that is competitive with strong recurrent models on these large-scale language tasks. Several gating mechanisms have been explored in modern convolutional architectures for sequential modeling [22,23]. Parallel to our work, to capture long- and short-term dependencies in f_0 contours and spectral envelopes, we use a gated CNN [21] to construct both the encoder and decoder networks of the VAE. Having linear units coupled to the gates reduces the vanishing gradient problem. This retains the non-linear capabilities of the layer while allowing the gradient to propagate through the linear unit without scaling. The output of the l_{th}

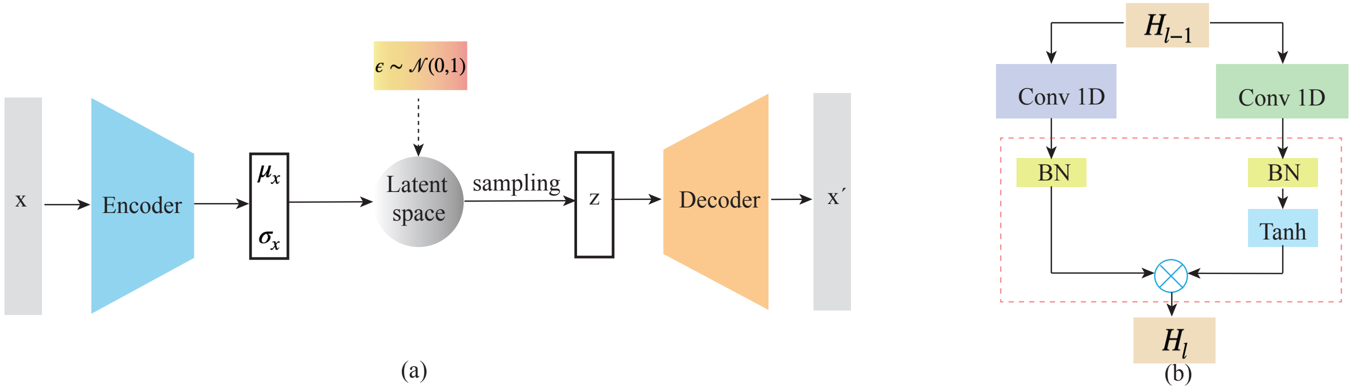

Figure 2. (a) A VAE architecture. Input x represents a key acoustic feature and x' is the reconstructed feature. The model is trained as follows: first, the input x is encoded as distribution $x \sim (\mu_x, \sigma_x)$ over the latent space; second, a point z from the latent space is sampled from that distribution; third, the sampled point x' is decoded and the reconstruction loss and KL loss can be computed; finally, the total loss is backpropagated through the network. (b) Gated CNN used in encoder and decoder.

hidden layer of a gated CNN is described as a linear projection $H_{l-1} * W_l + b_l$ modulated by an output gate $tanh(H_{l-1} * V_l + c_l)$ (as shown in Figure 2(b))

$$H_l = (H_{l-1} * W_l + b_l) \otimes tanh(H_{l-1} * V_l + c_l) \tag{3}$$

where W_l , V_l , b_l and c_l are the network parameters to be trainedand $\otimes$ indicates the element-wise product. Here, the input to the 1st layer is $H_0 = x$ for the encoder and $H_0 = z$ for the decoder whereas the output from the l_{th} layer is $H_l = [\mu_z; \sigma_z]$ for the encoder and $H_l = [\mu_x]$ for the decoder. Similar to LSTMs, the output gate multiplies each element of $H_{l-1} * W_l + b_l$ and controls what information should be propagated through the hierarchy of layers in a data-driven manner.

3. Experiments

Two sets of comparison experiments were conducted to synthesize the lexical tone continuum and the /z/-/l/ continuum, using our proposed approach based on the VAE and the traditional approach based on signal processing respectively.

3.1. The VAE approach

3.1.1. Dataset

The data for this study were based on recordings taken from the BLCU-SAIT speech corpus [24]. The corpus consists of both native and nonnative speech with monosyllabic and disyllabic words and multi-syllabic sentences. We selected the single-syllable speech data produced by a female native speaker, totaling 1520 monosyllabic words that cover all possible tones and initials in Mandarin.

3.1.2. Data preprocessing

The WORLD analyzer [25] was used to extract the required acoustic features. WORLD is a real-time processing analyzer consisting of three algorithms for obtaining three speech parameters, i.e., fundamental frequencies (f_0s), spectral envelopes (SPs) and aperiodic parameters (APs). As shown in classical speech perception studies, f_0 is the primary acoustic cue to lexical tones; therefore f_0 values were extracted for tonal continuum synthesis experiments. Similarly, we used the SP as another acoustic feature to carry out experiments on continuum synthesis of vocal tract parameters. The original speech recordings were downsampled to 22.05 kHz. Pitch parameters were set at a minimum of 50 Hz, a maximum of 600 Hz and the frame shift was 5 ms in the WORLD analyzer for f_0 extraction. We extracted 34 Mel-cepstral coefficients (MCEPs), fundamental frequency (f_0), and aperiodicities (APs) using the WORLD analyzer. As a pre-processing step, the extracted f_0 contours and MCEPs were normalized so that they ranged from -1 to 1. The detailed spectral analysis and synthesis settings were the same as in the previous work [27].

3.1.3. Training configuration

The deep learning toolkit used in this work is Pytorch [28]. The model was trained with the Adam optimizer and the initial learning rate was set to 0.001. All configurations were trained for a maximum of 20000 iterations with a batch size of 64 spoken words. Following the usual practice [27], we randomly cropped a segment (80 frames) from a randomly selected word instead of using the whole word directly, so as to increase the randomness of training data. Table 1 and Table 2 provide details of the network architectures of our proposed model for speaking voice pitch contours (f_0s) and spectral envelopes (SPs).

Layer	channel	Stride × Kernel	GLU
Input	1	-	-
Cov1d	32	1×61	GLU
Cov1d	16	1×21	GLU
Cov1d	8	1×5	GLU
Latent	1	-	-
Cov1d	32	1×1	GLU
Cov1d	16	1×21	GLU
Cov1d	8	1×5	GLU
Output	1	-	-

Table 1: VAE architecture to model the generative process of f_0. Conv1d refers to the 1D convolutional layer. Latent refers to the Gaussian parametric layer modeling z. GLU refers to gated linear unit.

Layer	channel	Stride × Kernel	GLU
Input	1	-	-
Cov2d	128	$(1,1) \times (5,9)$	GLU
Cov2d	256	$(2,2) \times (5,5)$	GLU
Cov2d	128	$(2,2) \times (5,5)$	GLU
Latent	1	-	-
Cov2d	128	$(1,1) \times (1,1)$	GLU
Cov2d	256	$(2,2) \times (5,5)$	GLU
Cov2d	128	$(2,2) \times (5,5)$	GLU
Output	1	-	-

Table 2: VAE architecture to model the generative process of SPs. Conv2d refers to the 2D convolutional layer.

3.1.4. Stimulus continuum generation

Stimulus continuum generation with VAE contains two phases: the training phase and the generation phase. Here we conducted tone continuum generation and /z/-/l/ continuum generation experiments using related acoustic features (f_0 and SP respectively) to demonstrate the effectiveness of our proposed approach.

3.1.5. Tone continuum generation

In the training phase, we trained a VAE framework on the extracted f_0 dataset (details in 3.1.2) to model the probabilistic generation process of fundamental frequencies. This process was done on a single Tesla K40c GPU, which took around an hour. Using the trained VAE model, we can generate a pitch continuum between any two reference conditions. To prove the effectiveness of our approach, we took Chinese monosyllables /a1/ and /a2/ as the endpoint stimuli to create a 9-interval lexical tone continuum.

In the generation phase, due to differences in duration of two spoken words, we did time normalization using PSOLA [29]. Similar to the training phase, we extracted the f_0s, SPs and APs from the recordings of /a1/ and /a2/ using the WORLD analyzer [25] and normalized the two f_0 contours to -1 and 1. Second, we sent the two preprocessed f_0 contours to the encoder of the VAE for f_0 and obtained two latent representations (normal distribution) of the original data. Third, we sampled latent representations equidistantly between the two reference distributions according to equation 4,

$$\hat{z} = \alpha * z_1 + (1 - \alpha) * z_2 \tag{4}$$

where α is an interval between [0, 1] and $\hat{z}$ refers to the latent representation which can walk in the continuous latent space when the parameter α is changing from 0 to 1. The interpolation code $\hat{z}$ is fed into the decoder of the trained VAE model, which outputs smooth transitions between the two original inputs. Finally, we used the WORLD analyzer [25] to apply the new f_0 contours to the speech signals and obtained equidistant pairs of stimuli along the pitch continuum.

3.1.6. /ʐ/-/l/ continuum generation

Since pitch is a suprasegmental acoustic feature, in order to show that our approach to continuum generation is also applicable to segmental features, /ʐ/-/l/ continuum generation experiment was performed using vocal tract parameters. In the training phase, we trained a VAE framework on the extracted SPs' dataset (details in 3.1.2) to model the generation process of vocal tract parameters. To prove the effectiveness of our approach, we created a /ʐ/-/l/ continuum between two Chinese monosyllables /re1/ and /le1/ as a case study. In the generation phase, all the steps were similar to the tone continuum generation, except that the acoustic parameters were spectral envelopes instead of f_0 contours.

3.2. Traditional approach: manual manipulation

For comparison, we referred to the common continuum stimulus synthesis method adopted in most of the classical speech perception studies [2,26]. The standard procedures of synthesizing the stimuli are: (1) adjusting the duration of the target syllables to 400 ms, (2) extracting the f_0, SP and AP parameters from two given speech signals using the WORLD analyzer, (3) reducing the number of pitch points to 10, with one at the starting position, one at the ending position, and eight intermediate points selected in equal steps, (4) synthesizing various stimuli by manually adjusting the above ten points.

As in the deep learning approach, the same target syllables were used as the reference stimuli for tone continuum generation and /ʐ/-/l/ continuum generation.

4. Results and Discussion

4.1. Objective evaluation: reconstruction error

Figure 3 shows the reconstruction of the latent representations of four lexical tones in Mandarin using the trained VAE. The generated f_0 contours seem to have almost the same shape compared to the original data. It can also be observed that the generated f_0 contours preserve the subtle variations in the original f_0 contours.

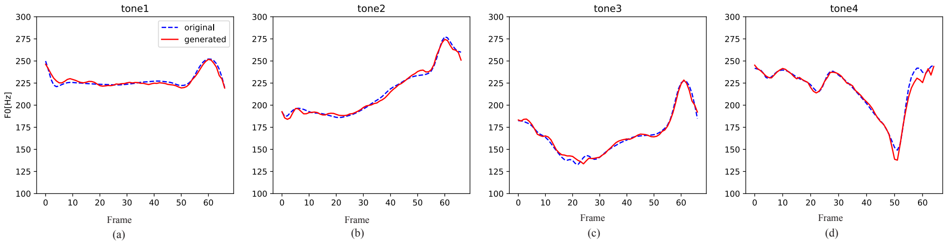

Figure 3. F_0 contours extracted from the original data (dashed blue line), and f_0 contours extracted from the reconstructed data (solid red line) obtained using our proposed generative model.

Figure 4 shows f_0 contours of the tone1-tone2 continuum. The pitch contours obtained by resampling in the latent space of the VAE have a smoother and more gradual transition between the two reference contours than those generated by the manual approach.

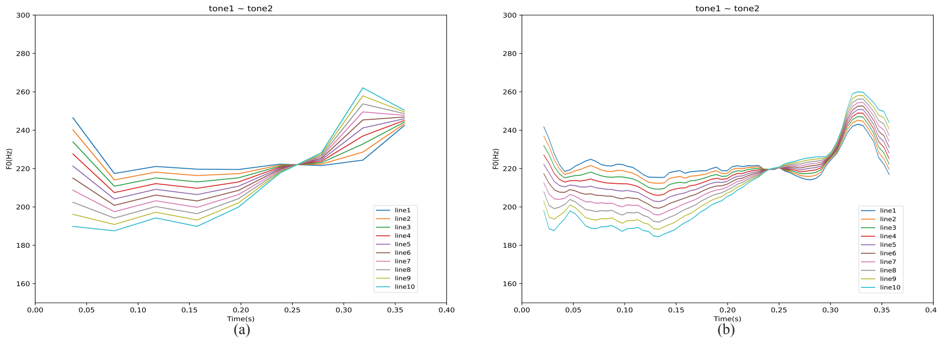

Figure 4. (a) f_0 contours of the tone1-tone2 continuum obtained by the traditional approach. (b) f_0 contours of the tone1-tone2 continuum obtained by resampling in the latent space of the trained VAE.

Figure 5 provides examples of the mel-spectrograms of training data and reconstructed data. The generated data were obtained using the VAE, which was trained with spectral envelopes. It can be seen that the reconstructed /re1/ and /le1/ preserve fine-grained details of spectral envelopes.

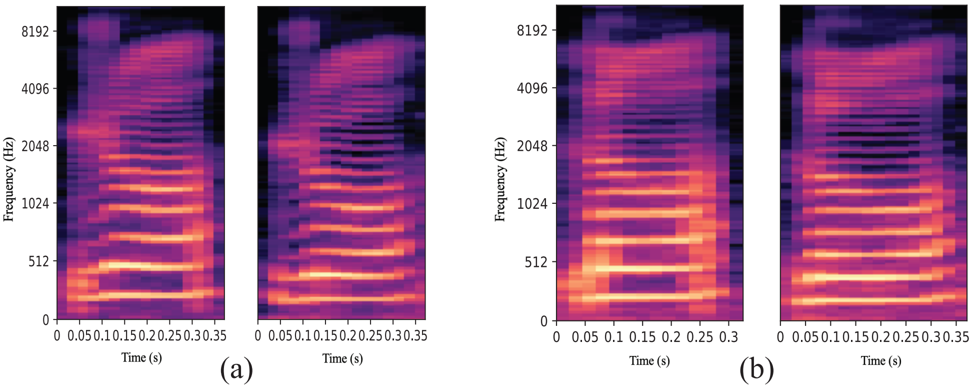

Figure 5. (a) the left subplot: original /re1/; the right subplot: reconstructed /re1/. (b) the left subplot: original /le1/; the right subplot: reconstructed /le1/.

Figure 6 illustrates the difference between the data-driven approach and manual manipulation of vocal tract parameters for stimulus continuum generation. It is noticeable that some tiny details between formant frequencies (green arrows in Figure 6) are

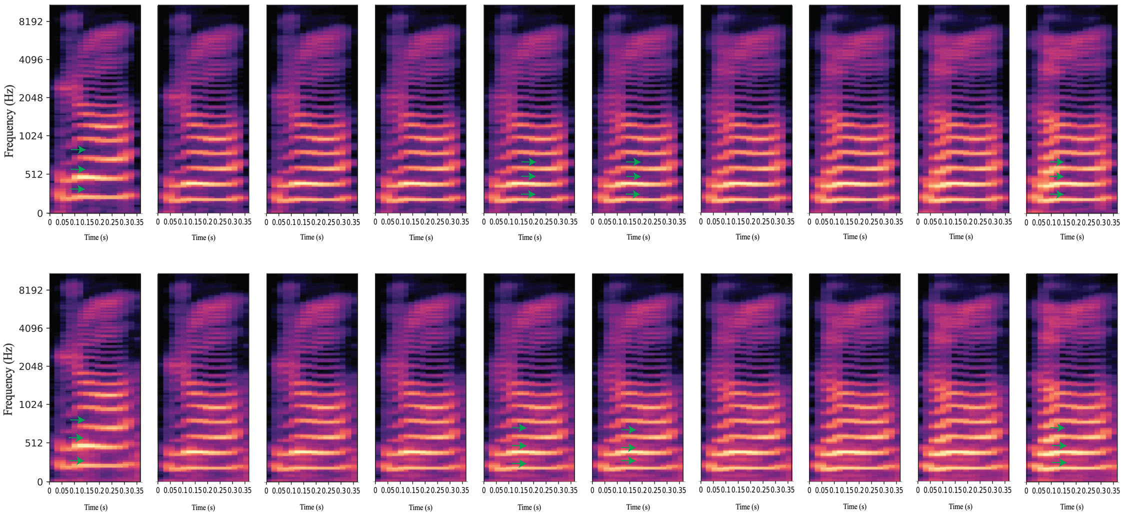

Figure 6. The top subplot shows mel-spectrograms of the /ʐ/-/l/ continuum obtained by manual manipulation; The bottom subplot shows mel-spectrograms of the /ʐ/-/l/ continuum obtained by resampling in the latent space.

obscured in the top mel-spectrograms, while this information is preserved in the bottom mel-spectrograms.

4.2. Subjective evaluation

Stimulus samples (9-interval tone1-tone2 continuum along the pitch dimension and /ʐ/-/l/ continuum along the vocal tract parameter dimension) were generated using our proposed method and manual manipulation. The subjective evaluation experiments were carried out via an online platform for behavioral research.

4.2.1. Identification test

To compare our proposed data-driven approach with the traditional method of directly adjusting the acoustic parameters, an identification experiment was conducted to explore whether differences exist in categorical boundary position and width using the stimulus continua created by these two techniques. Subjects were eight native speakers of Mandarin Chinese with a Mandarin level above 2A (Eight subjects participated in the tone1-tone2 perception study, and five of them participated in the /ʐ/-/l/ perception study). At the beginning of the test, two reference sounds (coded as "Sound 1" and "Sound 2" respectively) were played two times to participants, and they were instructed to familiarise themselves with the two representative sounds as best as possible. The stimulus samples of each continuum were presented to the participants randomly. Subjects were asked to press key "1" when they thought the sound was "Sound 1" or to press key "2" when they thought they had heard "Sound 2". The ten stimuli were played randomly in a block. There were five such testing blocks for each continuum generated by the two methods. Identification curves for the tone1-tone2 continuum and the /ʐ/-/l/ continuum are shown in Figure 7. The two curves show somewhat similar trends, especially for the tone1-tone2 continuum perception case. However, some subtle differences deserve further exploration. For example, for the /ʐ/-/l/ continuum perception case, the category boundary is closer to the middle stimulus when the stimuli generated by our VAE-based approach are used for the test. Also, there is a longer and more gradual curve for the transition part. One possible explanation is that transitions generated by the proposed method are smoother and more gradual. However, these patterns are not apparent in the tone1-tone2 example.

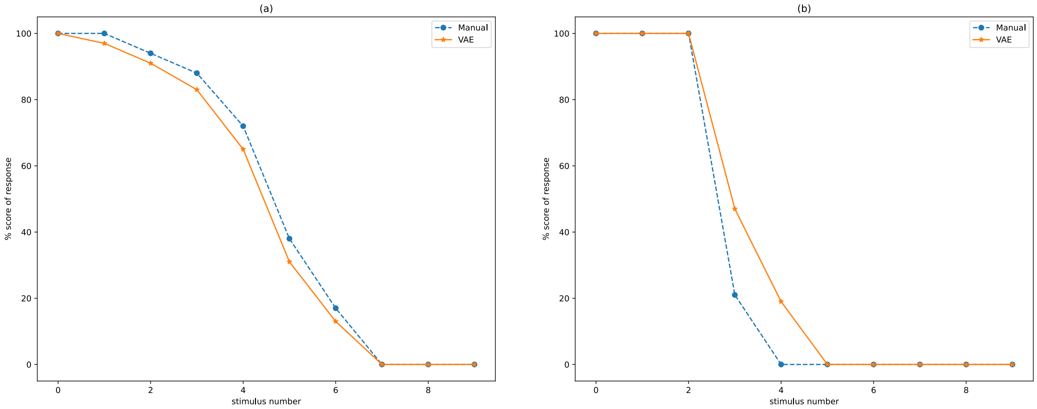

Figure 7. Identification curves pooled across participants. (**a**) Perception of the tone1-tone2 continuum. (**b**) Perception of the /ʐ/-/l/ continuum.

4.2.2. MOS evaluation

The overall quality of the stimulus samples generated by these two methods was evaluated using the mean opinion score (MOS). Eight native speakers of Mandarin Chinese were recruited, and none of them had participated in the previous experiment. Listeners were asked to rate the overall naturalness of the stimulus samples on a scale from 1 and 5. A total of 80 voice stimuli (8 continua) were mixed and presented to listeners in a randomized order. Based on the identification results, generated sounds were divided into within-category stimuli and between-category stimuli. For the tone continuum, the third to seventh stimuli were viewed as between-category. For the /ʐ/-/l/ continuum, the fourth and fifth stimuli were regarded as between-category. In accordance with this classification, the overall MOS, within-category MOS and between-category MOS were calculated. Tabel 3 summarizes the results of listeners' evaluation of the synthesized stimuli.

Table 3: MOS of stimulus samples

System	MOS (overall)	MOS (within category)	MOS (between category)
tone1-tone2 continuum (Manual)	3.81	4.21	3.32
tone1-tone2 continuum (VAE)	**3.92**	**4.18**	**3.47**
/ʐ/-/l/ continuum (Manual)	3.97	4.20	3.78
/ʐ/-/l/ continuum (VAE)	**4.06**	**4.25**	**3.89**

Pairwise comparisons using the paired Mann-Whitney U Tests [30] show that between-group differences in overall MOS are not significant ($p > 0.05$). This suggests that generally speaking, the quality of generated stimuli using the data-driven approach is comparable to that of the manual manipulation approach. Notably, the between-category MOS of the tone1-tone2 continuum and the between-category MOS of the /ʐ/-/l/ continuum based on the VAE model are slightly higher. These results indicate that both approaches can generate relatively natural stimulus samples with acceptable sound quality, but for those stimuli near the category boundary, the VAE-based method slightly improves the naturalness of generated speech over the manual manipulation baseline.

5. Conclusions

In this paper, we proposed a data-driven approach to generate stimulus continua based on VAEs. This work bridges the gap between stimulus continuum generation and state-of-the-art DL techniques. We used fundamental frequencies and vocal tract parameters to conduct stimulus continuum synthesis experiments using our proposed model. The results indicated that the proposed method can generate smoother and more gradual transitions between two endpoint reference stimuli, and yield more natural between-category stimuli compared to manual manipulation on the key acoustic feature.

Future directions include disentangling key acoustic features instead of using vocoders for feature extraction, and modeling on mel-spectrograms instead of directly modeling the acoustic features. In addition, we will experiment with using recurrent neural network architecture as the encoder to model speech sounds, so as to avoid possible information loss or distortion caused by time normalization. We will also try expanding the scale of perception experiments in order to obtain more convincing results. In addition to the perceptual experiments conducted in the current study, using other perceptual metrics, e.g., the perceptual evaluation of speech quality (PESQ) to compute the perceptual distance between intermediate stimuli is also a possible solution to evaluate the proposed approach.

Author Contributions: Conceptualization, Z.L. and Y.X.; methodology, Z.L. and Y.X.; validation, Z.L.; formal analysis, Z.L.; investigation, Z.L.; resources, Y.X.; writing—original draft preparation, Z.L. and Y.Z.; writing—review and editing, Z.L., Y.Z. and Y.X.; supervision, Y.X. and D.F.; funding acquisition, Y.X. and B.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by National Social Science Foundation of China (18BYY124), the Fundamental Research Funds for the Central Universities, and the Research Funds of Beijing Language and Culture University (21YCX167), Wutong Innovation Platform of Beijing Language and Culture University (19PT04), and the Education Reform Project of Beijing Language and Culture University (XJGZ 201904).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Acknowledgments: We thank all anonymous reviewers from conference NCMMSC2021 for their insightful comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. A. L. Francis, V. Ciocca, and B. K. C. Ng, "On the (non) categorical perception of lexical tones," *Perception & Psychophysics*, vol. 65, no. 7, pp. 1029–1044, 2003.
2. P. A. Hallé, Y.-C. Chang, and C. T. Best, "Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners," *Journal of phonetics*, vol. 32, no. 3, pp. 395–421, 2004.
3. W. Strange and S. Dittmann, "Effects of discrimination training on the perception of/rl/by Japanese adults learning English," *Perception & Psychophysics*, vol. 36, no. 2, pp. 131–145, 1984.
4. K. S. MacKain, C. T. Best, and W. Strange, "Categorical perception of English /r/and /l/by Japanese bilinguals," *Applied psycholinguistics*, vol. 2, no. 4, pp. 369–390, 1981.
5. V. G. Skuk and S. R. Schweinberger, "Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender," *Journal of speech, language, and hearing research: JSLHR*, vol. 57, no. 1, pp. 285–296, 2014.
6. M. Gubian, Y. Asano, S. Asaridou, and F. Cangemi, "Rapid and smooth pitch contour manipulation," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013, pp. 31–35.
7. M. Gubian, F. Cangemi, and L. Boves, "Automatic and data driven pitch contour manipulation with functional data analysis," in *Speech Prosody 2010-Fifth International Conference*, 2010.
8. T. White, "Sampling generative networks," *arXiv preprint arXiv:1609.04468*, 2016.
9. M. Thielk, T. Sainburg, T. Sharpee, and T. Gentner, "Combining biological and artificial approaches to understand perceptual spaces for categorizing natural acoustic signals," in *Conference on Cognitive Computational Neuroscience*, 2018, pp. 1–4.
10. T. Sainburg, M. Thielk, B. Theilman, B. Migliori, and T. Gentner, "Generative adversarial interpolative autoencoding: adversarial training on latent space interpolations encourage convex latent distributions," *arXiv preprint arXiv:1807.06650*, 2018.
11. D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
12. X. Liu, Y. Zou, L. Kong, Z. Diao, J. Yan, J. Wang, S. Li, P. Jia, and J. You, "Data augmentation via latent space interpolation for image classification," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 728–733.

13. Y.-C. Chen, X. Xu, Z. Tian, and J. Jia, "Homomorphic latent space interpolation for unpaired image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2408–2416.
14. Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of gans for semantic face editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9243–9252.
15. A. Roberts, J. Engel, C. Raffel, C. Hawthorne, and D. Eck, "A hierarchical latent vector model for learning long-term structure in music," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4364–4373.
16. I. Simon, A. Roberts, C. Raffel, J. Engel, C. Hawthorne, and D. Eck, "Learning a latent space of multitrack measures," *arXiv preprint arXiv:1806.00195*, 2018.
17. M. Tomczak, M. Goto, and J. Hockman, "Drum synthesis and rhythmic transformation with adversarial autoencoders," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2427–2435.
18. W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," *arXiv preprint arXiv:1704.04222*, 2017.
19. M. Blaauw and J. Bonada, "Modeling and transforming speech using variational autoencoders," *Morgan N, editor. Interspeech 2016; 2016 Sep 8-12; San Francisco, CA.: ISCA; 2016. p. 1770-4.*, 2016.
20. C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.
21. Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International conference on machine learning*. PMLR, 2017, pp. 933–941.
22. K. Tanaka, H. Kameoka, and K. Morikawa, "Vae-space: Deep generative model of voice fundamental frequency contours," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5779–5783.
23. W. Dai, J. Zhang, Y. Gao, W. Wei, D. Ke, B. Lin, and Y. Xie, "Formant tracking using dilated convolutional networks through dense connection with gating mechanism," *arXiv preprint arXiv:2005.10803*, 2020.
24. B. Wu, Y. Xie, L. Lu, C. Cao, and J. Zhang, "The construction of a Chinese interlanguage corpus," in *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 2016, pp. 183–187.
25. M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.
26. G. Peng, H.-Y. Zheng, T. Gong, R.-X. Yang, J.-P. Kong, and W. S.-Y. Wang, "The influence of language experience on categorical perception of pitch contours," *Journal of Phonetics*, vol. 38, no. 4, pp. 616–624, 2010.
27. T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6820–6824.
28. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.
29. H. Valbret, E. Moulines, and J.-P. Tubach, "Voice transformation using PSOLA technique," *Speech communication*, vol. 11, no. 2-3, pp. 175–187, 1992.
30. A. Rosenberg and B. Ramabhadran, "Bias and statistical significance in evaluating speech synthesis with mean opinion scores." in *Interspeech*, 2017, pp. 3976–3980.