VICNet: FaderNet-Based Voice Impression Conversion with Affective Dimensional Representation

Takuya Takahashi*, Saki Kugimoto*, and Toru Nakashika*
* The University of Electro-Communications, Japan
E-mail: {takahashi, s.kugimoto, nakashika}@uec.ac.jp

***Abstract*—This study proposes a voice impression conversion method based on voice impressions quantified using the valence-arousal-dominance (VAD) model commonly employed for emotional expression. Conventional techniques have limited voice impressions to discrete categorical representations, which presents challenges for versatility. In this study, we propose a method of quantifying voice impressions using a continuous three-dimensional space (VAD model) as a more versatile representation, and formulate VICNet, which enables voice impression conversion by incorporating the continuous impression representation and individual differences of listeners. In addition, a new voice impression dataset was constructed, where 100 human annotators assessed the voice impressions of 100 speakers from the JVS dataset in VAD space. Experimental evaluation demonstrated that utilising all VAD axes in the proposed model enabled clearer impression conversion; however, we also observed that impression conversion performance degraded when training data was limited for certain impression patterns.**

I. Introduction

Human voices are inherently unique and significantly influence interpersonal impressions. Recent advances in voice conversion technology have enabled several approaches: speaker conversion, which alters one's voice to resemble a specific person [1]–[3], and emotional conversion, which manipulates voice expression to convey particular emotions [4]–[6].

However, there is limited research on "voice impression conversion" that focuses solely on the impression conveyed by the voice itself, independent of speaking style or spoken content. As a first attempt, we studied FaderNetVIC, which realises voice impression conversion for seen speakers by replacing the speaker labels of FaderNetVC with impression labels [7]. Moreover, by introducing an external classifier that estimates impression labels from voice into FaderNetVIC, we have also achieved voice impression conversion for unseen speakers who do not have impression labels [8]. However, since the impression labels were defined as discrete categorical expressions based on adjectives related to impressions, these approaches could only convert voices into pre-defined impressions and had limitations in terms of versatility.

This study addresses the limitations of discrete categorical approaches by proposing a method for quantifying voice impressions using continuous dimensional models. Since impressions and emotions share similar perceptual properties despite being fundamentally different concepts, we adapted the Valence-Arousal-Dominance (VAD) model, which was originally proposed by Mehrabian and Russell [9] for emotional representation in three-dimensional continuous space, to quantify voice impressions. Based on this framework, we constructed a new voice dataset labelled with VAD-based impression annotations by collecting data from 100 human participants through a crowdsourcing service. We also extend FaderNetVIC to construct a new voice impression conversion model, VICNet, that incorporates both continuous dimensional representations for voice impressions and individual differences among listeners.

II. A Dataset of Voice Impressions Based on Affective Dimensional Representation

A. Quantification of voice impression

Since voice impression is a subjective phenomenon, it is reasonable to have humans listen to voices and label their impressions. Kido *et al.* [10] investigated everyday expressions related to voice impressions in normal speech and found that voice impressions can be described by seven primary pairs (high-pitched - low-pitched, hoarse - clear, calm - agitated, powerful - weak, thick - thin, resonant - flat, nasal - non-nasal) of expressions using statistical methods. Okadome *et al.* [7] and Kugimoto *et al.* [8] proposed a method to evaluate the voice impression of these seven expression pairs on a 7-point Likert scale, and constructed a subjective dataset of voice impressions from 100 speakers in VoxCeleb1 [11].

While the method to quantify voice impressions proposed by Kido *et al.* [10] is intuitively understandable, it does not necessarily cover all voice impressions. We believe that voice impressions cannot be adequately expressed by discrete categories and should be quantified as continuous variables.

In this study, we hypothesised that voice impressions could be quantified using the VAD model [9], which is one of the affective dimensional models. In the VAD model [9], "valence" represents the pleasantness or unpleasantness of emotion, "arousal" represents the activity level or energy of emotion, and "dominance" represents the sense of control or influence in emotion. The VAD model enables the numerical expression of complex emotional states and is widely used in emotion research with media [12]–[15]. Notably, we discovered that the seven primary expression word pairs describing

979-8-3315-7206-8/25/$31.00

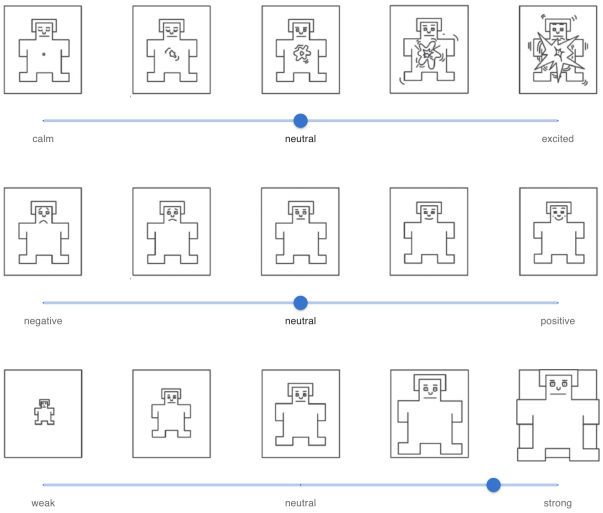

Fig. 1: Graphical interface for annotating voice impressions. After listening to the audio, participants annotate continuous values on sliders with Self-Assessment Manikin figures [18].

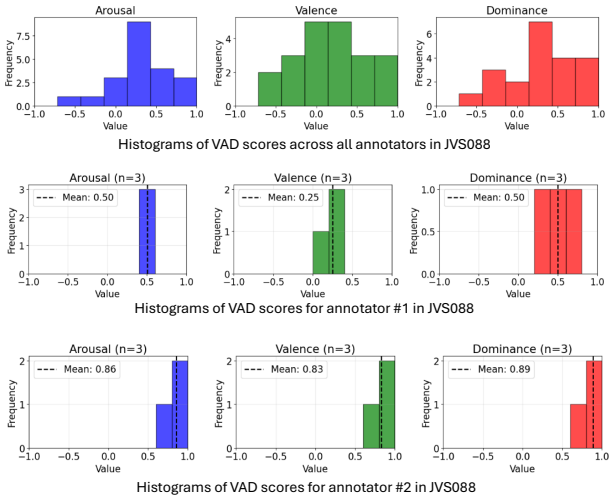

Fig. 2: Histograms of VAD scores for all annotators and specific annotators in speaker JVS088.

voice impressions identified by Kido *et al.* [10] can also be represented within the VAD model, according to Warriner *et al.*'s subjective investigation of VAD values for 13,915 English lemmas [16]. Therefore, it is expected that using the VAD model to quantify voice impressions will enable more diverse and flexible quantification than discrete categorical approaches.

B. Collection of a dataset for quantifying voice impressions using the VAD model

We proposed a unique method for collecting data in which humans subjectively evaluated voice impressions using the VAD model, and constructed a dataset of voice impressions for 100 speakers from the JVS dataset [17]. To the best of our knowledge, no published datasets exist that quantify voice impressions using the VAD model.

As audio data for annotators to evaluate, we prepared 4,200 short audio clips by randomly selecting 42 clips from each of the 100 speakers in the JVS dataset [17]. To evaluate voice impressions without being influenced by content or speaking style, we extracted short voiced sound segments (approximately 1 second) from the JVS dataset [17], a Japanese speech corpus, and had them evaluated by non-native Japanese speakers to ensure that linguistic understanding did not affect the voice impression assessments.

100 human annotators recruited online via Amazon Mechanical Turk[1] labelled 42 randomly selected clips using three scales of valence, arousal, and dominance with sliders ranging from -1 to +1 on the website[2]. Since previous emotion labeling experiments [19] have reported labeling difficulties due to unclear scale representations, we presented adjective pairs at both ends of each scale that best represent the respective dimensions, along with Self-Assessment Manikin (SAM) [18] figures for visual representation, as shown in Figure 1, following the approach of Soares et al. [19].

[1]https://www.mturk.com/
[2]http://sp.lab.uec.ac.jp:3000/adj3

C. Analysis of the constructed dataset

Figure 2 shows histograms of VAD scores for all annotators and specific annotators in speaker JVS088. Although there was variation between annotators (top of Figure 2), the distribution shape was unimodal. In contrast, we found that the variability in evaluations within annotators tends to be consistent, as can be seen in the centre and bottom of Figure 2.

The mean standard deviation between annotators was 0.395, while the mean standard deviation within annotators was 0.194 across the entire dataset. Since the variation within annotators is considerably smaller, this suggests that while voice impressions vary depending on the annotator, individual annotators evaluate particular speakers relatively consistently. Therefore, when performing voice impression conversion, it is necessary to construct a conversion model that accounts for individual subjective variability.

III. FaderNet for Voice Impression Conversion Based on VAD Model

A. Voice impression conversion models

In this study, we extended the model proposed by Kugimoto *et al.* [8] and conducted an initial investigation of VICNet, which applies the VAD model and takes into account the subjectivity of each annotator. The model structure of the proposed method is shown in Figure 3. Similar to FaderNetVIC, we employ an encoder-decoder model with mel-cepstrum input and output, introducing an adversarial discriminator that identifies voice impressions so that the latent space $\mathbf{z}$ represents features other than voice impressions. The decoder was trained with target impression labels estimated by an external classifier and latent representations, excluding voice impressions as input,

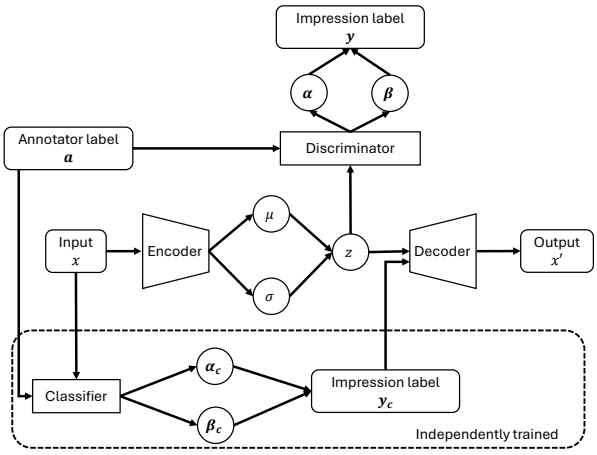

Fig. 3: VICNet architecture for voice impression conversion. Impression labels are represented as three-dimensional vectors based on the VAD model.

producing mel-cepstrum as output In addition, by inputting annotator labels into the discriminator and external classifier, it is possible to learn and infer impression labels that consider the individual subjective differences of each annotator. Since VAD values have upper and lower bounds, similar to Kugimoto *et al.* [8], we use beta distribution parameters as outputs of the discriminator and external classifier to explicitly model classification uncertainty and expect the model to learn the distributional characteristics of each scale per speaker.

B. Loss function of the VICNet

In this model, the encoder, decoder and discriminator (main part) and the external classifier are trained independently to avoid interference between the adversarial learning process and the impression label estimation task, ensuring stable convergence for both components.

The main part consists of an encoder E_θ, a decoder D_ϕ, and a discriminator A_ψ, and is trained as an adversarial optimisation problem as

$$\min_{\theta,\phi} \max_{\psi} \mathcal{L}(\theta, \phi, \psi), \tag{1}$$

where $\mathcal{L}(\theta, \phi, \psi)$ is the loss function of the main part. Let the mel-cepstrum of the input audio be denoted as $\mathbf{x} \in \mathbb{R}^{D\times T}$ (where D is the feature dimensionality and T is the number of frames), the annotator label as $\mathbf{a} \in \mathbb{R}^{100}$, the impression label as $\mathbf{y} \in \mathbb{R}^K$ with dimensionality K, the number of training samples as n, and the training dataset as $G = \{(\mathbf{x}_i, \mathbf{a}_i, \mathbf{y}_i) : i = 1, 2, \ldots, n\}$, then the loss function is expressed as:

$$\begin{aligned}\mathcal{L}(\theta, \phi, \psi) = \frac{1}{n} \sum_{(\mathbf{x},\mathbf{a},\mathbf{y})\in G} \{&\lambda_{\mathrm{MSE}}\mathcal{L}_{\mathrm{MSE}}(\mathbf{x}, \mathbf{y}) - \lambda_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}}(\boldsymbol{x}) \\ &- \lambda_{\mathrm{ADV}}\mathcal{L}_{\mathrm{ADV}}(\mathbf{x}, \mathbf{a}, \mathbf{y})\},\end{aligned} \tag{2}$$

where $\lambda_{\mathrm{MSE}}, \lambda_{\mathrm{ADV}}$ and λ_{KL} are trade-off parameters that adjust the balance of each loss (all empirically set to 1). With the encoder function $E_\theta(\mathbf{x})$ that outputs $\boldsymbol{\mu}_\theta(\boldsymbol{x}), \boldsymbol{\sigma}_\theta(\boldsymbol{x}) \in \mathbb{R}^{64}$ and samples the latent variable $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_\theta(\boldsymbol{x}), \boldsymbol{\sigma}^2_\theta(\boldsymbol{x}))$ and the decoder function $D_\phi(\mathbf{z}, \mathbf{y})$ that calculates the reconstructed $\mathbf{x}'$, the reconstruction loss $\mathcal{L}_{\mathrm{MSE}}$ can be formulated as

$$\mathcal{L}_{\mathrm{MSE}}(\mathbf{x}, \mathbf{y}) = ||D_\phi(E_\theta(\mathbf{x}), \mathbf{y}) - \mathbf{x}||_2^2. \tag{3}$$

The constraint term $\mathcal{L}_{\mathrm{KL}}(\boldsymbol{x})$ that brings the normal distribution of the encoder output closer to the standard normal distribution, i.e., the KL term of the VAE, is defined as

$$\mathcal{L}_{\mathrm{KL}}(\boldsymbol{x}) = \frac{1}{2} \left\| \mathbf{1} + \log(\boldsymbol{\sigma}_\theta(\boldsymbol{x})^2) - \boldsymbol{\mu}_\theta(\boldsymbol{x})^2 - \boldsymbol{\sigma}_\theta(\boldsymbol{x})^2 \right\|_1. \tag{4}$$

Let K be the dimensionality of the label, $\boldsymbol{\alpha}(\mathbf{x}, \mathbf{a}) \in \mathbb{R}_+^K$ and $\boldsymbol{\beta}(\mathbf{x}, \mathbf{a}) \in \mathbb{R}_+^K$ be the beta distribution parameters output by the discriminator. The loss function of the discriminator corresponding to the k-dimensional impression label is

$$\begin{aligned}\mathcal{L}_{ADV}(\mathbf{x}, \mathbf{a}, \mathbf{y}) = \sum_{k=1}^{K} \{&\log\Gamma(\alpha_k(\mathbf{x}, \mathbf{a})) + \log\Gamma(\beta_k(\mathbf{x}, \mathbf{a})) \\ &+ (1 - \alpha_k(\mathbf{x}, \mathbf{a})) \log(\frac{y_k + 1}{2}) \\ &+ (1 - \beta_k(\mathbf{x}, \mathbf{a})) \log(\frac{1 - y_k}{2}) \\ &- \log\Gamma(\alpha_k(\mathbf{x}, \mathbf{a}) + \beta_k(\mathbf{x}, \mathbf{a}))\},\end{aligned} \tag{5}$$

where the beta distribution is transformed to match the y_k domain, which is $[-1, 1]$, although its standard domain is $[0, 1]$.

The classifier is trained independently from the main part, but its loss function is similar to that of the discriminator Let $\boldsymbol{\alpha'}(\mathbf{x}, \mathbf{a}) \in \mathbb{R}_+^K$ and $\boldsymbol{\beta'}(\mathbf{x}, \mathbf{a}) \in \mathbb{R}_+^K$ be the beta distribution parameters output by the classifier, and let C represent the combination of the classifier function outputs $\boldsymbol{\alpha'}(\mathbf{x}, \mathbf{a})$ and $\boldsymbol{\beta'}(\mathbf{x}, \mathbf{a})$:

$$\begin{aligned}\mathcal{L}_C(\mathbf{x}, \mathbf{a}, \mathbf{y}) &= \log p(\mathbf{y}|\mathbf{x}, \mathbf{a}) \\ &= \sum_{k=1}^{K} \{\log\Gamma(\alpha'_k(\mathbf{x}, \mathbf{a})) + \log\Gamma(\beta'_k(\mathbf{x}, \mathbf{a})) \\ &\quad + (1 - \alpha'_k(\mathbf{x}, \mathbf{a})) \log(\frac{y_k + 1}{2}) \\ &\quad + (1 - \beta'_k(\mathbf{x}, \mathbf{a})) \log(\frac{1 - y_k}{2}) \\ &\quad - \log\Gamma(\alpha'_k(\mathbf{x}, \mathbf{a}) + \beta'_k(\mathbf{x}, \mathbf{a}))\}.\end{aligned} \tag{6}$$

Since the derivative of $\log\Gamma$ is the digamma function, the gradient can be calculated using a built-in function.

The classifier and discriminator output $\alpha(\mathbf{x}_{\mathrm{ref}}, \mathbf{a}_{\mathrm{ref}})$ and $\beta(\mathbf{x}_{\mathrm{ref}}, \mathbf{a}_{\mathrm{ref}})$ from the input audio $\mathbf{x}_{\mathrm{ref}}$ and the annotator label $\mathbf{a}_{\mathrm{ref}}$, and determine the impression label $\mathbf{y}_c$ using the expected value of the normalized beta distribution:

$$\mathbf{y}_c = \mathbb{E}_{p(\mathbf{y}|\mathbf{x}_{\mathrm{ref}}, \mathbf{a}_{\mathrm{ref}})}[\mathbf{y}] = \frac{\boldsymbol{\alpha}(\mathbf{x}_{\mathrm{ref}}, \mathbf{a}_{\mathrm{ref}}) - \boldsymbol{\beta}(\mathbf{x}_{\mathrm{ref}}, \mathbf{a}_{\mathrm{ref}})}{\boldsymbol{\alpha}(\mathbf{x}_{\mathrm{ref}}, \mathbf{a}_{\mathrm{ref}}) + \boldsymbol{\beta}(\mathbf{x}_{\mathrm{ref}}, \mathbf{a}_{\mathrm{ref}})}, \tag{7}$$

where the fraction bar represents element-wise division. Additionally, by using an average $\mathbf{a}_{\mathrm{ref}}$ over all annotators, such as setting all elements of $\mathbf{a}_{\mathrm{ref}} = 1/(\text{number of annotators})$, it is also possible to convert voice impression that does not depend on individual annotator differences.

IV. Experimental Evaluation

A. Experimental setup

In this chapter, we used the dataset constructed in Chapter II to train the voice impression conversion model discussed in Chapter III. The voice impressions of the converted speech generated by the trained model were evaluated objectively and subjectively.

From the VAD-labelled voice dataset described in Chapter II, speech from 13 speakers with particularly small variance in annotation values was used for the experiments, with training, validation, and test data split at 80%, 10%, and 10%, respectively. The audio sampling rate was 24 kHz for JVS, with a shift width of 5 ms, and the input consisted of 32-dimensional, 32-frame mel-cepstrum segment features. The generated mel-cepstrum was converted to audio using the World vocoder [20].

Since no previous work has applied the VAD model to speech impression conversion, we evaluated our method by comparing different configurations within our proposed approach. To conduct an ablation study on which individual dimensions of the VAD model contribute most effectively to voice impression conversion performance, training was performed for the following four models: the V model, A model and D model trained using only valence, arousal and dominance respectively, and the VAD model trained using all three dimensions — valence, arousal, and dominance.

B. Performance of classifier and discriminator

After training the models, the accuracy of impression label estimation by the external classifier and discriminator was compared. For the four model types — V model, A model, D model, and VAD model — accuracy was calculated as the cosine similarity between impression labels estimated from test speech by either the external classifier or discriminator and the groundtruth impression labels predefined for the test speech, because the VAD model is a circumplex model.

The results shown in Table I indicate that the A, D, and VAD models achieved higher classification accuracy, while the V model showed notably lower accuracy. Although direct comparison between individual models (V, A, D) and the VAD model is not entirely fair due to dimensional differences, the relatively high similarity scores on the VAD model indicate that joint training across all three dimensions can potentially improve estimation accuracy.

Such difficulty in valence classification has yielded similar results in the field of emotion recognition [21], [22]. Valence is strongly related to the linguistic content of speech [23] and is considered difficult to capture using only acoustic features [24]. On the other hand, a report [25] indicated that spectral features and F_0 features are useful for valence classification in acoustic features. Since only mel-cepstrum was input to the model, adding features such as F_0 could potentially improve the estimation accuracy of valence.

TABLE I: Cosine similarity of classifier and discriminator.

model	Classifier	Discriminator
V model	0.250	0.313
A model	0.781	0.781
D model	0.844	0.688
VAD model	0.703	0.621

C. A/B test of impression identification

1) Experimental setup for A/B test: In this test, target impressions are presented to participants as text, and participants are asked to select from options A and B the voice that is closer to the target impression, thereby evaluating the extent to which the impression-converted speech generated by the proposed method can express the target impression. In the following, VAD-based impression labels will be expressed in the form of a tuple with three values, such as $(\text{valence}, \text{arousal}, \text{dominance}) = (0.5, 0.8, -0.1)$ where the domain of each scale was $[-1, 1]$.

We prepared 7 pairs of converted voices by pairing the target impression $(1, 1, 1)$ with each of 7 different comparison impressions. These comparison impressions were categorized into three groups based on the number of dimensions that differed from the target: **single-dimension (SD) manipulation group** with one dimension changed $\{(-1, 1, 1), (1, -1, 1), (1, 1, -1)\}$, **multiple-dimension (MD) manipulation group** with two dimensions changed $\{(1, -1, -1), (-1, 1, -1), (-1, -1, 1)\}$, and **full-dimension (FD) manipulation group** with all three dimensions changed $\{(-1, -1, -1)\}$, and each pair was randomly placed on A/B. We set such extreme target impressions to create clear differences between comparison speech sounds and test whether the model can effectively learn each axis.

As text representations of the impressions to be selected, the impression words "excited" for $\text{arousal} = 1$, "positive" for $\text{valence} = 1$, and "strong" for $\text{dominance} = 1$ were used. For example, for a pair of audio generated from target impressions $(1, -1, 1)$ and $(1, 1, 1)$, participants would be asked to select the audio that matches the text "Choose the audio from A or B that is closest to the impression of excited", and the correct answer would be the audio generated from the target impression of $(1, 1, 1)$. However, the V, A, and D models, trained on their respective single dimensions, were tested only on impression conversions that involved changes in the corresponding dimension.

Participants were 100 people recruited through Amazon Mechanical Turk[1], and each person completed a 54-question survey. After excluding participants who failed to properly respond to attention check questions designed to verify participant engagement and response quality inserted throughout the survey, 78 valid responses remained for analysis. The following results are based on these 78 valid responses.

2) Experimental results and discussion: The results of the accuracy rate for the target impression are shown in Table II. While the voice impression converted by the V, A, and D models was below the chance rate of 0.5, the voice impression converted by the VAD model exceeded the chance rate of 0.5, except for the case of $(-1, -1, -1)$. These results suggest that considering all VAD dimensions simultaneously rather than

TABLE II: Results of subjective evaluation experiments. Column 1: Proportion of samples from (valence, arousal, dominance) = (1,1,1) selected in A/B tests against target impressions shown in the rows. Column 2: Cosine similarity between target impressions and human re-annotations of converted speech samples (range: -1 to 1, higher values indicate greater similarity).

	(valence, arousal, dominance)	AB test accuracy (section IV.C)	cosine similarity (section IV.D)
Single-dimension (SD) manipulation group	(-1, 1, 1) on V model	0.374	-0.696
	(1, -1, 1) on A model	0.400	-0.711
	(1, 1, -1) on D model	0.451	-0.674
Single-dimension (SD) manipulation group	(-1, 1, 1) on VAD model	0.687	0.337
	(1, -1, 1) on VAD model	0.654	0.250
	(1, 1, -1) on VAD model	0.651	0.299
Multiple-dimension (MD) manipulation group	(1, -1, -1) on VAD model	0.672	-0.197
	(-1, 1, -1) on VAD model	0.628	-0.102
	(-1, -1, 1) on VAD model	0.654	-0.121
Full-dimension (FD) manipulation group	(-1, -1, -1) on VAD model	0.359	-0.693

using them separately can facilitate the training of voice and impression labels and allow more accurate impressions to be applied to speech.

The V model had the lowest correct answer rate, likely due to its low classification accuracy shown in Section IV.B. The low estimation accuracy for $(-1, -1, -1)$ in the VAD model has resulted from the difficulty of answering questions about whether the audio matched the intersection of "positive", "excited" and "strong". Moreover, Figure 4 shows the data distribution across VAD scales in our dataset. All scales exhibit fewer data points with negative values compared to positive values. This data imbalance likely causes the reduced accuracy in impression expression for negative VAD values.

In this way, by using the VAD model as an impression representation, it is possible to effectively manipulate impressions for each dimension. However, this experiment tested only extreme values on each axis, yielding limited results that serve to verify the effective functioning of each axis. Further experiments are needed to determine the feasibility of more detailed impression conversion.

D. Re-VAD annotation of speech after impression conversion

1) Experimental setup for Re-annotation: We conducted an experiment in which participants listened to the audio converted using the proposed method with the same target impressions as in the previous section and re-annotated the VAD using the same method as in Chapter II. By comparing the cosine similarity between the target impression and the annotations, we can quantitatively evaluate the extent to which VICNet expresses the target impression similar to IV.B.

The participants were 100 people different from those who participated in the AB test in the previous section, and were recruited through Amazon Mechanical Turk[1]. Each person answered 54 questions and performed VAD annotations using an interface similar to that shown in Figure 1.

2) Experimental results and discussion: The cosine similarities between the target impression and the re-annotated impression are shown in the second column of Table II. The V, A, and D models showed significantly lower similarity than the VAD model. This result indicates that using all VAD dimensions for training can achieve more accurate voice impression conversion.

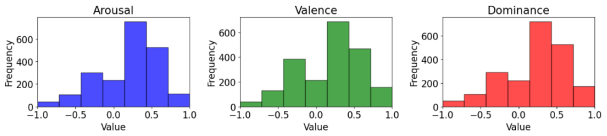

Fig. 4: Histograms showing the number of data points at each scale for all data in the dataset constructed in Chapter II.

In the VAD model, similarity exceeded 0 for the SD manipulation group, while lower similarity scores were observed for the other two groups. These results suggest that smaller VAD values lead to a decrease in impression expressiveness, and the data imbalance shown in Figure 4 likely causes the limitations of the current model.

V. Conclusions

In this paper, we proposed VICNet, a novel continuous voice impression conversion method that applied the Valence-Arousal-Dominance (VAD) model. We constructed a unique dataset of 100 speakers from the JVS corpus with VAD-based impression annotations collected from 100 human annotators, addressing the limitations of previous discrete categorical approaches. Experiments showed that all VAD dimensions yielded the highest impression conversion accuracy, though arousal and dominance correlated more strongly with voice impressions than valence.

Future work should address data imbalance for negative VAD values, explore additional acoustic features (particularly for valence estimation), conduct cross-corpus validation, and implement more comprehensive evaluation experiments with intermediate VAD values and detailed error analysis to validate the robustness and effectiveness of the proposed approach. Furthermore, the proposed network architecture is applicable not only to speech impression conversion but also to music domains such as emotion/impression-based music generation and timbre impression conversion, and we plan to explore these possibilities in future work.

Acknowledgment

This research was partly funded by JSPS Grants-in-Aid for Scientific Research 24H00715 and JSPS Special Fellowship 24KJ1125.

References

[1] A. Sayadian and F. Mozaffari, "A novel method for voice conversion based on non-parallel corpus," *International Journal of Speech Technology*, vol. 20, pp. 587–592, 2017.

[2] Y. Xinyuan and B. Mak, "Non-parallel many-to-many voice conversion by knowledge transfer from a text-to-speech model," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5924–5928.

[3] T. Kuwabara and T. Nakashika, "Voice gender conversion with speaker identity preservation based on FaderNets," *Dissertation at University of Electro-Communications*, 2020.

[4] T. Qi, W. Zheng, C. Lu, Y. Zong, and H. Lian, "PAVITS: Exploring prosody-aware VITS for end-to-end emotional voice conversion," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12 697–12 701.

[5] H. Tang, X. Zhang, J. Wang, N. Cheng, and J. Xiao, "EmoMix: Emotion mixing via diffusion models for emotional speech synthesis," *arXiv preprint arXiv:2306.00648*, 2023.

[6] Y. Xu, H. Chen, J. Yu, *et al.*, "SECap: Speech emotion captioning with large language model," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 19 323–19 331.

[7] Y. Okadome and T. Nakashika, "Voice impression transformation based on FaderNets with impression expression word labels," *Dissertation at University of Electro-Communications*, 2022.

[8] S. Kugimoto and T. Nakashika, "Voice impression conversion for unseen speakers using FaderNets," *Proceedings of the Spoken Language Processing in Information Processing Society of Japan*, vol. 2023, no. 21, pp. 1–4, 2023.

[9] A. Mehrabian and J. A. Russell, "The basic emotional impact of environments," *Perceptual and motor skills*, vol. 38, no. 1, pp. 283–301, 1974.

[10] H. Kido and H. Kasuya, "Everyday expressions associated with vioce quiality of normal utterance -extraction by perceptual evaluation-," *The Journal of the Acoustical Society of Japan*, vol. 57, no. 5, pp. 337–344, 2001.

[11] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[12] R. Parke, E. Chew, and C. Kyriakakis, "Quantitative and visual analysis of the impact of music on perceived emotion of film," *Computers in Entertainment (CIE)*, vol. 5, no. 3, p. 5, 2007.

[13] M. Barthet, G. Fazekas, and M. Sandler, "Music emotion recognition: From content-to context-based models," in *the 9th International Symposium on Computer Music Multidisciplinary Research*, 2013, pp. 228–252.

[14] H. H. Tan and D. Herremans, "Music FaderNets: Controllable music generation based on high-level features via low-level feature modelling," *arXiv preprint arXiv:2007.15474*, 2020.

[15] T. Takahashi and M. Barthet, "Emotion-driven harmonisation and tempo arrangement of melodies using transfer learning.," in *International Society for Music Information Retrieval Conference*, 2022, pp. 741–748.

[16] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior research methods*, vol. 45, pp. 1191–1207, 2013.

[17] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: Free Japanese multi-speaker voice corpus," *arXiv preprint arXiv:1908.06248*, 2019.

[18] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[19] A. P. Soares, A. P. Pinheiro, A. Costa, C. S. Frade, M. Comesaña, and R. Pureza, "Affective auditory stimuli: Adaptation of the international affective digitized sounds (IADS-2) for european portuguese," *Behavior research methods*, vol. 45, pp. 1168–1181, 2013.

[20] 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.

[21] M. Grimm, K. Kroschel, and S. Narayanan, "Support vector regression for automatic recognition of spontaneous emotions in speech," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, 2007, pp. IV–1085.

[22] M. Asif, N. Ali, S. Mishra, A. Dandawate, and U. S. Tiwary, "Deep fuzzy framework for emotion recognition using eeg signals and emotion representation in type-2 fuzzy vad space," *arXiv preprint arXiv:2401.07892*, 2024.

[23] Z. Aldeneh, S. Khorram, D. Dimitriadis, and E. M. Provost, "Pooling acoustic and lexical features for the prediction of valence," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 68–72.

[24] B. T. Atmaja and M. Akagi, "Dimensional speech emotion recognition from speech features and word embeddings by using multitask learning," *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, vol. 9, e17, 2020.

[25] C. Busso and T. Rahman, "Unveiling the acoustic properties that describe the valence dimension.," in *Interspeech*, 2012, pp. 1179–1182.