BWSNET: AUTOMATIC PERCEPTUAL ASSESSMENT OF AUDIO SIGNALS

Clément Le Moine Veillon[⋆1]*, Victor Rosi*[⋆2]*, Pablo Arias Sarah*[3]*, Léane Salais*[1]*, Nicolas Obin*[1]

[1]STMS Lab - IRCAM, CNRS, Sorbonne Université, Paris, France
[2]Department of Speech Hearing and Phonetic Sciences, University College London, London, UK
[3]School of Psychology and Neuroscience, University of Glasgow, Glasgow, UK

ABSTRACT

This paper introduces BWSNet, a model that can be trained from raw human judgements obtained through a Best-Worst scaling (BWS) experiment. It maps sound samples into an embedded space that represents the perception of a studied attribute. To this end, we propose a set of cost functions and constraints, interpreting trial-wise ordinal relations as distance comparisons in a metric learning task. We tested our proposal on data from two BWS studies investigating the perception of speech social attitudes and timbral qualities. For both datasets, our results show that the structure of the latent space is faithful to human judgements.

Index Terms— Automatic Perceptual Assessment, Best-Worst Scaling, Metric Learning, Social Attitudes, Timbre

1. INTRODUCTION

Access to a perceptual representation of data typically involves an experiment in which stimuli are judged by human participants according to a specific criterion. In particular, several methods to subjectively assess audio stimuli have been proposed, including pairwise comparison, MUSHRA and rating scales, the latter being widely praised in experimental psychology. With the advent of synthesis algorithms and the subsequent need to finely assess their outputs, such rating scales are widely employed to complement, or even replace, objective measures such as MCD or RMSE that are not necessarily correlated with human perception [1]. Typically, the Mean Opinion Score (MOS) has been used to assess speech synthesis models' performance by asking participants to rate the quality or naturalness of output samples [2] and their similarity to a reference, according to specific speech attributes, e.g., speaker identity [2, 3], emotion [4], or attitude [5]. Although favoured, scale-based methods present potential biases [6, 7], making the choice of evaluation method an important issue in experimental psychology.

Recently, attention was given to Best-Worst Scaling (BWS) [8], another method in which participants are presented with trials of N (e.g. N=4,5) items and asked to judge which ones are the best and worst according to a studied attribute. Once the experiment completed, each item receives a score - computed using more or less elaborate techniques on the basis of raw judgements - representing how it is perceived in regards with this attribute. BWS has proven to yield more reliable results than rating scales to gather perceptual scores [9, 10, 11] and has been found effective for various audio-related tasks such as the perception of speech emotions [4] and attitudes [12] as well as timbral qualities [13].

Unfortunately, all these subjective assessment methods always require a large number of human ratings to be reliable, making them costly and time-consuming. Facing this challenge, an entire field of research has emerged with the aim of automating perceptual evaluation. Thus, several methods have been proposed to learn a regression model that predicts MOS scores such as AutoMOS [14], Quality-Net [15], and MOSNet [16, 17, 18]. To our knowledge and despite its aforementioned benefits, there is no existing method for predicting BWS judgements.

In this paper, we introduce BWSNet, a model for automatic perceptual assessment based on BWS data. In contrast with existing MOSNet approaches, we do not seek to predict perceptual scores in a regression task. Indeed, the determination of BWS scores entails a dimensional reduction of the perceptual space underlying raw judgements, which involves a potential loss of information. To predict these judgements, that infer ordinal relations between items in each trial, we design a metric learning task in which these relations (see Fig. 1) are interpreted as distances comparisons. BWSNet is thus trained to learn a function that maps sound samples into a latent space in which the distance represents samples dissimilarity with respect to the studied attribute.

We present two contributions, firstly BWSNet, which consists of a set of cost functions adapted to the ordinal and relative nature of BWS judgements. Secondly, we apply it in the specific context of two studies previously led by the authors, investigating the perception of speech social attitudes [12] and timbral concepts [13].

2. BWSNET

2.1. Problem Positioning

A BWS trial is a tuple of N sounds $t^a = \{x^1, ..., x^N\}$ set for judgement. For each trial, a judgment consists of choos-

ing the best and worst items in the tuple, the $N-2$ remaining items are then considered neutrals. We denote $\mathcal{T}^a$ the set containing all the trials considered for the BWS experiment that investigates a.

Each sample is renamed with respect to the trial it lies in and the judgement it has been assigned. The best, worst and neutrals of trial t^a can be indexed as t^a_b, t^a_w and $t^a_{n_i}$ with $i \in \{1, N-2\}$, respectively. We denote $\succ_a$, the relation such that $x \succ_a y$ is equivalent to "x is more perceived as a than y". Then, t^a judgement is represented by $2(N-1)$ relations - ordinal in nature - between sample triplets, as shown on the right part of Fig. 1.

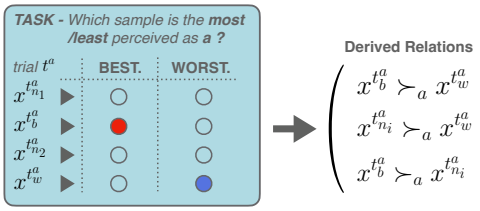

Fig. 1. A BWS trial $t^a \in \mathcal{T}^a$ of $N = 4$ sounds judged with respect to the attribute a (left) and the derived relations (right).

2.2. Concept of a BWSNet

Our goal is to design a model able to learn perceptual representations underlying human BWS judgements. We use mel-spectrograms as sound representations, as it has proven to account for many perceptually relevant aspects of speech. As shown in Fig. 2, for any sample x, the BWSNet takes a mel-spectrogram $\mathbf{X}$ as input and produce a BWS embedding $\mathbf{h}^x$. Relative positioning of embeddings, in the latent space they lie in, must be true to BWS judgments. To achieve this, ordinal relations between sound samples in BWS trials are translated into distances comparisons of the corresponding embeddings. Thus, introducing a distance $\|.\|: \mathbb{R}^d \to \mathbb{R}$, the relations for a trial t^a displayed in Fig. 1 can be translated into the following inequalities for $i \in \{1, ..., N-2\}$:

$$\|\mathbf{h}^{t^a_b} - \mathbf{h}^{t^a_w}\| \geq \|\mathbf{h}^{t^a_b} - \mathbf{h}^{t^a_{n_i}}\| \quad (1)$$

$$\|\mathbf{h}^{t^a_b} - \mathbf{h}^{t^a_w}\| \geq \|\mathbf{h}^{t^a_w} - \mathbf{h}^{t^a_{n_i}}\| \quad (2)$$

2.3. Losses and Optimisation

To train the model to match trials relations within its latent space, we designed our own training criterion, a cost function inspired by the metric learning literature and notably the triplet loss proposed in [19].

The relations to match are ordinal and, like in triplet loss, each one involves three items in a trial as formalized in inequalities 1 and 2. However, they only prevail within a given trial, making the problem even more complex. To avoid mode collapse, i.e., the model turning all samples into one single point in the latent space, aforementioned inequalities must be considered strict. We thus introduce a positive margin α and define the RC loss $\mathcal{L}_{t^a}$ for any trial $t^a \in \mathcal{T}^a$ as:

$$\mathcal{L}^{t^a}_{rc} = \frac{1}{n^{t^a}_v} \sum_{i=1}^{N-2} \max\left(\|\mathbf{h}^{t^a_b} - \mathbf{h}^{t^a_{n_i}}\| - \|\mathbf{h}^{t^a_b} - \mathbf{h}^{t^a_w}\| + \alpha, 0\right) + \frac{1}{n^{t^a}_v} \sum_{i=1}^{N-2} \max\left(\|\mathbf{h}^{t^a_w} - \mathbf{h}^{t^a_{n_i}}\| - \|\mathbf{h}^{t^a_b} - \mathbf{h}^{t^a_w}\| + \alpha, 0\right) \quad (3)$$

where $n^{t^a}_v$ is the number of relations that remain to be fulfilled within trial t^a.

Perceptual differences represented in BWS trials can be more or less substantial, quantifying them requires dynamic margins rather than fixed ones. We thus introduce a network $\mathcal{M}$ dedicated to margin learning. It takes two parameters as arguments, a mean μ and an amplitude δ value, such that learnt margin lies between $\mu - \delta$ and $\mu + \delta$. With all trials' embeddings in the batch as input, it produces $N-2$ distinct margins $\{\alpha_{b,n_i}, \alpha_{w,n_i}\}_{i \in \{1,N-2\}}$ related to each trial relation. This impacts the RC loss formulated in equation 3 as it takes learnt margins as additional input. Thus, for a given trial t^a, the Dynamic margin (Dm)-RC loss can be expressed as follows:

$$\mathcal{L}^{t^a}_{drc} = \frac{1}{n^{t^a}_v} \sum_{i=1}^{N-2} \max\left(\|\mathbf{h}^{t^a_b} - \mathbf{h}^{t^a_{n_i}}\| - \|\mathbf{h}^{t^a_b} - \mathbf{h}^{t^a_w}\| + \alpha_{b,n_i}, 0\right) + \frac{1}{n^{t^a}_v} \sum_{i=1}^{N-2} \max\left(\|\mathbf{h}^{t^a_w} - \mathbf{h}^{t^a_{n_i}}\| - \|\mathbf{h}^{t^a_b} - \mathbf{h}^{t^a_w}\| + \alpha_{w,n_i}, 0\right) \quad (4)$$

We assume a Gaussian distribution of margins and propose an additional constraint to penalize our model's tendency to learn low-value margins. This constraint is formalized through a function γ that takes the learned margins as an argument and outputs a scalar loss. Various functions can be tested, resulting in different learnt margin distributions. For a trial t^a, the Dynamic Margin Constraint (DMC) can be formulated as:

$$\mathcal{L}^{t^a}_{dmc} = \sum_{i=1}^{N-2} \gamma(\alpha_{b,n_i} - \mu) + \gamma(\alpha_{w,n_i} - \mu) \quad (5)$$

Decrease in the Dm-RC loss does not guarantee an increase in the number of fulfilled relations. Since margins can decrease overall without affecting order relationships. To avoid this, we have added a final loss directly derived from the Dm-RC loss which accounts, for a given trial t^a, to the number of unfulfilled relations within the trial divided by the number N of elements in the trial.

$$\mathcal{L}^{t^a}_{fr} = \frac{n^{t^a}_v}{N} \quad (6)$$

We introduce the scalars λ_{dmc}, $\lambda_{fr} \geq 0$ and define the global BWSNet loss for a trial t^a as:

$$\mathcal{L}^{t^a} = \mathcal{L}^{t^a}_{dmrc} + \lambda_{dmc}\mathcal{L}^{t^a}_{dmc} + \lambda_{fr}\mathcal{L}^{t^a}_{fr} \quad (7)$$

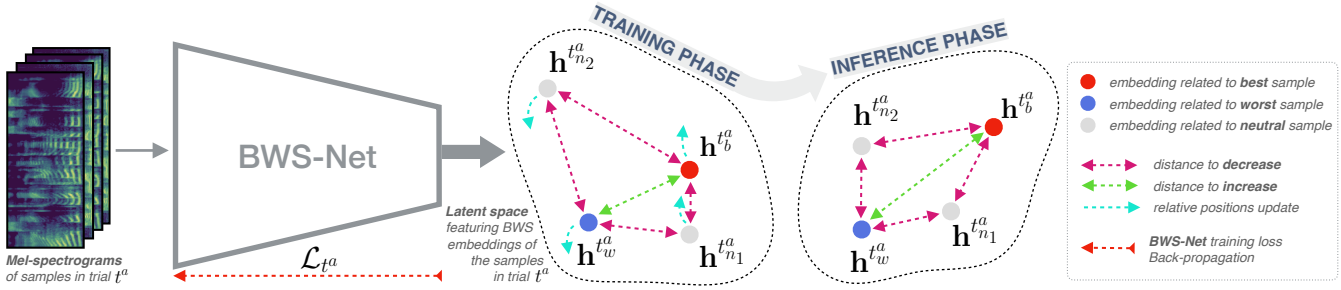

Fig. 2. The Mel-Spectrograms of the $N = 4$ samples related to a BWS trial t^a investigating the attribute a are passed to BWSNet. The model yields BWS embeddings of which relative position is changed over training.

3. EXPERIMENTS

We used data from two previous BWS studies to train and test our model. We provide short descriptions of these in the following.

3.1. Data for Experiment

3.1.1. Study I: Speech Social Attitude

We carried out this first study on Att-HACK [20], a 30-hour speech dataset composed of 9 male and 11 female actors portraying four different social attitudes - *friendliness*, *dominance*, *distance* and *seduction* - over 100 sentences in French. 96 participants (48 women) were recruited to evaluate the perception of its related sounds for each a priori attitude, i.e. the ones already produced with aim of conveying this specific attitude. Model trainings were thus conducted on the four attitudes separately. For the sake of feasibility, only a fourth of the dataset has been assessed with each sound being evaluated eight times.

3.1.2. Study II: Instrumental Timbre

With the purpose of giving acoustic portraits of timbre concepts, sound expert participants (N=16, sound engineers and conductors) evaluated a dataset of musical instruments (N=520) on four well-known timbral concepts, namely *brightness*, *warmth*, *roundness* and *roughness*. The sounds showcase a great diversity of instruments, playing techniques, dynamics and registers. Unlike Study I, each sample is evaluated with respect to all concepts.

3.2. Implementation Details

3.2.1. Input Pipeline

Mel-Spectrograms are obtained through computing Short-Term-Fourier-Transform (STFT) with FFT 2048, hop 200, window 800 and 80 mel filters. Our custom RC loss takes two tensors as input, in addition to the BWS embeddings: a tensor made of the corresponding trial names - which ensures the loss is computed trial-wise even if several trials lie in a single batch - and a tensor made of each element's corresponding judgement labels (b, w, n).

3.2.2. Architecture Design

The BWSNet architecture was inspired by a promising version of ACRNN [21] initially proposed for speech emotion recognition. We previously assessed the role of its various components for speech attitude recognition in [12] using the same attitudinal dataset [20]. Based on this study, the best configuration used 2 convolutional layers with 64 filters, temporal and feature kernel of sizes 5 and 3 respectively, and an 8 head attention mechanism with 512 dimensions. Here, we reuse the same architecture for our task - which is close to the one for which it has proved optimal - without further experimenting. Finer adjustments to the architecture, e.g. latent space dimension d, could improve performance depending on the input data, especially for timbral concepts. Here we choose $d = 32$ and focus on the relative importance of the aforementioned cost functions and constraints.

3.2.3. Training Procedure

In both experiments, we split data into three groups, first we selected 10% of the samples and dropped out any trial they were involved in. Then, we split the remaining data trial-wise into training (80%) and validation (20%) trial sets. We fed our model with batches of size 80 with as many data of each concept for Experiment II and used ADAM optimizer with 0.0001 as learning rate. We found the euclidean distance for $\|.\|, \gamma : x \to ReLU(-x)$ for DMC and $\mu = \gamma = 1$ as margin learning module $\mathcal{M}$ parameters to yield the best results.

3.2.4. Evaluation Criteria

To evaluate the performance of BWSNet, we used two metrics, reflecting the arrangement of speech samples in the latent space at two levels: **FR** and **WAT** - respectively the percentages of fulfilled relations and well-arranged trials within the set - both computed on relations from dropped trial set, involving at least one unseen sample. We chose lowest **FR** value as early stopping criterion.

4. RESULTS & DISCUSSION

We carried out an ablation study to demonstrate the influence of each cost function and constraint on the model's performance. Then, we sought to explore the latent spaces yielded by BWSNet for both studies.

4.1. Ablation Study

Table 1 displays **FR** and **WAT** mean and standard deviation values for various BWSNet configurations evaluated on unseen samples across both BWS studies' sets of attributes. First, as expected, the fixed-margin configuration (**A-f**) did not generalize well, as some trials' best and worst could be either very distant in the latent space or rather close. Then, when considering learnt margin configurations, it appeared that with no constraint on margins (**A-l**) the latent space was collapsing into one single point, turning distances between any pair of points null. Adding such a constraint (**A-l-d**) tended to help the model converge and fulfill just under one relations in two on attitudes and slightly more than one relations in two for timbre, which is insufficient to claim that our model accurately predicts BWS judgements. However, we found that BWSNet alternatively uses two strategies to lower Dm-RC loss: it could seek to fulfill more relations within trials otherwise it could reduce the margins. The addition of FR loss as training criterion (**A-l-d-fr**) appeared to prevent the model from engaging in the second strategy and led to improvements by 27.7% and 27.5% in both **FR** and **WAT** respectively for attitudes and by 4.7% in **FR** for timbre. The model performed better for attitudinal than for timbral data, which may be due to the choice of neural architecture that we deliberately adapted to the former.

			Study I: Attitudes		*Study II: Instrument Timbre*	
Model	λ_{dmc}	λ_{fr}	**FR** (%)	**WAT** (%)	**FR** (%)	**WAT** (%)
A-f	-	-	21.4 ± 8.5	5.8 ± 3.6	37.5 ± 2.5	4.7 ± 0.5
A-l	0	0	1.0 ± 0.4	0.2 ± 0.0	26.1 ± 2.1	$\mathbf{26.0 \pm 1.3}$
A-l-d	1	0	40.1 ± 18.7	22.4 ± 17.1	51.6 ± 3.5	19.9 ± 2.6
A-l-d-fr	1	1	$\mathbf{67.7 \pm 4.5}$	$\mathbf{49.9 \pm 8.9}$	$\mathbf{56.3 \pm 2.4}$	23.9 ± 1.2

Table 1. **FR** & **WAT** mean and standard deviation values for various BWSNet configurations evaluated on unseen samples across both BWS studies' sets of attributes.

4.2. Exploring BWSNet's Latent Space

To further analyse our results, we investigated the relation between item scores, obtained with the original scoring algorithm [10] used in both studies, with our latent space's dimensions. Figure 3 shows the BWSNet latent spaces corresponding to each attitude (left) and timbre qualities (right) with the original scores characterised with color and size respectively.

By assessing attitudes independently in the BWS experiment, we obtained four distinct latent spaces with different degrees of polarisation with regard to BWS scores. For instance, for *friendliness* and *seductiveness* - that were found the most consensual [12] - high and low score samples are located in opposite ends of the space. Furthermore, regarding the latent spaces obtained for *distance* and *dominance*, BWS scores may be an inaccurate depiction of their multidimensional nature.

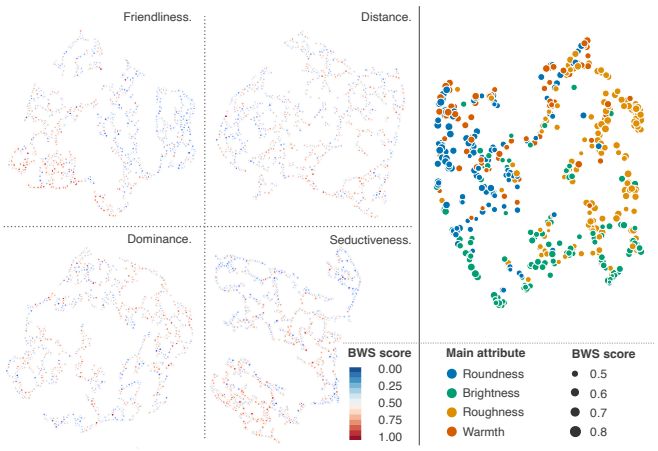

Fig. 3. BWSNet's latent space UMAP vizualization for social attitudes (left) and timbral qualities (right). Each point is a sample whose BWS score is represented by its colour (left) and size (right) respectively.

As for timbral concepts, the original study sought to uncover their mutual interactions which leads to have all samples lying in the same space. To report on these interactions, we associated each sound with its most salient attribute (i.e., the highest scored concept for each sound). We observe similar interactions to those in the original study, e.g. *warmth* and *roundness* are blended together in the timbral space while showing strong opposition to *brighthness*.

5. CONCLUSION

This paper presents BWSNet, a model dedicated to automatic audio perceptual assessment based on BWS judgements. Distances between learnt sample embeddings in the resulting latent space represent their perceptual similarity. By fulfilling almost 70% of relations involving an unseen sample for attitudinal speech data, BWSNet provides a rather accurate estimation of how this sample is perceived based on its distance with previously judged samples in the latent space. Its performance on timbral data (56% fulfilled relations) also indicates potential for application to a manifold of judgement tasks. These results, obtained on two very different datasets, mark a first step towards automating the BWS perceptual assessment. Furthermore, as the analysis of its latent space suggests, BWSNet could be a relevant tool for representing and understanding human perception. Provided that participants agree to a certain extent, using more BWS judgements for training would likely yield a more comprehensive map of the perceptual space for a chosen sound attribute.

6. REFERENCES

[1] Lucas Theis, Aäron van den Oord, and Matthias Bethge, "A note on the evaluation of generative models," *arXiv preprint arXiv:1511.01844*, 2016.

[2] Mingyang Zhang, Xin Wang, Fuming Fang, Haizhou Li, and Junichi Yamagishi, "Joint training framework for text-to-speech and voice conversion using multi-source tacotron and wavenet," *arXiv preprint arXiv:1903.12389*, 2019.

[3] Hirokazu Kameoka, Kou Tanaka, Damian Kwaśny, Takuhiro Kaneko, and Nobukatsu Hojo, "Convs2s-vc: Fully convolutional sequence-to-sequence voice conversion," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 28, pp. 1849–1863, 2020.

[4] Kun Zhou, Berrak Sisman, Rajib Rana, Björn W Schuller, and Haizhou Li, "Speech synthesis with mixed emotions," *IEEE Transactions on Affective Computing*, 2022.

[5] Clément Le Moine, Nicolas Obin, and Axel Roebel, "Towards end-to-end F0 voice conversion based on Dual-GAN with convolutional wavelet kernels," in *EUSIPCO*, Dublin (virtual), Ireland, 2021.

[6] Hans Baumgartner and Jan-Benedict EM Steenkamp, "Response styles in marketing research: A cross-national investigation," *Journal of marketing research*, vol. 38, no. 2, pp. 143–156, 2001.

[7] Howard Schuman and Stanley Presser, *Questions and answers in attitude surveys: Experiments on question form, wording, and context*, Sage, 1996.

[8] Jordan J. Louviere, Terry N. Flynn, and A. A. J. Marley, *Best-Worst Scaling: Theory, Methods and Applications*, Cambridge University Press, 2015.

[9] Svetlana Kiritchenko and Saif Mohammad, "Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, Canada, July 2017, pp. 465–470, Association for Computational Linguistics.

[10] Geoff Hollis, "Scoring best-worst data in unbalanced many-item designs, with applications to crowdsourcing semantic judgments," *Behavior research methods*, vol. 50, no. 2, pp. 711–729, 2018.

[11] Victor Rosi, Aliette Ravillion, Olivier Houix, and Patrick Susini, "Best-worst scaling, an alternative method to assess perceptual sound qualities," *The Journal of the Acoustical Society of America*, vol. 2, pp. 064404, 06 2022.

[12] Clément Le Moine Veillon, *Neural Conversion of Social Attitudes in Speech Signals*, Ph.D. thesis, 2023, Thèse de doctorat dirigée par Roebel, Axel et Obin, Nicolas Informatique Sorbonne université 2023.

[13] Victor Rosi, Pablo Arias Sarah, Olivier Houix, Nicolas Misdariis, and Patrick Susini, "Shared mental representations underlie metaphorical sound concepts," *Scientific Reports*, vol. 13, no. 1, pp. 5180, 2023.

[14] Brian Patton, Yannis Agiomyrgiannakis, Michael Terry, Kevin Wilson, Rif A. Saurous, and D. Sculley, "Automos: Learning a non-intrusive assessor of naturalness-of-speech," in *NIPS 2016 End-to-end Learning for Speech and Audio Processing Workshop*, 2016.

[15] Szu-Wei Fu, Yu Tsao, Hsin-Te Hwang, and Hsin-min Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," 09 2018, pp. 1873–1877.

[16] Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang, "MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion," in *Proc. Interspeech*, 2019, pp. 1541–1545.

[17] Yeunju Choi, Youngmoon Jung, and Hoirin Kim, "Deep MOS Predictor for Synthetic Speech Using Cluster-Based Modeling," in *Proc. Interspeech*, 2020, pp. 1743–1747.

[18] Yeunju Choi, Youngmoon Jung, and Hoirin Kim, "Neural MOS prediction for synthesized speech using multi-task learning with spoofing detection and spoofing type classification," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 462–469.

[19] Elad Hoffer and Nir Ailon, "Deep metric learning using triplet network," in *Similarity-Based Pattern Recognition*, Aasa Feragen, Marcello Pelillo, and Marco Loog, Eds., Cham, 2015, pp. 84–92, Springer International Publishing.

[20] Clément Le Moine and Nicolas Obin, "Att-HACK: An Expressive Speech Database with Social Attitudes," in *Speech Prosody*, Tokyo, Japan, 2020.

[21] Yuanchao Li, Tianyu Zhao, and Tatsuya Kawahara, "Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning," in *Proc. Interspeech*, 2019, pp. 2803–2807.