BWSNet: Automatic Perceptual Assessment of Audio Signals

Clement 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 Science, University College London, London, UK

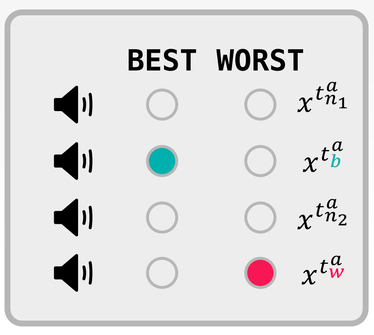
3 School of Psychology and Neuroscience, University of Glasgow, Glasgow, UK

Context

Revealing the **perceptual representation** of any sensory object usually requires conducting an experiment with human participants evaluating stimuli, based on a specific criterion. Recently, attention was given to **Best-Worst Scaling**[1] (BWS) as a method to assess perceptual qualities of sounds[2,3]. In this paper, we introduce **BWSNet**, a model for automatic perceptual assessment based on BWS data in a **metric learning** task.

From BWS trial t^a...

*« Which samples are perceived as <u>most</u> and <u>least</u> **A**? »*



Derived ordinal relations[4]

$$\begin{cases} x^{t^a_b} \succ_a x^{t^a_w} \\ x^{t^a_b} \succ_a x^{t^a_{n_i}} \\ x^{t^a_{n_i}} \succ_a x^{t^a_w} \end{cases}$$

$x \succ_a y$:
x is more **A** than y

... To distances

$\mathbf{h}^x$: **BWS embedding**

Inequality **(1)**

$$\left\|\mathbf{h}^{t^a_b} - \mathbf{h}^{t^a_w}\right\| \geq \left\|\mathbf{h}^{t^a_b} - \mathbf{h}^{t^a_{n_i}}\right\|$$

Inequality **(2)**

$$\left\|\mathbf{h}^{t^a_b} - \mathbf{h}^{t^a_w}\right\| \geq \left\|\mathbf{h}^{t^a_w} - \mathbf{h}^{t^a_{n_i}}\right\|$$

Metric learning losses and optimisations

- **Relative Contrastive loss with dynamic margins** $\mathcal{L}^{t^a}_{drc}$

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

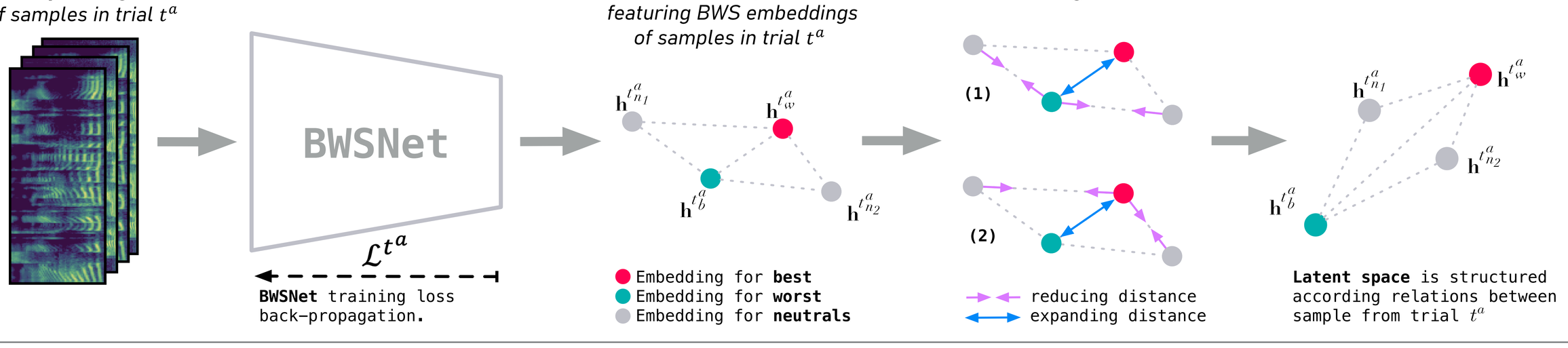
- **Dynamic margin constraint** $\mathcal{L}^{t^a}_{dmc}$

$$\mathcal{L}^{t^a}_{dmc} = \sum_{i=1}^{N-2}\gamma(\alpha_{b,n_i} - \mu) + \gamma(\alpha_{w,n_i} - \mu)$$ → *penalise the model's tendency to learn low-value margins.*

- **Fulfilled relations constraint** $\mathcal{L}^{t^a}_{frc}$

$$\mathcal{L}^{t^a}_{frc} = \frac{n^{t^a}_v}{N}$$ → *Ensures that a decrease of $\mathcal{L}^{t^a}_{dmc}$ corresponds to more fulfilled relations in t^a.*

Global BWSNet loss → $\mathcal{L}^{t^a} = \mathcal{L}^{t^a}_{drc} + \lambda_{dmc}\,\mathcal{L}^{t^a}_{dmc} + \lambda_{frc}\,\mathcal{L}^{t^a}_{frc}$ $\quad \lambda_{dmc}, \lambda_{frc} \geq 0$



Application 1: Vocal Attitudes[2]

Dataset

30-hour speech dataset showcasing four social attitudes: *friendliness*, *dominance*, *distance* and *seductiveness* in French.

BWS experiment

96 participants evaluated four distinct subsets (N=2400) corresponding to each intended attitude.

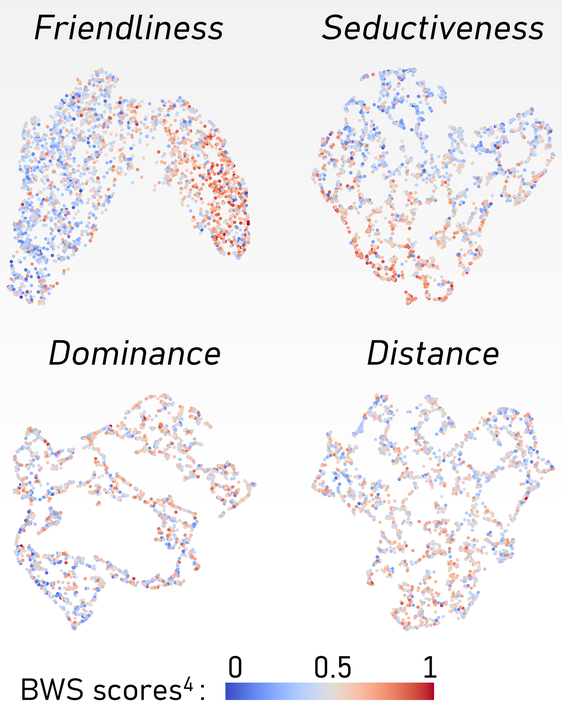
Metric learning task

One perceptually structured latent space *per attitude/dataset* with BWSNet.

Ablation study

Model	λ_{dmc}	λ_{frc}	FR* (%)
A-fixed	-	-	21.4 ± 8.5
A-l	0	0	1.0 ± 0.4
A-l-d	1	0	40.1 ± 18.7
A-l-d-fr	1	1	$\mathbf{67.7 \pm 4.5}$

Latent space



Application 2: Timbre Attributes[3]

Dataset

Orchestral sounds dataset (N=520) showcasing multiple instruments and playing techniques.

BWS experiment

16 sound experts evaluated the dataset on four timbral attributes: *brightness, warmth, roundness, and roughness*

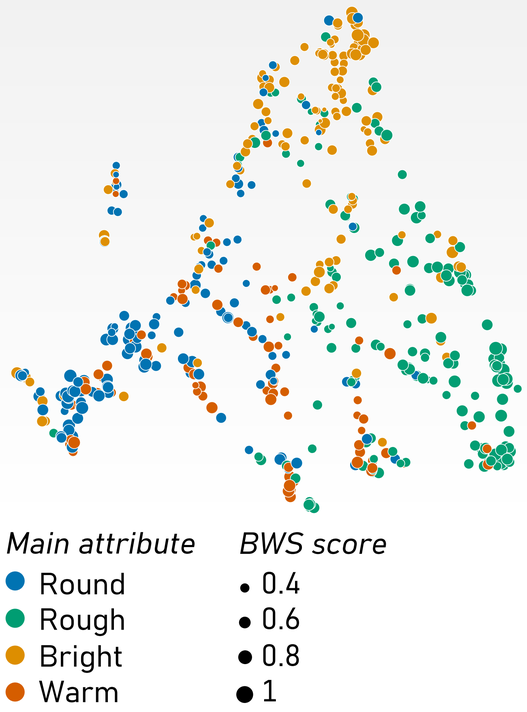
Metric learning task

One perceptually structured latent space *for all attributes* with BWSNet.

Ablation study

Model	λ_{dmc}	λ_{frc}	FR* (%)
A-fixed	-	-	37.5 ± 2.5
A-l	0	0	26.1 ± 2.1
A-l-d	1	0	51.6 ± 3.5
A-l-d-fr	1	1	$\mathbf{56.3 \pm 2.4}$

Latent space



FR*: Fulfilled relations | **A-fixed**: fixed-margin configuration | **A-l**: no margin constraint | **A-l-d**: margin constraint ($\mathcal{L}^{t^a}_{dmc}$) | **A-l-d-fr**: margin constraint ($\mathcal{L}^{t^a}_{dmc}$) and fulfilled relations constraint ($\mathcal{L}^{t^a}_{frc}$)

Conclusion

- BWSNet, a novel method designed for automated audio perceptual assessment based on human judgements.
- Perceptual organisation of latent spaces for vocal attitudes and timbral attributes.
- Perspectives of enhanced analyses of perceptual data by leveraging the dimensions of the latent space.
- Possible applications for conditioning synthesis/conversion models of speech and musical data.

References

[1] Louviere, J. J., Flynn, T. N., & Marley, A. A. J. (2015). *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

[2] Le Moine, C. (2023). *Neural conversion of social attitudes in speech signals* (Doctoral dissertation, sorbonne université).

[3] Rosi, V., Arias Sarah, P., Houix, O., Misdariis, N., & Susini, P. (2023). *Shared mental representations underlie metaphorical sound concepts*. Scientific Reports, *13*(1), 5180.

[4] Hollis, G., & Westbury, C. (2018). *When is best-worst best? A comparison of best-worst scaling, numeric estimation, and rating scales for collection of semantic norms*. Behavior research methods, *50*, 115-133.

ircam Centre Pompidou · MINISTÈRE DE LA CULTURE Liberté Égalité Fraternité · cnrs · UCL · UNIVERSITY of GLASGOW · ICASSP 2024 KOREA