Evaluating context-invariance in unsupervised speech representations

Mark Hallap1, Emmanuel Dupoux2, Ewan Dunbar1,2

1 University of Toronto
2 CoML, ENS/CNRS/EHESS/INRIA/PSL Research University
mark.hallap@utoronto.ca, emmanuel.dupoux@gmail.com, ewan.dunbar@utoronto.ca

Abstract

Unsupervised speech representations have taken off with benchmarks demonstrating major progress on semi-supervised speech recognition, speech synthesis, and speech-only language modelling. Inspiration comes from the promise of discovering the phonemes of a language or a similar low-bitrate encoding. However, one of the critical properties of phoneme transcriptions is context-invariance: the phonetic context of a speech sound can have massive influence on the way it is pronounced while text remains stable. This is why tokens of the same word have the same transcriptions—key to language understanding. Current benchmarks do not measure context-stability. We develop a new version of the ZeroSpeech ABX benchmark that does, and apply it to recent self-supervised representations. We show that context-independence of representations is predictive of the stability of word-level representations. We suggest research concentrate on improving context-independence of unsupervised representations.

Index Terms: pre-trained acoustic models, self-supervised speech, unsupervised speech, invariance

1. Introduction

Self-supervised and unsupervised learning of speech representations have their roots in attempts to discover phone- or phoneme-like units, or similarly linguistically relevant representations, either in the interest of working with lower-resource languages or of cognitive modelling [1]. These representations have already shown practical success. The use of recent models such as CPC [2], Wav2vec 2.0 [3], HuBERT [4], or WavLM [5] for pre-training has been shown to greatly reduce the amount of labelled speech data needed to build a recognizer. But the promise of self-supervised speech representations is much greater. In principle, if we were truly able to discover low-bitrate representations for a language similar to phonemes or letters, higher-level tasks could be done directly from speech. Currently, however, the performance of self-supervised representations on even mid-level tasks such as word segmentation lags behind that of text, to say nothing of higher-level tasks such as language modelling [6].

What makes text or phoneme representations so fundamental in language processing is their dual nature, linking signifier (form) and signified (content), which otherwise bear no meaningful relation. A language might use any string of sounds to refer to canines, but the transcription of the English word dog refers to canines, but the transcription of the English word dog is linguistic-ally insignificant variations in the signal and pinpoint those necessary for understanding.

Critically, phonemes change pronunciation substantially depending on their surrounding phoneme context due to coarticulation and allophony. Yet, current evaluations of intrinsic quality for unsupervised and self-supervised representations do not directly measure whether representations remain the same when the surrounding context changes. In this paper, we propose a novel evaluation based on the ABX phone error rate which directly assesses context-invariance.

In Experiment 1, we evaluate systems submitted to the 2021 Zero Resource Speech Challenge and demonstrate that effects of context are the most significant source of instability for current models—much greater than lack of invariance to speaker or resistance to less-clean speech. In Experiment 2, we demonstrate that this result is not an artefact of the models’ warping of time. Finally, in Experiment 3, we demonstrate that context-invariance predicts representations’ ability to consistently encode tokens of the same word type.

2. Background

In addition to benchmarking on downstream tasks (SUPERB: [7]), a large part of self-supervised speech representation evaluation consists of measures of intrinsic quality. Linguistically-motivated measures of intrinsic quality have been proposed as part of the ZeroSpeech challenge [6]. Notably, the ABX phone discriminability score [8] attempts to measure, for a model trained on a given language, how distinctly the model represents the linguistically relevant sound (phoneme) categories.

The task is inspired by human psychophysics and measures discriminability between two sound categories. \( \Delta \), the ABX-discriminability of sound category \( A \) from category \( B \), is defined as the probability that tokens \( a, x \in A \) are further apart than token \( b \in B \) is from \( x \), according to a dissimilarity function \( d \). Thus \( \Delta(A, B) \) is

\[
\Delta(A, B) = \frac{\sum_{a \in A} \sum_{b \in B} \sum_{x \in A} \mathbb{1}_{d(a, x) < d(b, x)} + \frac{1}{2} \sum_{a \in A} \mathbb{1}_{d(a, x) = d(b, x)}}{|A|(|A| - 1)/2}.
\]

where \( \mathbb{1} \) is the indicator function and \(|A| (|B|)\) the number of tokens in category \( A (B) \). The discriminability score is symmetrized by averaging \( \Delta(A, B) \) and \( \Delta(B, A) \).

Evaluating a model for a language begins with the model’s encoding for a test corpus. Using the gold alignment, the evaluation splits the encoding into one sequence of frames for each token of each category. To calculate \( d \) for two tokens, dynamic time warping is used to realign them, and frame-level similarities are averaged along the alignment path. Submissions to

1. The \( d(a, x) = d(b, x) \) clause is particularly important if the representations are symbolic. They need not be, and here for many systems it is continuous embeddings that are used: but the ABX score is compatible with both discrete and continuous representations.
the ZeroSpeech challenge specify whether to use angular dis-
similarity (arccos of the normalized dot product of frame em-
beddings) or KL divergence to calculate \( \Delta \); see [6]. \( \Delta \) (discrim-
innability) for all pairs of categories are averaged and subtracted
from 1 to obtain an overall ABX error rate.

Importantly, however, in the ZeroSpeech ABX phone eval-
uation, each token is a triphone: to compare two phoneme cat-
egories (for example, /l/ and /r/) we use triphones where only
the middle phoneme varies in the critical phone contrast—like
/lul–/flu/. The immediate context is thus included in the to-
ken, and is constant across \( a, b, \) and \( x \). This was originally to
avoid relying on very short, one-phone sequences and to avoid
demanding that the representations meet the (presumably dif-
ferent) requirement of context-independence. The strong scores
on the within-context version of the new benchmark, which uses
one-phone sequences, demonstrate that the first problem is not
an issue with current systems.

We base our novel evaluation on the ABX-LS benchmark
[6] for English, one of the ZeroSpeech ABX benchmark evalua-
toins, itself based on LibriSpeech (LS) [9]. This benchmark ad-
ditionally measures speaker invariance by comparing a within-
speaker score, in which all of the phone triplets belong to the
same speaker (e.g., \( a = \text{flu}_T, b = \text{flu}_T, x = \text{flu}_T \)) to an
across-speaker score, in which \( a \) and \( b \) are the same speaker
and \( x \) a different one (e.g., \( a = \text{flu}_T, b = \text{flu}_T, x = \text{flu}_T \)). It also
compares the clean speech part of the LibriSpeech corpus to the
other part. The ABX phone discriminability measure has been
demonstrated to correlate with the performance of speech rep-
resentations on a number of other quality measures and down-
stream tasks [6].

Another intrinsic quality measure sometimes applied to
speech representations is the mean average precision, or MAP
on spoken word embeddings [10, 11]. MAP assesses how well
spoken word embeddings separate different word types using
a method similar (but not equivalent) to the ABX score. All
pairs of word tokens in the test corpus are compared using a
dissimilarity; then, a precision–recall curve for same/different
word type classification is calculated across different dissimilar-
ity thresholds. The MAP score is the area under the precision–
recall curve. A simple (albeit lossy) way of constructing spo-
ken word embeddings from a self-supervised representation is
to simply do mean pooling (time average), which allows for
an approximate assessment of the quality of the representation for
encoding a spoken lexicon.

We propose an ABX-discriminability measure of how
well learned speech representations track context-independent
phonemes. The most closely related work is [12], which applies
a closely-related ABX-discriminability score to MFCC repre-
sentations to measure how much different classes of phonemes
are acoustically influenced by context, rather than to evaluate
learned representations. Additionally, a number of papers have
used phoneme classification as an evaluation for the quality of
learned speech representations [2, 13]. Correct phoneme classi-
fication for segments extracted from a representation requires
text-context independent representations. However, these classifi-
cation analyses do not measure context-independence: here,
we introduce the critical comparison between a single mea-
sure (ABX) in a context-dependent versus -independent mode
to measure context-independence.

3. Methods
Starting from ABX-LS, we develop a novel variant to evalu-
ate context-invariance.\(^2\) Rather than extracting triphone tokens,
we extract phonemes in isolation. In the within-context con-
dition, the immediately preceding and following phoneme are
held constant across \( a, b, \) and \( x \). In the without-context con-
dition, there are no constraints on the surrounding context; in
general, it varies. Comparing the two measures invariance to
changes in context, i.e., to coarticulation and allophony.

We evaluate all of the self-supervised speech representa-
tions submitted to the 2021 ZeroSpeech benchmark [6], which
appear on the current leaderboard. HuBERT is an implementa-
tion of [4] described in [14]; CPC is an implementation of [2]
as described in [15]. This model was the baseline model for
the benchmark, and many submitted systems re-used these baseline
features unchanged (taking different approaches to the language
modelling component of the benchmark, which does not inter-
est us here). We exclude these submissions. S-CPC [16] at-
tempts to push CPC to learn representations of phone segments
that are stable across time; P&H VG is a “visually grounded” model [17]
that trains end-to-end on a masked language mod-
elling objective using spoken picture captions; ZR VG and ZR
VG-CPC [18] are also trained with picture captions, based on
[19], with the first using MFCCs as input and the second using
CPC representations; ResDaveNet, also visually-grounded, is
an implementation of [20]; finally CPC+Seg discovers bound-
aries on CPC units and applies pooling.

4. Experiment 1
Results of applying the novel evaluation to these models are
shown in Table 1 and Figure 1. The separation between the
\( \Delta \) and the \( \bullet \)-pointed lines in Figure 1 shows the gap between
within-speaker and across-speaker to vary between a modest
penalty and a doubling of the error rate, depending on the
model and the evaluation condition. The gap between solid
and dashed lines compares the clean speech and other
subsets, and is of a similar magnitude. However, the gap between the
darker purple within-context and the lighter orange without-context
scores is much greater, tripling or even quadrupling the error
rate. This lack of context-invariance disproportionately affects
high-performing models such as HuBERT and CPC, suggest-
ing these models’ loss encourages them to perform particularly
well within context.

5. Experiment 2
While Experiment 1 suggests the representations evaluated
are not context-invariant, an alternate explanation is that their
time-scales are incompatible with that of the ABX evaluation.
The ABX evaluation extracts representations from time-stamps
based on gold-standard alignments. Not all model loss func-
tions penalize time warping, and so some representations may
preserve the order of phonemes without preserving their exact
position in the sequence. If the evaluation is using the “wrong”
time-alignment for the representation, this would have a dispro-
portional impact on the without-context condition. Because
the context around the phoneme is different, small disagree-
ments in alignment between the representation and the gold
transcription will be more important.

To assess this, we change the way we deal with time in the

\(^2\)In what follows, we report performance only on the dev subset; results on the test subset are qualitatively the same.
Table 1: ABX context-independence evaluation. ABX error scores (%). Lower scores are better. Best scores in each column are bolded.

|          | Clean Within-speaker | Across-speaker | Other Within-speaker | Across-speaker |
|----------|----------------------|----------------|----------------------|----------------|
|          | W/in-ctx | W/out-ctx | W/in-ctx | W/out-ctx | W/in-ctx | W/out-ctx | W/in-ctx | W/out-ctx |
| HuBERT   | 1.56     | 7.26     | 2.13     | 8.04     | 3.08     | 8.64     | 4.78     | 10.09     |
| CPC      | 1.99     | 7.15     | 2.72     | 12.73    | 4.21     | 10.21    | 6.80     | 11.08     |
| S-CPC    | 1.99     | 7.17     | 2.69     | 12.96    | 4.08     | 11.07    | 6.65     | 11.05     |
| P&H VG   | 2.32     | 8.97     | 2.80     | 10.25    | 5.75     | 15.23    | 9.23     | 16.53     |
| ZR VG-CPC| 3.43     | 12.26    | 4.74     | 12.73    | 7.86     | 14.29    | 11.99    | 16.89     |
| ResDaveNet | 5.31   | 11.58    | 6.80     | 12.96    | 7.12     | 17.78    | 11.41    | 19.42     |
| ZR VG    | 5.31     | 15.45    | 6.89     | 16.35    | 10.36    | 18.21    | 14.77    | 19.73     |
| CPC+Seg  | 5.64     | 13.65    | 7.24     | 13.97    | 11.99    | 18.21    | 14.77    | 19.73     |
| Spectrogram | 12.41 | 21.13    | 19.09    | 24.85    | 14.95    | 23.32    | 23.80    | 28.48     |

Figure 1: ABX context-independence evaluation scores. Scale is logarithmic. Purple (darker) lines are within-context and orange (lighter) lines are without-context. The ▲ points are within-speaker while the ● points are across-speaker. Finally, the solid lines are clean and the dashed lines are other.

Table 2: ABX scores (%) comparing DTW (less sensitive to alignment) and Hamming (more sensitive) methods.

|          | Within-context | Without-context |
|----------|----------------|-----------------|
|          | DTW | Hamming | DTW | Hamming |
| Gold +2  | 1.84 | 3.70    | 2.95 | 4.46    |
| Gold +4  | 7.91 | 10.05   | 9.73 | 10.23   |
| Gold +6  | 14.67| 16.17   | 16.26| 16.67   |
| Gold +8  | 21.09| 21.46   | 21.62| 21.96   |
| Spectrogram | 12.41 | 14.05 | 21.13 | 20.34 |
| HuBERT   | 1.56 | 1.91    | 7.26 | 7.86    |
| CPC      | 1.99 | 2.18    | 7.15 | 6.91    |
| S-CPC    | 1.99 | 2.18    | 7.17 | 6.97    |
| P&H VG   | 2.32 | 2.51    | 8.97 | 9.52    |
| ZR VG-CPC| 3.43 | 3.42    | 12.26| 12.59   |
| ResDaveNet | 5.31 | 6.05   | 11.58| 12.44   |
| ZR VG    | 5.31 | 5.48    | 15.45| 15.71   |
| CPC+Seg  | 5.64 | 6.18    | 13.65| 13.20   |

ABX calculation. Instead of aligning a and x (b and x) using DTW and averaging dissimilarities for the aligned frames, we perform pooling to obtain a single representation for a, for b, and for x, weighted using a Hamming window, which peaks at the centre of the sequence; d is angular dissimilarity or KL divergence. This is less forgiving than DTW, which can compensate for minor errors in the alignment. If the alignment is problematic, ABX scores should worsen when the Hamming window is used. In particular, some pairs of representations A–X may be similar to each other but temporally misaligned with each other. Pooling with a filter that peaks in the middle therefore tends to make the inconsistency between A and X worse, particularly if the surrounding context material is not necessarily constant. DTW, on the other hand, can compensate to some degree.

We perform the experiment on the within-speaker, clean condition. To validate the experiment, we evaluate manipulated gold phoneme transcriptions (one-hot, by frame) wherein phoneme boundaries were sometimes ($p = 0.5$) shifted right by 4, 6, or 8 frames. We expect scores to be non-zero for the manipulated representations.

Results are shown in Table 2. Frame-shifting the gold transcription indeed makes scores worse and without-context evaluation is more impacted; pooling amplifies the effect. The spectrogram does not appear to be perfectly “aligned,” nor do the self-supervised representations (Hamming pooling makes the scores worse): information about phone identity is not all concentrated at the centre of the gold segment. However, unlike for the manipulated alignments, the difference (<14% of the original error rate) falls far short of explaining the penalty incurred in the without-context condition. We conclude that alignment is at most a small part of the reason for systems’ issues with context-invariance.

6. Experiment 3

For tasks such as unsupervised discovery of a lexicon of word types [21], it is critical to accurately represent the phonemic content of word tokens without colouring by adjacent context. Following [10], we use the mean average precision (MAP) as defined above to assess the discriminability of word embeddings based on the self-supervised speech representations above. We construct embeddings for each word token in the

[^3]: Using pooling on the +8 frame-shifting has no clear effect, perhaps because the representation is already so degraded that it only makes coarse-grained contrasts that are somewhat resistant to further perturbation.
gold annotation using mean pooling. While this method is necessarily lossy, it is sufficient to demonstrate the importance of context-invariance for speech word embeddings.

For the self-supervised systems, within- and without-context ABX scores are strongly correlated. Thus, to evaluate the role of context-independence by itself, we artificially dissociate the two. We do so by applying a square filter to each of the representations: we replace each frame by the surrounding time average in a window of 3, 5, or 7 frames, pulling parts of the surrounding context into the representation of a given frame. The filter is a blurring operation to make the representations more dependent on context—in other words, less context-robust. This will have an outsized impact on the without-context condition: for within-context, the surrounding phones are identical for A, B, and X, while, for without-context, filtering will introduce substantial noise. We assess whether these degraded representations have lower MAP scores. We perform the experiment on the clean subset again, this time examining the across-speaker ABX as it is more relevant to the representationally coherence of words (within-speaker results are qualitatively similar).

Results are shown in Table 3. Across systems, the within-context ABX scores worsen slightly as the filter width increases (by around 0.1% in general), while the without-context ABX scores show much greater degradation (e.g., HuBERT gets about 50% worse in the within-context condition at width 7. CPC around 12% worse, while in the without-context condition they degrade by 90% and 28% respectively). The exception is width 3, which has little impact.

In general, we see a relation between the without-context ABX and MAP scores. Across systems, we note much unexplained variance—the MAP scores are not entirely predictable from the ABX scores—consistent with the result of [11]. Within systems, however, we see a notable degradation in the MAP scores as the (without-context) ABX scores decline. This suggests that further improving the context-independence of these units would lead to more consistent representations of word types, thus more useful for tasks that require accurate representation of individual word tokens.

### Table 3: ABX and MAP scores (%) for filtered representations.

| Filter | ↓ABX w/in | ↓ABX w/out | ↑MAP |
|--------|-----------|------------|------|
| HuBERT None | 5.13 | 8.04 | 48.25 |
| HuBERT 3 | 2.06 | 9.24 | 47.63 |
| HuBERT 5 | 2.50 | 11.92 | 45.86 |
| HuBERT 7 | 3.20 | 15.34 | 42.90 |
| CPC None | 2.72 | 2.79 | 32.93 |
| CPC 2 | 3.75 | 7.55 | 32.73 |
| CPC 5 | 2.86 | 8.25 | 32.31 |
| CPC 7 | 3.05 | 9.34 | 31.66 |
| S-CPC None | 2.69 | 2.79 | 32.93 |
| S-CPC 2 | 2.69 | 7.56 | 32.73 |
| S-CPC 5 | 2.80 | 8.25 | 32.31 |
| S-CPC 7 | 2.98 | 9.34 | 31.66 |
| P&H VG None | 2.80 | 10.25 | 42.77 |
| P&H VG 2 | 2.95 | 11.63 | 41.90 |
| P&H VG 5 | 3.54 | 14.60 | 39.91 |
| P&H VG 7 | 4.61 | 18.18 | 36.88 |
| ZR VG-CPC None | 4.74 | 12.73 | 33.55 |
| ZR VG-CPC 2 | 4.85 | 12.23 | 33.19 |
| ZR VG-CPC 5 | 5.43 | 16.63 | 32.47 |
| ZR VG-CPC 7 | 6.16 | 19.31 | 31.41 |

7. **Summary of contributions**

This paper introduces a new version of the ZeroSpeech ABX-LS evaluation measure for self-supervised representations in English. The evaluation is freely available at [https://zerospeech.com/](https://zerospeech.com/) and code for the additional analyses in the paper is available at [https://github.com/perceptimatic/context-invariance-paper](https://github.com/perceptimatic/context-invariance-paper). While the previous benchmark measures whether representations are consistent with the phonemic contrasts of the language only within specific phonetic contexts, the new benchmark measures the context-independence of representations. We demonstrate that current systems show poor context independence: the typical case is a 300-400% drop in performance on a context-independent task, far larger than the gap seen for speaker-independence or for clean versus less-clean speech. We propose that future research address this gap.

8. **Acknowledgements**

Supported by the Connaught Fund and the Arts and Science Bridging Fund, U. of Toronto, Natural Sciences and Engineering Research Council of Canada (NSERC) RGPIN-2022-04431, ANR grants ANR-17-EURE-0017 Frontcong. ANR-10-IDEX-0001-02 PSL* and ANR-19-P3IA-0001 PRAIRIE 3IA Institute, by a Meta AI Research Gift, and by the Pronovost Morgan Family Foundation Fellowship in Ethical AI.

9. **References**

[1] E. Dupoux, “Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner,” Cognition, vol. 173, pp. 43–59, 2018.

[2] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” CoRR, vol. abs/1807.03748, 2018.

[3] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” arXiv preprint arXiv:2006.11477, 2020.

[4] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakttia, R. Salakhutdinov, and A. Mohamed, “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” IEEE Trans. on Audio, Speech, and Lang. Proc., vol. 29, pp. 3451–3460, 2021.

[5] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshoka, X. Xiao et al., “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” IEEE Journal of Selected Topics in Signal Processing, 2022.

[6] E. Dunbar, N. Hamilakis, and E. Dupoux, “Self-supervised language learning from raw audio: Lessons from the zero resource speech challenge,” IEEE Journal of Selected Topics in Signal Processing, 2022.

[7] S.-W. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakttia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin et al., “SUPERR: Speech processing universal performance benchmark,” in INTERSPEECH, 2021. [Online]. Available: https://arxiv.org/abs/2105.01051

[8] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux, “Evaluating speech features with the Minimal-Pair ABX task (I): Analysis of the classical MFC/PLP pipeline,” in INTERSPEECH, 2015.

[9] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “LibriSpeech: an ASR corpus based on public domain audio books,” in ICASSP, 2015, pp. 5206–5210.
[10] M. A. Carlin, S. Thomas, A. Jansen, and H. Hermansky, “Rapid evaluation of speech representations for spoken term discovery,” in Twelfth Annual Conference of the International Speech Communication Association, 2011.

[11] R. Algayres, M. S. Zaiem, B. Sagot, and E. Dupoux, “Evaluating the reliability of acoustic speech embeddings,” in INTERSPEECH, 2020.

[12] T. Schatz, R. Turnbull, F. Bach, and E. Dupoux, “A quantitative measure of the impact of coarticulation on phone discriminability,” Interspeech 2017, 2017.

[13] B. Higy, L. Gelderloos, A. Alishahi, and G. Chrupała, “Discrete representations in neural models of spoken language,” arXiv preprint arXiv:2105.05582, 2021.

[14] T. A. Nguyen, B. Sagot, and E. Dupoux, “Are discrete units necessary for spoken language modeling?” IEEE J. Sel. Top. in Sig. Proc., 2022.

[15] T. A. Nguyen, M. de Seyssel, P. Rozé, M. Rivière, E. Khartonov, A. Baevski, E. Dunbar, and E. Dupoux, “The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling,” arXiv preprint arXiv:2011.11588, 2020.

[16] S. Bhati, J. Villalba, P. Żelasko, L. Moro-Velazquez, and N. Dehak, “Segmental contrastive predictive coding for unsupervised word segmentation,” arXiv preprint arXiv:2106.02170, 2021.

[17] P. Peng and D. Harwath, “Self-supervised representation learning for speech using visual grounding and masked language modeling,” arXiv preprint arXiv:2202.03543, 2022.

[18] A. Alishahi, G. Chrupała, A. Cristia, E. Dupoux, B. Higy, M. Lavechin, O. Räsänen, and C. Yu, “Zr-2021vg: Zero-resource speech challenge, visually-grounded language modelling track,” arXiv preprint arXiv:2107.06546, 2021.

[19] B. Higy, D. Elliott, and G. Chrupała, “Textual Supervision for Visually Grounded Spoken Language Understanding,” in Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, Nov. 2020, pp. 2698–2709.

[20] D. Harwath, A. Recasens, D. Surtis, G. Chuang, A. Torralba, and J. Glass, “Jointly discovering visual objects and spoken words from raw sensory input,” in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 649–665.

[21] O. Räsänen and M. A. C. Blandón, “Unsupervised discovery of recurring speech patterns using probabilistic adaptive metrics,” arXiv preprint arXiv:2008.00731, 2020.