Establishing degrees of closeness between audio recordings along different dimensions using large-scale cross-lingual models

Anonymous EACL submission

Abstract

In the highly constrained context of low-resource language studies, we propose a new unsupervised method using ABX tests on audio recordings with carefully curated metadata to shed light on the type of information present in the representations. ABX tests determine if the representations computed by a multilingual speech model encode a given characteristic. Two experiments are devised: one on acoustic aspects, specifically room acoustic characteristics, and one on phonetic aspects. The results confirm that the representations extracted from recordings with different linguistic/extra-linguistic characteristics differ along the same lines. Embedding more audio signal in one vector better discriminates extra-linguistic characteristics, whereas shorter snippets are better to distinguish segmental information. The method is fully unsupervised, potentially opening new research avenues for comparative work on under-documented languages.

1 Introduction

In recent improvements in speech processing,[1] the amount of data used at pre-training has been instrumental (Wei et al., 2022), which makes it more challenging – if not impossible – to reach similar levels of performance for endangered languages. Developing new unsupervised approaches, in addition to being cost-effective (Bender et al., 2021), helps us better understand the models.

Training a speech model often results in changing the weights of the parameter matrices to specialize it for a task. But speech, when accessed via audio recordings, is highly multifactorial: a recorded voice tells a message and an intention; the audio contains information about the surroundings.

Our experimental setup relies on tailored datasets to see how specific differences in the input signal are reflected in the output vectors. ABX tests are used on carefully selected data in Na language (ISO-639-3: nru). Two series of experiments explore different dimensions to assess differences between recordings. The *folk-tale series* aims to explore an extra-linguistic dimension by comparing seven versions of the same tale by the same speaker, and the *phonetics series* explores segmental dimensions by comparing sentences (some identical, some different) from different speakers.

The results provide an insight into the nature of the information encoded in the representations of a model such as XLSR-53 (Baevski et al., 2020; Babu et al., 2021). Our findings suggest that ABX tests can be leveraged to bring out differences in the acoustic setup (room, microphone) or in the (segmental) linguistic content. A parametric study shows that processing audio by snippets[2] of 10 s is sufficient to bring out differences on the acoustic setup, while 1 s snippets are better for segmental characteristics.

This study offers an innovative method to detect confounding factors in corpora intended for unsupervised machine learning, and provides a means to accelerate the classification of recordings (e.g. by noise level or genre) where such metadata are unavailable.

2 Method

We propose a method based on two components: (i) ABX tests to determine the presence/absence of a given characteristic in a representation and (ii) audio corpora with precise metadata. These metadata allow us to build datasets based on one characteristic: language name, speaker ID, room acoustics, microphone type or segmental content.

ABX tests To find out, in an unsupervised manner, if a multilingual speech model encodes a characteristic $\mathcal{C}$ of the speech signal, we use the ABX

[1] in ASR, TTS, and even on corpora/languages/tasks not seen at pre-training (Guillaume et al., 2022).

[2] The term ‘snippet’ is preferred over ‘segment’, reserving the latter to refer to phonetic segments.

tests introduced by Carlin et al. (2011) and Schatz et al. (2013). The test relies on vector representations built by a pre-trained model for three audio snippets. Let A and X denote the snippets that share the characteristic $\mathcal{C}$, while B is the one that does not. The test checks whether the distance $d(A, X)$ is smaller than $d(A, B)$.

The ABX score corresponds to the proportion of triplets for which the condition $d(A, X) < d(A, B)$ holds true. An ABX score close to 50 % indicates that, on average, the distance between A and X is the same as the distance between A and B, suggesting that $\mathcal{C}$ is not encoded in the audio representation. Conversely, the closer the score is to 100 %, the more the representation captures the characteristic $\mathcal{C}$.

ABX tests are interesting for low-resource scenarios because they require no additional training, so they can be directly applied to the representations (unlike linguistic probes: Belinkov and Glass 2019; Yin and Neubig 2022).

Corpora All recordings come from the Pangloss Collection, an open-access archive of 'little-documented languages'.[3] Two series of recordings selected for their characteristics are considered:

(i) The ***folk tale series*** consists of seven recording sessions of the same folk tale in Na, told by the same speaker. These experiments focus on the effect of the recording conditions, which are slightly different from one version to another.

The first batch studied comprises three versions: V1, V2 and V3. V1 was recorded in a room with perceptible reverberation, while V2 and V3 were recorded in a damped room.

The second batch is made up of V6 and V7. These two versions were recorded in the same acoustic conditions. The audio was captured simultaneously by two microphones: a headset microphone and a handheld microphone placed on a small stand.

The third batch compares V4 and V5, which have a native listener acting as respondent, to all the other recordings of the *folk tale series*.

These recordings are particularly interesting because some potential confounding factors (typically the topic and the speaker) are controlled, which makes it possible to focus on the influence of certain specific factors (e.g. room acoustics).

(ii) The ***phonetics series*** is made up of five recordings of phonetic elicitations and one recording of words in a carrier sentence (lexical elicitations). The language is Na. Three speakers identified as AS, RS and TLT are considered. We included two recording sessions, which allows for intra-speaker comparison. We thus arrive at a fine-grained heatmap of ABX scores.

The five recordings of phonetic elicitations have the same content (apart from the variation inherent to the experimental process in fieldwork conditions: Niebuhr and Michaud 2015) whereas lexical elicitations are a completely different content. Only AS participated in both the phonetic and lexical elicitation sessions.

Table 1 and 2 in App. A provide a more exhaustive outline of the above mentioned metadata.

Experimental Setting In all our experiments, we use the `XLSR-53`[4] model, a `wav2vec2` architecture trained on 56 kh of (raw) audio data in 53 languages (Conneau et al., 2020). For the comparisons, we consider audio snippets of length 1 s, 5 s, 10 s and 20 s in order to study the effect of snippet length on our ABX test. We use max-pooling to build a single vector representing the snippet because we are interested in assessing differences between vectors. As advocated by Schatz et al. (2013), we use the cosine distance in all our experiments.

We used the representations from the 21[st] layer, following several recent results (Pasad et al., 2021; Li et al., 2022, 2023) which show that the ability of `wav2vec2` representations to capture linguistic information declines in the last two layers.

3 Results

Using ABX tests with carefully selected audio recordings, we investigate whether or not the audio representations computed by `wav2vec2` capture specific information from the audio signal.

3.1 Study of various versions of the same tale

The aim of our first experiment is to determine whether certain extra-linguistic variables (e.g. room acoustics, type of microphone, ...) are captured in the neural representations. For that, we consider recordings from the *folk tale series* and use ABX tests to distinguish between different versions of the tale: these scores are calculated from triplets

[3] For reproducibility reasons, an exhaustive list of the resources' DOI is provided in App. E.

[4] The HuggingFace API was used (signature `facebook/wav2vec2-large-xlsr-53`).

consisting of two snippets of 10 s from the same version and one snippet from a different version.[5]

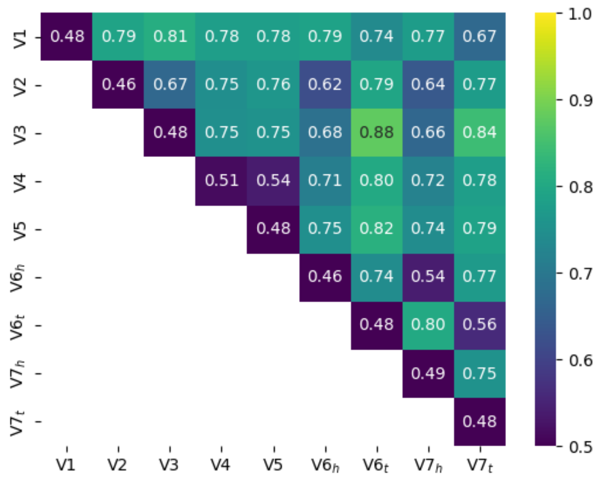

Figure 1: ABX scores when distinguishing different versions of the *folk tale series*.

Figure 1 shows that, in most cases, with a 10 s snippet-length it is possible to distinguish between the different recordings, although it is always the same speaker telling the same story. It suggests that neural representations capture much more than the linguistic information needed to understand speech, and it seems possible to use them to retrieve information related to the recording conditions. This observation is surprising: the ABX tests only use the raw representation constructed by a pre-trained model on a very large quantity of recordings covering a wide array of speakers, languages and recording conditions, and we would have expected that the speech representations be cut off from an information deemed irrelevant.

A more precise analysis of the scores between two recording conditions provides a better understanding of the information that is or is not captured by the representations. Note that all our observations are the most visible with 10 s snippets, which suggest that this is the proper setting to reveal differences at a broad acoustic level.

The first batch, a comparison between V1, V2 and V3 (NW corner of Figure 1) is very interesting: the ABX scores show that the representation of V2 and V3 are indistinguishable when compared to the representations of V1. We know from Section 2 that the main difference between these three recordings is related to the recording venue: V2 and V3 were recorded in the same place, less reverberating than the place where V1 was recorded. To confirm the influence of this parameter, we carried out a complementary experiment by artificially adding *reverb*[6] to the V2 recordings and measuring the ABX score between the V1 and modified V2 recordings. Figure 2 shows the evolution of the ABX score as a function of the amount of reverb added. One interesting observation is that when gradually increasing the amount of reverb in V2, the ABX score decreases first before increasing again. It means that V1 is closer to V2 with 5 % reverb, which suggests a relation of causality between the amount of reverberation and the degree of closeness between the recordings of this batch.

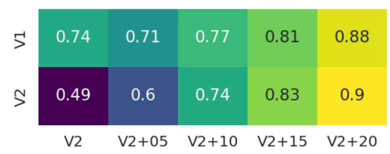

Figure 2: Reproducing V1 room tone with artificial room tone applied on V2 (snippet length = 5 s).

In the second batch, the sub-versions of V6 and V7 are labeled as h for *headset* and t for *table*. Figure 1 shows that he XLSR-53 representations can effectively distinguish between these two microphone types with high precision. For instance, the ABX scores between V6$_h$ and V6$_t$ are some of the highest in our experiment. However, when it comes to distinguishing between two different recordings made with the same microphone (i.e. V6$_h$-V7$_h$ and V6$_t$-V7$_t$), the ABX scores are only slightly better than scores for the same recording. This suggests that these representations strongly depend on the microphone used: two vectors representing the same audio signal but recorded by different microphones will be more dissimilar than those representing two different audio signals recorded by the same microphone.

Finally, the results in Figure 1 also show that the representations of recordings V4 and V5 are very similar: the ABX score between these two versions is only 54%, whereas it is at least 71% with all the other versions. One possible explanation for this observation is that these two sessions were conducted by with a local listener. This observation suggests that the neural representations capture information about the context in which the recording took place that is potentially very distant from the audio produced by the speaker. Further experiments are necessary to confirm this conclusion.

[5]Results for other snippet lengths are reported in App. C.

[6]We use Audacity to add 5, 10, 15 or 20 % reverb.

3.2 Study of a phonetics corpus

While it is quite obvious that two sentences with a different linguistic content in perfectly controlled conditions will come out as different when submitted to an ABX test, the answer is not immediate when it comes to a whole recording. It is also not obvious that two different sentences uttered by two different speakers are distinguished solely due to a difference in the linguistic content: speaker ID acts as a confounding factor.

The aim of this second experiment is to perform ABX tests on data with differences on the phonetic segments. To do this, we rely on a phonetics corpus recorded in a controlled manner, where each speaker received similar instructions. The scores are calculated from triplets consisting of two snippets of 1 s from the same recording and one snippet of 1 s from a different recording.[7]

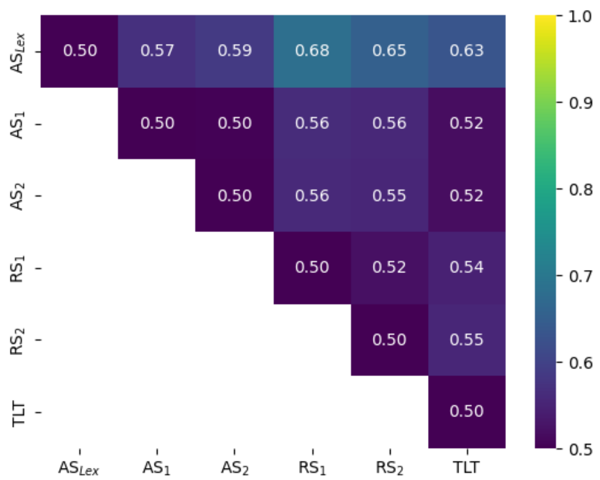

Figure 3: ABX scores for the comparisons between elements of the *phonetics series*.Speaker AS has three recordings, and has three recordings (AS_1, AS_2, AS_{Lex}), RS has two (RS_1, RS_2) and TLT has one.

First, Figure 3 shows that with a 1 s snippet-length it is nearly not possible to distinguish between the different recordings of the same sentences, even when the speakers differ. It suggests that neural representations, in this configuration, effectively 'centrifugate' the extra-linguistic information. This observation is not surprising given how the models are pre-trained, and it is a convenient springboard for the second part of the analysis, which consists in comparing these recordings of identical sentences to another one with different sentences.

[7]Results for other snippet lengths are reported in App. D.

The results in the first row of Figure 3 suggest that the ABX tests reveal differences in linguistic content. The magnitude of the discrepancy (between row 1 and the others) depends on whether or not the speaker is different. The fixed-speaker discrepancy is around 0.07, while the cross-speaker discrepancy is around 0.11, which means that even with 1 s snippets the effect of the speaker is not much less than the effect of the different content.

In this study, ABX scores are averaged over an entire recording. For phonetic differences, it would be interesting to be able to perform comparisons on a per-sentence basis, but that would constitute a departure from a fully unsupervised approach.

4 Discussion and conclusion

When one undertakes the task of comparing vector representations of audio, differences are expected, too many of them rather than too few. We adopted an experimental method to submit a given model to different experiments with test variables.

In the first experiment, the recordings are distinguished according to their technical acoustic properties (room acoustics, microphone) or interview method. A 10 s snippet length seems to reveal differences in these characteristics.

In the phonetics experiment, we focused on 1 s snippet lengths. The recordings of three speakers who participated in a phonetics experiment, quasi-identical to one another, are distinguished from a recording with a different content, but the distinction is not very strong.

The study of the *folk tale series* suggests that recordings can be distinguished based on extra-linguistic variables, and this is achieved using long snippet lengths. We think that with appropriate data, long-range variables such as genre or typological properties of the language could also be detected in the representations. These results provide a means for automatically classifying recordings e.g. by noise level or genre.

The results from the *phonetics series* suggest that smaller snippets encompass less information, which results is smaller differences on the ABX score. This observation presents an interest for cross-linguistic comparison, but it would require additional investigations to devise a method more suited to phonetic segments. Among the possible improvements, using segmented corpora would be an interesting way to pursue.

Limitations

As is often the case for endangered languages (Liu et al., 2022), our corpora rely on a few speakers of the same gender. In our case, we exploit a resource with rich metadata to build experiments with minimal differences and observe sets that differ by one characteristic only. The conclusions drawn on the speaker-independent setting in Section 3 may need to be reanalyzed when we run the experiment on cross-gender data.

Our study does not perform comparisons with other methods for identifying characteristics, because other methods require more data than the amount treated here (typically linguistic probes using classifiers).

We have not investigated how the model reacts to a superposition of variables sensitive to a given snippet length. Therefore, we would need to extend our experiments further, e.g., to check how a 10 s snippet length is handled when assessing a discrepancy in speaker and room acoustics.

We plan to extend this study by adding data from experimental phonetics experiments related to second language acquisition, as they often include productions from the same speaker in multiple languages. Experimental phonetics corpora are devised under highly controlled conditions, which is beneficial for our study as it removes potential confounding factors.

Ethics Statement

The study presented here relies on small-sized corpora because the methods are meant for low-resource languages, i.e., without a significant amount of data available. This limitation is offset by the wealth of metadata available for each recording in the Pangloss Collection. Pangloss is a world language open-access archive developed in a Dublin-core compliant framework (Weibel et al., 1998).

The data used in this study are first-hand, collected by researchers working with the communities to document and describe their language. They are the result of months of collaborative work in the field to transcribe and translate the data with native speakers (typically the speaker himself/herself). The speakers all consented to the use of these data for scientific purposes and were compensated for their work as linguistic consultants.

All data and models in this study are open-access under a Creative Commons license stated on the consultation page for each resource (which is also the landing page of its DOI listed in Table 3). The information needed for reproducibility is present in the text (model information) or the appendices (data). The metadata collected were directly collected via questionnaires during the fieldwork. Gender, for example, corresponds to the gender the speaker provided in the questionnaire.

Acknowledgements

Section removed for anonymization purposes.

References

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Michael A Carlin, Samuel Thomas, Aren Jansen, and Hynek Hermansky. 2011. Rapid evaluation of speech representations for spoken term discovery. In *Twelfth Annual Conference of the International Speech Communication Association*.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *CoRR*, abs/2006.13979.

Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux, Solange Rossato, Minh-Châu Nguyên, and Maxime Fily. 2022. Fine-tuning pre-trained models for Automatic Speech Recognition: experiments on a fieldwork corpus of Japhug (Trans-Himalayan family). In *ComputEL-5 5th Workshop on Computational Methods for Endangered Languages (ComputEL-5)*, Proceedings of ComputEL-5: Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages, Dublin, Ireland.

Yuanchao Li, Peter Bell, and Catherine Lai. 2022. Fusing ASR outputs in joint training for speech emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7362–7366. IEEE.

Yuanchao Li, Yumnah Mohamied, Peter Bell, and Catherine Lai. 2023. Exploration of a self-supervised speech model: A study on emotional corpora. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 868–875. IEEE.

Zoey Liu, Justin Spence, and Emily Prud'hommeaux. 2022. Investigating data partitioning strategies for crosslinguistic low-resource ASR evaluation. *arXiv preprint arXiv:2208.12888*.

Oliver Niebuhr and Alexis Michaud. 2015. Speech data acquisition: the underestimated challenge. *KALIPHO-Kieler Arbeiten zur Linguistik und Phonetik*, 3:1–42.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921. IEEE.

Thomas Schatz, Vijayaditya Peddinti, Francis Bach, Aren Jansen, Hynek Hermansky, and Emmanuel Dupoux. 2013. Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. In *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association*, pages 1–5.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Stuart Weibel, John Kunze, Carl Lagoze, and Misha Wolf. 1998. Dublin core metadata for resource discovery. Technical report, IETF RFC.

Kayo Yin and Graham Neubig. 2022. Interpreting language models with contrastive explanations. *arXiv preprint arXiv:2202.10419*.

A Metadata for the experiments

The list of metadata for the experiments conducted is given in Table 1 for the *folk tale series* and in Table 2 for the *phonetics series*.

REC ID	Year	DUR (s)	MIC	ITV	Acoust.
V1	2006	518	Tab	out	ND
V2	2007	440	Tab	out	D
V3	2008	707	Tab	out	D
V4	2014	527	Hea	Na	D
V5	2014	423	Hea	Na	D
$V6_h$	2018	348	Hea	out	ND
$V6_t$	2018	348	Tab	out	ND
$V7_h$	2018	635	Hea	out	ND
$V7_t$	2018	635	Tab	out	ND

Table 1: Metadata for the *folk tale* series. MIC = microphone: Headset or Table; ITV = interviewer: outsider or Na (local). Acoustics: non-damped (ND), or damped (D).

REC ID	DUR (s)	SPK	SESSION TYPE
AS_1	1567	AS (F)	Phonetic elicit.
AS_2	952	AS (F)	Phonetic elicit.
RS_1	681	RS (F)	Phonetic elicit.
RS_2	786	RS (F)	Phonetic elicit.
TLT	897	TLT (F)	Phonetic elicit.
AS_{Lex}	1216	AS (F)	Lexical elicit.

Table 2: Metadata for the *phonetics series*. SPK = speaker; (F) = Female. Data collected in 2019

B M and SD values showing that ABX tests can be used to measure differences between our corpora

Figure 4 shows mean and standard deviation values for a comparison between inter-recordings scores (*phonetics series* and *folk-tale series* barplots) and intra-recording scores (*same-recording*), for different snippet lengths. For all snippet lengths, the average inter-recording ABX score is always significantly higher than the average intra-recording score, even for 1 s snippet-length. This shows that ABX tests can be used to measure differences in our experiments.

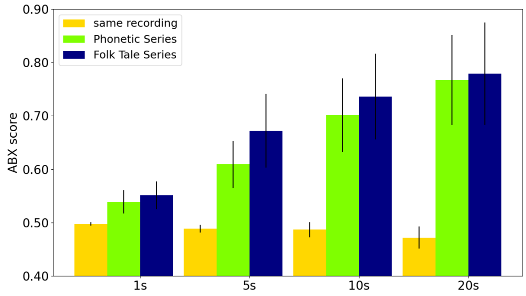

Figure 4: Average ABX scores for 1, 5, 10, 20 s snippets.

C ABX scores when distinguishing different versions of the *folk tale series* by the same speaker.

The 20 s value for snippet length has been investigated, and it does not bring out much more than the 10 s snippet length. In addition a 20 s snippet length with max-pooling tackles the limits of the max-pooling method. Indeed, we believe there is a limit to the amount of audio we can have in an embedding. Indeed, with the max pooling extraction method, each of the 980 vectors before pooling the 20 s of audio will only occupy, on average, 1.04 cells per final vector since it only has 1024 components. The results can be seen in Figure 5 for 20 s snippets, Figure 6 for 10 s snippets, Figure 7 for 5 s snippets, Figure 8 for 1 s snippets.

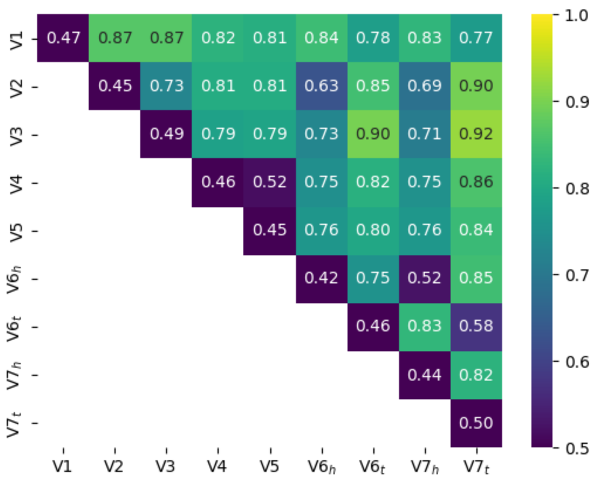

Figure 5: ABX scores for the *folk tale series*. (snippet length = 20 s).

Figure 6: ABX scores for the *folk tale series* (snippet length = 10 s).

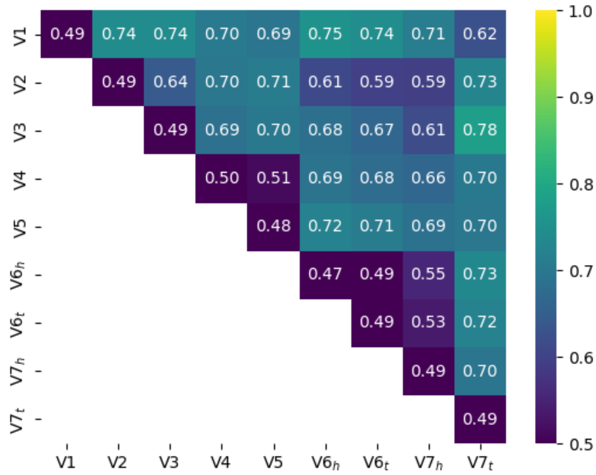

Figure 7: ABX scores for the *folk tale series* (snippet length = 5 s).

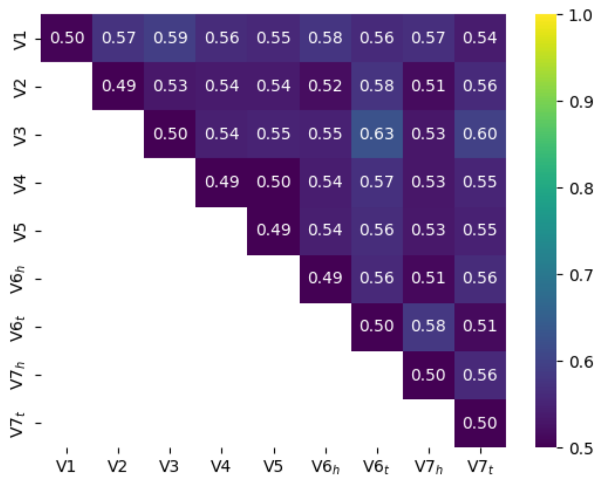

Figure 8: ABX scores for the *folk tale series* (snippet length = 1 s).

D ABX scores when distinguishing between elements of the *phonetics series*

The results can be seen in Figure 9 for 20 s snippets, Figure 10 for 10 s snippets, Figure 11 for 5 s snippets, Figure 12 for 1 s snippets.

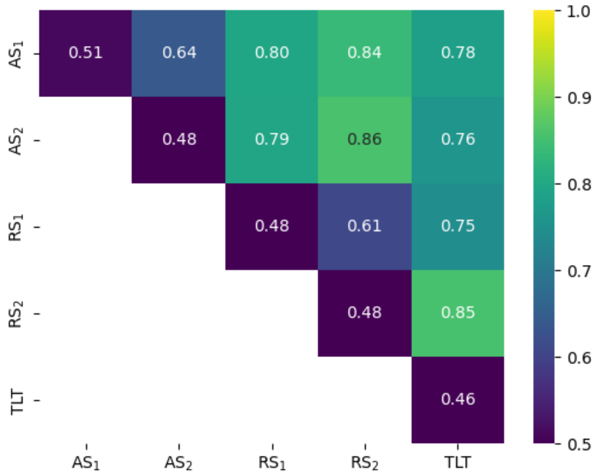

Figure 9: ABX scores for the comparisons between elements of the *phonetics series* (snippet length = 20 s).

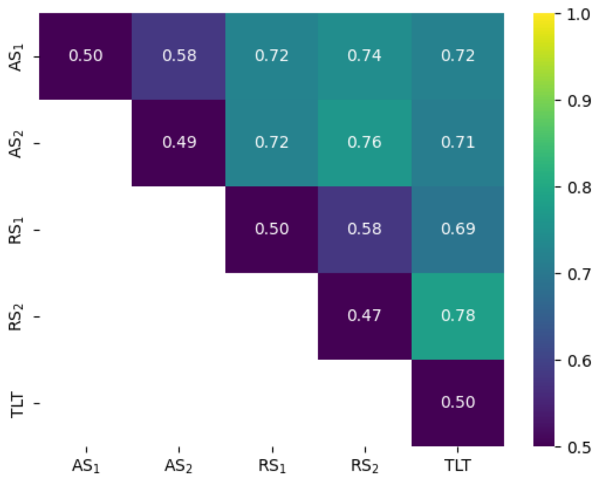

Figure 10: ABX scores for the comparisons between elements of the *phonetics series* (snippet length = 10 s).

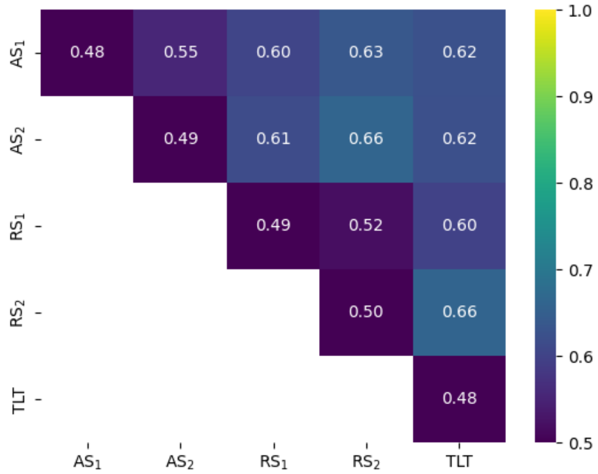

Figure 11: ABX scores for the comparisons between elements of the *phonetics series* (snippet length = 5 s).

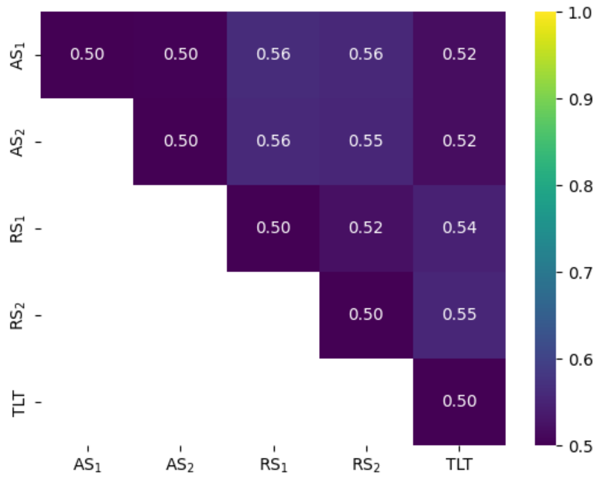

Figure 12: ABX scores for the comparisons between elements of the *phonetics series* (snippet length = 1 s).

E Audio resource: list of the recordings used for the study, with their DOI

Folk-tale series:

REC ID	DOI
V1	doi.org/10.24397/PANGLOSS-0004341
V2	doi.org/10.24397/PANGLOSS-0004343
V3	doi.org/10.24397/PANGLOSS-0004344
V4	doi.org/10.24397/pangloss-0004938
V5	doi.org/10.24397/pangloss-0004940
V6	doi.org/10.24397/pangloss-0007695
V7	doi.org/10.24397/pangloss-0007698

Phonetics series

REC ID	DOI
AS_2	doi.org/10.24397/pangloss-0008663
RS_2	doi.org/10.24397/pangloss-0008667
AS_1	doi.org/10.24397/pangloss-0008662
	doi.org/10.24397/pangloss-0008664
RS_1	doi.org/10.24397/pangloss-0008665
	doi.org/10.24397/pangloss-0008666
TLT	doi.org/10.24397/pangloss-0008668
	doi.org/10.24397/pangloss-0008669
AS_{Lex}	doi.org/10.24397/pangloss-0008670
	doi.org/10.24397/pangloss-0008671

Table 3: List of the DOIs for the recordings used in this study.