Speech Prosody 2022
23-26 May 2022, Lisbon, Portugal

Can Prosody Transfer Embeddings be Used for Prosody Assessment?

Mariana Julião[1,2], *Alberto Abad*[1,2], *Helena Moniz*[1,3]

[1]INESC-ID, Lisbon, Portugal
[2]Instituto Superior Técnico, University of Lisbon, Portugal
[3]FLUL - School of Arts and Humanities, University of Lisbon, Portugal

mariana.juliao@tecnico.ulisboa.pt, alberto.abad@inesc-id.pt, helena.moniz@inesc-id.pt

Abstract

In voice conversion, it is possible to transfer some characteristic components of a (target) speech utterance, such as the content, pitch, or speaker identity, from the corresponding component from another (source) utterance. This has recently been achieved by characterizing these components through neural-based vector embeddings which encode the specific information to be transferred. In the particular case of neural prosody embeddings, to the best of our knowledge, no work has explored the informativeness of these embeddings for other purposes, such as prosody assessment or comparison of prosodic patterns. In this work, we use an intonation data set and a voice conversion corpus to explore how these neural prosody embeddings group for utterances of different intonation, content, and speaker identity. We compare these neural prosody embeddings to hand-crafted acoustic-prosodic features and to content embeddings. We found that neural prosody embeddings can achieve a geometrical separability index as high as 0.956 for highly contrastive intonations, and 0.706 for different sentence types.

Index Terms: neural prosody embeddings, prosody transfer, prosody assessment

1. Introduction

The full expressive potential of prosody is by now widely acknowledged, even if not exhaustively understood. In a quest for naturalness in speech, Text-to-Speech (TTS) technology is now investing in prosody production [1, 2]. Meanwhile, the technology for Computer Assisted Language Learning (CAPT) is now strengthening its offer in terms of prosody teaching [3, 4].

Prosody encompasses much information that is crucial to spoken communication. Limitations in its production are considerably detrimental to speakers, due to illnesses, as Parkinson's [5, 6, 7], or to a limited command of the language, as in second language speakers (v.g., [8, 9, 10]). For this reason, it is very important to have good methods for the assessment of prosody. In general, it is also desirable that these methods do not exclusively rely on human annotation, as this may be expensive and cumbersome.

In this work, given the above mentioned limitations and challenges, we test the hypothesis that neural prosody embeddings can be useful for prosody assessment, in particular, prosody transfer embeddings. Prosody transfer tasks are currently possible due to the use of embeddings which encode prosodic content. Because of this, these embeddings might be directly and easily compared to each other. This could allow for an easy comparison of test and reference utterances, thus leading to an understanding of how the test utterance differs from equivalent utterances by reference (native, healthy) speakers.

To the best of our knowledge, the use of neural-based embeddings beyond its initial purpose remains vastly unexplored. Here, we extracted neural prosody embeddings using a voice conversion module [11], and considered an intonation data set and VCTK, a corpus for voice conversion [12]. We compared neural prosody embeddings, acoustic-prosodic features and content embeddings by using utterances with the same intonation (from the intonation data set) and utterances of contrasting intonation categories (from VCTK). We search for trends of separability, using clustering assessment metrics and visualizing the projections of the embeddings in a reduced dimensionality space.

2. Related Work

Much of the research on intonation for L2 is based on the direct comparison with a reference production of an intonation. This is the case of [13, 14, 15, 16, 17], where systems are trained to assess how good a sentence is when compared to a given utterance. In parallel to this, other approaches focus on the classification of the intonation of segments, in which a segment is classified according to a prosodic label, as in [18, 19]. Typically, the data used in the aforementioned tasks have been annotated by human experts.

Lately, the speech synthesis field has devoted much attention to prosody in a quest for naturalness in the synthesized speech. For example, in Tacotron [20], prosody evolved from being implicitly learnt to being represented in dedicated embeddings, by adding a prosody-specific encoder to the model [1, 2]. Other works, instead of relying on implicit learning of prosody, turn to explicit prosody annotation, while still using neural-based models, as [21, 22]. According to the authors of [21], most neural TTS methods fail to encode text-based prosodic-linguistic content, when compared to Statistical Parametric Speech Synthesis (SPSS), where linguistic and prosodic features are extracted. To bridge this gap, the authors encode information about stressed syllable and pitch accent into a phoneme embedding given at the input to the model. In [22], the model is fed with ToBI [23] labels as well.

Speech decomposition denotes the separation of speaker identity, spoken content, pitch and rhythm, for instance, or any desired combination of these. These decompositions normally rely on encoder-decoder architectures, which separately encode the parts to differentiate. The resulting embeddings can then be used as representations of the respective parts: for speaker identity [24, 25, 26]; for rhythm, content, pitch and speaker [11]; and to disentangle speaker from language [27].

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT), with reference UIDB/50021/2020 and PhD grant SFRH/BD/139473/2018.

10.21437/SpeechProsody.2022-60

3. Methodology

The methodology we follow consists of analysing data corresponding to different intonation types with different features: neural prosody embeddings, acoustic-prosody features and content embeddings. We take as premise that a set of features is as good to encode prosody as it is able to lead to well-separated clusters of contrastive utterances. We analyse this by using specific metrics to assess clustering, as well as by inspecting the utterances' t-SNE projections [28], a method that permits visualization of data in a two-dimensional space.

3.1. Features

3.1.1. Neural Prosody Embeddings

SpeechSplit [11] is a system for unsupervised decomposition of speech into pitch, rhythm, content and speaker, with applications to voice conversion. The framework consists on an autoencoder with three encoders (for pitch, rhythm, and content – speaker is only considered at the decoder) and one decoder. The information of the components is corrupted in such a way that the encoder of component X is the only encoder to have full access to the uncorrupted information of X. Furthermore, the model includes a mechanism of information bottleneck, which aims at forcing the output of each encoder to keep no more than the information of the corresponding component. We refer the reader to the original work [11] for implementation details.

In order to obtain fixed-length embeddings, the SpeechSplit model expects input sequences of 3 seconds, and shorter input sequences are padded with zeros. Since we are interested in extracting embeddings for sentences of arbitrary length, while keeping original model architecture, we trained the model regularly and afterwards altered the encoders at inference time. In the regular model, the last step of the encoders is the concatenation of the backward and forward outputs of an LSTM, after being sampled at regular spaced intervals. We, instead, changed this step so as to sample a fixed number of equally spaced points (the same as before), but with variable distance. This way we managed to keep a fixed length on the embeddings with a total size of 1536 components. In this work, we only used the embeddings of the pitch encoder, given that rhythm embeddings performed considerably worse than the pitch ones in preliminary experiments. Hereinafter, we refer to these pitch embeddings as neural prosody embeddings (NPE).

3.1.2. Acoustic-Prosody Features

The Geneva Minimalistic Acoustic Parameter Set [29] is a set of functionals computed on top of low-level descriptors, designed specifically to provide a common ground for research in various areas of automatic voice analysis, namely paralinguistics. Here, we use the extended set (88 features). It has been widely used in paralinguistic tasks. In particular, [30] used it to represent style (prosody). We use these features here as a means to understand how much we gain from using pretrained models to extract features (transfer learning). In this work, we use the name acoustic-prosody features (APF) to refer to these features. APF features were scaled for zero mean and unit variance.

3.1.3. Content Embeddings

To encode content information, we extracted the embeddings from an ASR model at the end of the encoder for each utterance. We used a pre-trained model provided by SpeechBrain [31], with an acoustic model made of a transformer encoder and a joint decoder with CTC + transformer. This model was trained with LibriSpeech (EN) [32] and attains a WER of 2.46 for test-clean and 5.86 for test-other. As these embeddings have lengths proportional to the size of the utterance, to directly compare the content of utterances of different lengths, we computed the time average of these embeddings for each utterance. We used these embeddings to check whether the information encoded in the NPE was actually not content instead of pitch. The resulting vectors, Content Embeddings (CE) have length 768. As all sounds of the words in the intonation data set exist in English, we used the same model to extract data for both datasets.

3.2. Clustering metrics

3.2.1. Silhouette Coefficient

The Silhouette Coefficient [33] measures how the samples in a data set are similar to the ones of its own class when compared to samples in other classes. It ranges from -1 to 1, where negative values mean assignments to the wrong cluster, 0 indicates cluster overlap, and 1 indicates the best separation possible. For each sample x, the Silhouette Score is

$$S = \frac{b(x) - a(x)}{max(a(x), b(x))}\,, \tag{1}$$

where $a(x)$ is the mean intra-cluster distance and $b(x)$ is mean nearest-cluster distance. The Silhouette Coefficient is the mean Silhouette Score of all samples.

3.2.2. Geometrical Separability Index

Thornton Index (TH) [34], also known as Geometrical Separability Index (GSI), corresponds to the fraction of a set of points which have the same classification labels as their first neighbour. It can be described as

$$GSI = \frac{1}{n}\Sigma_{i=1}^{n} f(x_i, x_i')\,, \tag{2}$$

where x_i' is the nearest neighbour of x_i, n is the number of points, f is a function that is 1 if x_i and x_i' belong to the same class and zero otherwise. Therefore, GSI will tend to 1 if opposite labels exist in well-separated groups [35].

4. Experiments

4.1. Corpora

4.1.1. Intonation Data Set

We used a small intonation data set, comprising 20 original stimuli recorded by a native female speaker of Standard European Portuguese and reproductions of it by 17 different speakers. The 20 original stimuli correspond to the possible combinations of five words: *Banana*, *Bolo*, *Gelado*, *Leite*, and *Ovo*[1] with four different intonations: *Declarative*, *Interrogative*[2], *Pleasure*, and *Displeasure* (v. Figure 1). Each utterance corresponds to one word uttered with one particular intonation. In total, there are 340 imitation utterances, of which 320 are labelled as good, which are the ones we considered here.

Although limited, the Intonation Data Set is the most reliable set we have in terms of prosody assessment, as it con-

[1]The translation of these words in English is, respectively, *Banana*, *Cake*, *Ice Cream*, *Milk*, and *Egg*.

[2]Although single-worded, the intonation of these interrogative utterances corresponds to typical yes-no question.

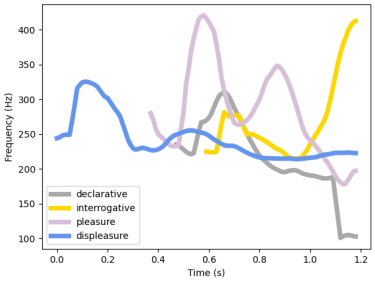

Figure 1: *F0 contours of the stimuli for the word "banana".*

tains different intonation patterns for the same word, which allows for a better control of what is actually varying (prosody or phonemes).

4.1.2. VCTK Data Set

We have also used the CSTR VCTK Corpus [12]. It comprises 110 English speakers of various accents, with approximately 400 utterances per speaker. These utterances correspond to the rainbow passage, an accent elicitation paragraph ("Please call Stella.") – common to all speakers – , and a newspaper text, different for each speaker. We use "sentence" to refer to a specific text, and "utterance" or "sample" to each realization of a sentence.

4.2. Experimental Set Up

4.2.1. Model training

The SpeechSplit model was trained with the default parameters, using 20 speakers from the VCTK [12] corpus during 400k iterations. Like in [11], we use 80% of the sentences from these speakers for training the model. Note that all the data from these 20 speakers is removed for the following analysis experiments.

4.2.2. Intonation Comparison

We considered the four classes in the intonation data set as well as the interrogative class against the other three. We took the later split as we know that these sets bear contrasting intonation patterns: whereas intonation has a final up pitch, the other three have mostly a final down pitch (v. Figure 1).

4.2.3. Contrastive Sentence Comparison

Although it is known that syntax does not have a one-to-one mapping with prosody, it is also known that there is some correspondence through the syntax-prosody interface [36]. For this reason, we analysed contrastive sentence groups: simple vs. complex (S/C), i.e. sentences with more than one main verb vs sentences with only one verb; yes-no questions vs. declaratives (Y-N/D); and sentiment-contrastive sentences, i.e. positive vs. negative (+/-). The sentences were selected and validated by an expert linguist.

In some of these combinations, one class largely outnumbered the other. We therefore randomly chose samples from the largest class, so as to equal the size of the smallest. The results correspond to the average results for five different random selections.

4.3. Results

For the intonation experiments and the contrastive sentence experiments, results are in Table 1 and in Table 2, respectively. The plots of the t-SNE projections are in Figure 2 and Figure 3 (only for NPE). The distributions on the t-SNE plots were verified to be constant throughout different seedings.

Table 1: *Clustering metrics for the Intonation data set.*

	NPE		APF		CE	
	all	int/oth	all	int/oth	all	int/oth
SC ↑	0.086	**0.214**	0.019	0.027	0.006	0.047
GSI ↑	0.762	**0.956**	0.406	0.668	0.682	0.850

5. Discussion

5.1. Intonation Comparison

In Table 1 values are always better when comparing interrogatives to other intonations. As well, values for NPE are always better than their APF and CE counterparts. Since we know that these intonation types are contrastive, we assert that NPE provide the best clustering. Differences between pleasure and displeasure are strongly encoded with rhythm and energy, which do not seem to be sufficiently represented in these embeddings, according to the plot in Figure 2. This is a good indicator that these embeddings are actually encoding intonation and not other prosodic components.

In the t-SNE projections of APF and CE (Figure 2, columns 2 and 3), there is also some tendency for points of the same colour to be grouped together, corresponding to the values in Table 1. This indicates that there is also prosodic encoding in these embeddings. However, in none of these are the intonation clusters as separate as they are in NPE (even when we consider the four of them). The fact that male and female utterances are mixed in the plots (Figure 2, column 1) seems to show that what is being encoded is actually the variation of pitch (intonation) and not the absolute values of pitch, which is desirable for the purpose of assessment.

On the t-SNE plot of APF, not only there is little overlap between utterances of male and female speakers, but also there is a small cluster of the darker points, corresponding to the same (reference) speaker. This indicates that these features model other speaker traits than gender more accurately than intonation. These features are functionals of low-level descriptors, which leads to the loss of relevant local variations in the utterance, thus better modeling constant traces of the utterance (as speaker identity). Indeed, intonation seems to be secondary for

Table 2: *Clustering metrics for contrastive sentences of VCTK: Simple vs Complex (S/C), Yes/No vs. Declaratives (Y-N/D) and Positive vs negative (+/-).*

		NPE	APF	CE
S/C	SC ↑	0.046	0.010	**0.054**
	GSI ↑	**0.705**	0.564	0.681
Y-N/D	SC ↑	**0.035**	0.009	0.020
	GSI ↑	0.706	0.590	**0.883**
+/-	SC ↑	0.001	0.005	**0.033**
	GSI ↑	0.542	0.550	**0.924**

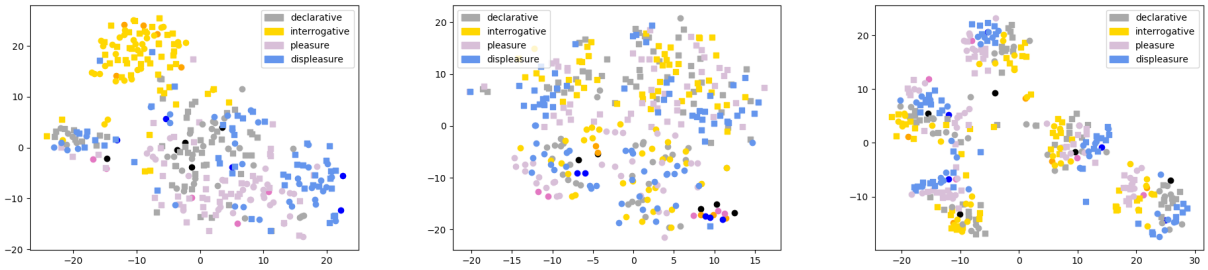

Figure 2: *t-SNE projections of the Intonation data set. Left to right: NPE, PF, CE. Darker points correspond to stimuli. Squares: male speakers, circles: female speakers.*

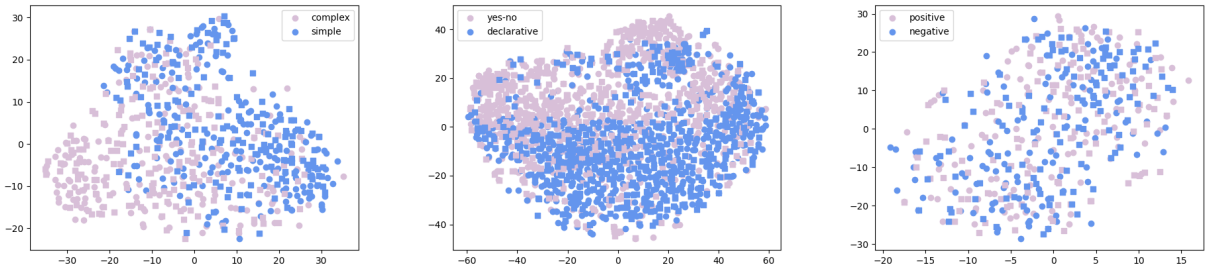

Figure 3: *t-SNE projections of NPE for contrastive sentences of VCTK. Left to right: Simple vs Complex; Yes-no questions vs Declarative; Positive vs Negative Sentiment. Squares: male speakers, circles: female speakers.*

the clusterings of APF and CE, as gender and word tend to account for most of it (respectively).

5.2. Contrastive Sentence Comparison

Regarding the NPE features, we find that SC and GSI metrics are considerably higher for S/C and Y-N/D groups of sentences than for +/-. This means that these features allow for some clustering in the cases of S/C and Y-N/D, but not for +/-, where results hardly surpass chance level. We consider these results acceptable, if we consider that there is not necessarily a contrast intonation between positive and negative sentences, mainly in the case of read speech. In the case of S/C and Y-N/D contrastive groups, the GSI scores close to 0.7 are considerably above chance level. This means that the separability of these classes is relevant. In the corresponding plots, we can find some regions where there is a clear separation, although there is also a considerable overlap between the two classes.

For the APF features, we notice that the SC score is always small, while GSI is quite close to chance level, meaning that there is no considerable tendency for sentences of the same class to stay closer than further from the ones of a different class.

The CE features provide results which tend to be similar or even better that NPE. This can be related to lexical cues that allow to discriminate between sentence-types.

5.3. Intonation vs. Contrastive Sentence Comparison

The data used in the two sets of experiments are very different in what comes to prosodic content. The intonation data set is intentional and therefore prosodic cues are accentuated, whereas VCTK is only read speech, which leads to less expressiveness, and to flatter intonation in general. It is also important to keep in mind that the neural prosody embeddings used in this work have been designed for fixed length utterances of 3s maximum, and afterwards adapted by us for variable length. Currently, with this architecture, it is possible that prosodic cues are not fully encompassed, as the relevance of the intonation is uniformly sampled throughout the utterance, which is not aligned with normal prosodic production.

6. Conclusions

In this work, we have compared neural prosody embeddings trained for prosody transfer with paralinguistics handcrafted features and content embeddings, in order to understand whether these embeddings could be useful for prosody assessment tasks, and how they compared to other features. We have understood that they provide good separability for small utterances, as we saw in the intonation data set. We attained a GSI of 0.956 when comparing interrogatives to other sentence-types. For longer utterances, we have evidence that prosody is being encoded, with a GSI as high as 0.706 for yes-no questions vs declaratives, where prosodic differences are evident. Although promising, more research is needed to find reliable ways to compare the prosody of long variable-length sentences. This work has shown that NPE are informative, but it also raises several questions since embbeddings are known to capture context, but intonation variance is still not well captured with this method. Future work will tackle distinct processes to capture pitch variance and its implications to L2 assessment and will also tackle variable sentence length, areas which are still very challenging in the literature.

7. References

[1] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron," in *ICML*, 2018.

[2] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[3] E. Pyshkin, J. Blake, A. Lamtev, I. Lezhenin, A. Zhuikov, and N. Bogach, "Prosody training mobile application: Early design assessment and lessons learned," in *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, vol. 2. IEEE, 2019, pp. 735–740.

[4] K. Hirschi, O. Kang, C. Cucchiarini, J. Hansen, K. Evanini, and H. Strik, "Mobile-Assisted Prosody Training for Limited English Proficiency. Learner Background and Speech Learning Pattern," 2020.

[5] A. W. Darkins, V. A. Fromkin, and D. F. Benson, "A characterization of the prosodic loss in Parkinson's disease," *Brain and Language*, vol. 34, no. 2, pp. 315–327, 1988.

[6] H. N. Jones, "Prosody in Parkinson's disease," *Perspectives on Neurophysiology and Neurogenic Speech and Language Disorders*, vol. 19, no. 3, pp. 77–82, 2009.

[7] S. Frota, M. Cruz, R. Cardoso, I. Guimarães, J. J. Ferreira, S. Pinto, and M. Vigário, "(Dys) Prosody in Parkinson's Disease: Effects of Medication and Disease Duration on Intonation and Prosodic Phrasing," *Brain Sciences*, vol. 11, no. 8, p. 1100, 2021.

[8] D. M. Chun, *Discourse intonation in L2: From theory and research to practice*. John Benjamins Publishing, 2002, vol. 1.

[9] A. Cutler, *Native listening: Language experience and the recognition of spoken words*. Mit Press, 2012.

[10] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "Automatic assessment of non-native prosody for English as L2," 2010.

[11] K. Qian, Y. Zhang, S. Chang, M. Hasegawa-Johnson, and D. Cox, "Unsupervised speech decomposition via triple information bottleneck," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7836–7846.

[12] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2016.

[13] J. P. Arias, N. B. Yoma, and H. Vivanco, "Automatic intonation assessment for computer aided language learning," *Speech communication*, vol. 52, no. 3, pp. 254–267, 2010.

[14] J. Cheng, "Automatic assessment of prosody in high-stakes English tests," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[15] M. Ma, K. Evanini, A. Loukina, X. Wang, and K. Zechner, "Using f0 contours to assess nativeness in a sentence repeat task," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[16] Q.-T. Truong, T. Kato, and S. Yamamoto, "Automatic assessment of L2 English word prosody using weighted distances of f0 and intensity contours." in *Interspeech*, 2018, pp. 2186–2190.

[17] M. Julião, A. Abad, and H. Moniz, "Comparison of heterogeneous feature sets for intonation verification," *International Conference on Computational Processing of the Portuguese Language*, pp. 13–22, 2020.

[18] D. Escudero-Mancebo, C. González-Ferreras, L. Aguilar, E. Estebas-Vilaplana, and V. Cardeñoso-Payo, "Exploratory use of automatic prosodic labels for the evaluation of Japanese speakers of L2 Spanish," 2016.

[19] K. Li, X. Wu, and H. Meng, "Intonation classification for L2 English speech using multi-distribution deep neural networks," *Computer Speech & Language*, vol. 43, pp. 18–33, 2017.

[20] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[21] M. Elyasi and G. Bharaj, "Flavored Tacotron: Conditional learning for prosodic-linguistic features," *arXiv preprint arXiv:2104.04050*, 2021.

[22] Y. Zou, S. Liu, X. Yin, H. Lin, C. Wang, H. Zhang, and Z. Ma, "Fine-grained prosody modelling in neural speech synthesis using ToBI representation," in *Interspeech*, 2021.

[23] K. E. Silverman, M. E. Beckman, J. F. Pitrelli, M. Ostendorf, C. W. Wightman, P. Price, J. B. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody." in *ICSLP*, vol. 2, 1992, pp. 867–870.

[24] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5210–5219.

[25] J. Chou, C. Yeh, and H. Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," *arXiv preprint arXiv:1904.05742*, 2019.

[26] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," *arXiv preprint arXiv:1905.13561*, 2019.

[27] D. Xin, T. Komatsu, S. Takamichi, and H. Saruwatari, "Disentangled Speaker and Language Representations Using Mutual Information Minimization and Domain Adaptation for Cross-Lingual TTS," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6608–6612.

[28] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[29] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[30] Z. Hodari, O. Watts, S. Ronanki, and S. King, "Learning interpretable control dimensions for speech synthesis by using external data." in *Interspeech*, 2018, pp. 32–36.

[31] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[32] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[33] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[34] C. Thornton, "Separability is a learner's best friend," in *4th Neural Computation and Psychology Workshop, London, 9–11 April 1997*. Springer, 1998, pp. 40–46.

[35] A. Acevedo, S. Ciucci, M. Kuo, C. Durán, and C. V. Cannistraci, "Measuring group-separability in geometrical space for evaluation of pattern recognition and embedding algorithms," *arXiv preprint arXiv:1912.12418*, 2019.

[36] R. Bennett and E. Elfner, "The syntax–prosody interface," *Annual Review of Linguistics*, vol. 5, pp. 151–171, 2019.