Characterizing speaking style with automatic prosodic labels, applications in computer assisted pronunciation training

David Escudero-Mancebo

ECA-SIMM Research Group
University of Valladolid, Spain
descuder@infor.uva.es

Abstract

This communication presents a novel methodology to characterize the style of different speakers or groups of speakers described in detail in [1] and its application to L2 pronunciation scoring. This methodology uses sequences of prosodic labels (automatic Sp_ToBI labels) to compare and differentiate these speaking styles. A set of metrics based on conditional entropy is used to compute the distance between two speakers or group of speakers depending on the use of sequences of prosodic labels. Additionally, the most contrastive sequences of labels are identified as characteristic patterns of the speaking styles represented in a given corpus. When this methodology is applied to a corpus of radio news items, the result is that the most frequent prosodic patterns coincide with those previously characterized in studies about radio style.

There are several approaches in the state of the art that face up the problem of evaluating L2 prosody [2]. Most systems are based on comparing the prosodic acoustic characteristics of L2 utterances (like F0, duration and energy) with the corresponding features of native speakers (generally with the ones of a golden speaker who is considered to use *the correct pronunciation*). These approaches have an important limitation that has to do with the under representation of variety in prosody: the same prosodic function can be represented with more than one prosodic form [3]. This is challenging for CAPT systems because two prosodic productions of the same text can be different but valid at the same time. To face up this problem, we defend a double strategy: on the one hand, we have used prosodic labels (no directly prosodic acoustic features) to compare utterances; on the other hand, L2 utterances have not only been compared with those of a single golden speaker but with the productions of a set of reference speakers.

The efficiency of using prosodic labels (a set of symbols for transcribing the intonation patterns and other aspects of the prosody of utterances) has been well established in the context of L2 assessment [4, 5, 6]. Related to this, the ToBI system is a broadly accepted framework for the transcription of prosodic phenomena. It was originally developed for English, based on Pierrehumberts autosegmental model, but since then it has been applied to a large number of languages, among them Spanish [7].

In [4], an experiment of style identification was presented by using the Automatic ToBI labels described in [8]: the results showed 95% of accuracy. When a given utterance is labeled with prosodic labels, its representation is simplified since the labels include symbolic information that specifies the relevant prosodic functions present in the utterance. The automatic prosodic labeling systems are prepared to process prosodic variety as they are trained with data that reflects the form-function multiplicity. In [1], we used the automatic Sp_ToBI classifier presented in [9] to characterize radio broadcasting prosodic style by measuring the mutual information between sequences of prosodic labels. In [10] we followed a similar approach to compute distances between native and non-native speakers by improving the mutual information metric used in [1] and by applying normalization that takes into account the joint entropy of the labels of the different type of speakers. The results show that these new metrics permit to identify non-native speakers with a degree of confidence that is statistically significant. The results are consistent with the a-priori expected improvement on the pronunciation as the pronunciation exercises are repeated.

1. References

[1] D. Escudero, C. González, Y. Gutiérrez, and E. Rodero, “Identifying characteristic prosodic patterns through the analysis of the information of sp_tobi label sequences,” *Computer Speech and Language*, vol. 45, pp. 39 – 57, 2017.

[2] O. Kang and D. Johnson, “The roles of suprasegmental features in predicting english oral proficiency with an automated system,” *Language Assessment Quarterly*, pp. 1–19, 2018.

[3] D. Hirst, “Form and function in the representation of speech prosody,” *Speech Communication*, vol. 46, no. 34, pp. 334 – 347, 2005.

[4] A. Rosenberg, “Symbolic and direct sequential modeling of prosody for classification of speaking-style and nativeness.” in *INTERSPEECH*, 2011, pp. 1065–1068.

[5] J.-m. Kim, “Annotation of a non-native english speech database by korean speakers,” *Speech Sciences*, vol. 9, no. 1, pp. 111–135, 2002.

[6] J. Tepperman, A. Kazemzadeh, and S. Narayanan, “A text-free approach to assessing nonnative intonation.” in *INTERSPEECH*, 2007, pp. 2169–2172.

[7] P. Prieto, *Transcription of intonation of the Spanish language*. Lincom Europa, 2010.

[8] A. Rosenberg, “AuToBI-a tool for automatic ToBI annotation.” in *Interspeech*, 2010, pp. 146–149.

[9] C. Gonzalez-Ferreras, D. Escudero-Mancebo, C. Vivaracho-Pascual, and V. Cardeñoso Payo, “Improving automatic classification of prosodic events by pairwise coupling,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 7, pp. 2045 –2058, sept. 2012.

[10] D. Escudero-Mancebo, C. Gonzlez-Ferreras, L. Aguilar, and E. Estebas-Vilaplana, “Automatic assessment of non-native prosody by measuring distances on prosodic label sequences,” in *Proc. Interspeech 2017*, 2017, pp. 1442–1446.