Interspeech 2018
2-6 September 2018, Hyderabad

Language-Dependent Melody Embeddings

Daniil Kocharov, Alla Menshikova

Saint Petersburg State University, Russia

kocharov@phonetics.pu.ru, menshikova.alla2016@yandex.ru

Abstract

The paper explores the perspectives of applying the distributional approach to prosodic typology of languages. The method discussed here is an adaptation of the distributional semantics approach, as suggested by Mikolov, to melodic features of speech. The paper contains a detailed description of the new method, as well as a comparison of five European languages (English, Czech, German, Russian, and Finnish) in terms of melody embeddings. The total amount of speech data was over 500 hours. The experimental results show that melody embeddings are language dependent. The proposed melody embedding model has shown reasonable results in language comparison.

Index Terms: prosodic typology, melody, distributed representations, embeddings

1. Introduction

The amount of digitized speech data for various languages increases every day. Nevertheless, the number of annotated speech resources grows in a much slower tempo. This causes an increase in interest towards unsupervised methods of speech analysis.

Recently several unsupervised data-driven methods for comparison of prosodic features of different languages have been proposed. M. Vainio and colleagues presented positive results on modelling prosody by means of Continuous Wavelet Transform of speech signal and using this information to compare several languages from Finno-Ugric and Indo-European families [1]. Based on small amount of speech data they achieved results reflecting genetic and contact relations among the processed languages. P. N. Zulu used prosodic features—pitch and intensity—to cluster South African languages from Germanic and Bantu languages [2]. His results on using pitch showed positive results in clustering languages in accordance with the genetic classification of languages.

The main issue of data-driven prosodic typology is that prosody is a complex object for analysis due too complex interactions between the units functioning at different levels, such as syllables, words and phrases. Similarly, written text combines syntax, morphology and lexis, and together they are used to convey the meaning of a sentence.

In textual domain the method of word embeddings is successfully used to model semantics of units on various levels: strings of symbols, words, phrases and sentences [3], [4], [5]. Word embeddings represent words as continuous vectors in a multi-dimensional space based on the hypothesis that words with similar meaning are used in similar contexts. The method has been successfully applied for language classification [6].

Thus we propose to apply the embeddings approach to the task of language comparison. The melody embedding model was introduced in [7]. The novelty of the approach is to train multidimensional melody embeddings purely on melodic information. It has shown good potential of capturing contextual information in the melodic domain.

We used speech data of five European languages (English, Czech, German, Russian, and Finnish) to test the perspectives of applying the distributional approach to prosodic typology. These languages represent two families: Finno-Ugric and Indo-European (Slavic and Germanic branches). They have distinct prosodic properties at all levels of prosodic hierarchy. The overview of prosody of English, Finnish, German and Russian is given in [8]. The prosody of Czech language is described in [9] and [10]. Other information on prosody of Finnish and Russian may be found in [11], [12]. The prosodic diversity of these languages makes them a good choice for our experiments. The total amount of speech data was over 500 hours with about 100 hours of speech material for each language.

The rest of the paper contains a detailed description of the new method; the information about speech material; the experimental results on language dependency of melody embedding models and on language comparison by means of embeddings of the most frequent melodic contours.

2. Method

The procedure of building the melody embedding model included the following three steps:

1. calculating the stylized melodic contour;
2. coding melodic information;
3. calculating the vector representation of melody.

2.1. Melodic contour stylization

At the first stage, fundamental frequency (F_0) was calculated. After the F_0 errors and periods of microprosody were automatically detected and eliminated from melodic contours, the F_0 values for the voiceless parts of the signal were added by means of linear interpolation. Then the contour was smoothed using Savitzky-Golay filtering with a second order polynomial in 5 sample windows [13].

Smoothed melodic contours were processed in non-overlapping 50 ms frames. Within each frame the F_0 movement range was calculated in semitones. The movement was defined as rising, falling or level based on the relative position of the F_0 maximum and minimum within the frame. Then the contour was split into sequences of frames with identical direction of F_0 movement; the range values were summed up across each sequence of frames.

2.2. Coding melodic information

In order to be able to apply text processing techniques, we must develop a system of translating the melodic information into some textual form—that is, a system for coding F_0 movements with textual characters.

10.21437/Interspeech.2018-1962

The existing coding methods, such as ToBI [14], Tilt model [15], INTSINT [16] and SLAM [17], have turned out to be too general for our purposes. ToBI and INTSINT describe only the most important points of the melodic contour. Tilt and SLAM describe contours in terms of several symbols and quantize the F_0 values in very broad intervals.

Thus, for our specific task, we decided to use our own coding scheme. Each value in semitones corresponded to a letter: 1—a, 2—b, 3—c, ..., 26—z. Positive values were coded by lower-case letters, and negative values—by upper-case letters. Level slopes, i.e. slopes with a range of zero semitones, was coded by the symbol '='. Thus, for example, the melodic contour consisting of three slopes '*2, 0, -3*' was coded by the string '*b=C*'. The resulting coded melodic contour has no information on slopes duration, it contains only information on their range.

After a series of preliminary experiments on relatively small amount of speech data we added some adjustments to the algorithm. First, we decided to discard those slopes that were single frame long, as they added random noise to the final model. Second, we confirmed that using temporal information in coding the melodic information significantly decreased the efficiency of the resulting embedding model; this might be due to a huge increase in the inventory of basic symbols—from 53 to about 400—as in this case each coding 'symbol' consisted of a letter and a value for slope duration (e.g. '*b2 =1 C5*').

2.3. Embedding

The melodic contours coded as strings of symbols were used as input for the embedding procedure. Following the tokenization-free approach of embedding representation [4], we split the symbol representation of each melodic contour into sequences of non-overlapping segments of random length ranging from k_{min} to k_{max} ('n-grams'). The result of such segmentation was a sequence of n-grams up to k_{max} symbols long. In order to provide a better coverage of symbol n-grams, each symbol representation was split *m* times. Then all the resulting variants of random segmentation were concatenated into a whole 'text'.

The embeddings learning was performed using the skip-gram objective—the method for predicting the surrounding n-grams from each symbol n-gram [3]. Words were represented as n-dimensional vectors, and the model was built using a neural network with a single hidden layer. The neural network was trained to minimize the negative log-likelihood:

$$\begin{aligned} & -\log P(w_{c-h},...,w_{c-1},w_{c+1},...,w_{c+h}|w_c) \\ & = -\sum_{j=0,j\neq h}^{2h} u_{c-h+j}^T v_c + 2h\log\sum_{k=1}^{|V|}\exp(u_k^T v_c), \end{aligned} \tag{1}$$

where w_i is a word in the vocabulary V; word w_c is taken within each context $c-h,...,c+h$; v_c and u_c are the input and the output vector representations of word w_c.

2.4. Implementation

In our experiments, a word length defined by k_{min} and k_{max} was from 1 to 4 symbols. The corpus was segmented 20 times. The dimension of v_c and u_c vectors was set to 50, while the considered context length equalled to 3 words.

We used a Python implementation of word2vec—Gensim [18] for learning embeddings of symbol n-grams. We trained word2vec skip-gram model on the material, with inter-pausal units presented as sentences and 'melodic units' (short melodic contours from 1 to 4 slopes long) as words.

Table 1: *Speech material used for model training per language*

Language	Total duration (hours)	Content
Czech	120	120 fragments of audio books, 177 podcasts, 83 broadcasts
English	130	29 interviews, 175 broadcasts
Finnish	90	105 podcasts
German	93	27 podcasts, 81 lectures, 19 interviews
Russian	120	150 podcasts

2.5. Material

The material used for training language-dependent melody embedding models consisted of unannotated speech recordings in five languages: Finnish, English, German, Czech, and Russian. Speech included reading, spontaneous and prepared speech. The material included podcasts, news broadcasts, lectures, audio books and interview recordings. The total duration of recordings was more than 550 hours of speech. Table 1 shows the distribution of speech data among languages.

No preprocessing of data was conducted, except for cutting off the first minute of each recording, in order to avoid the inclusion of musical introductions in the material.

3. Results

3.1. Language-Dependent Models

We assume that the distribution of melodic contours in speech depends on the language. On the one hand, the same F_0 contour in different languages can serve different linguistic functions. As a result, the frequency of a melodic contour depends on the frequency of the word or phrase over which the contour typically stretches. On the other hand, the distribution of melodic contours may depend on the general prosodic features of speech characterizing a given language.

To illustrate language dependency of the presented model, we selected one melodic contour and calculated all the contours whose representation vectors were closest to the vector of the given melodic contour. We have chosen one of the most frequent rising movements: 3 semitone rise followed by level tone, labelled as '3,0' ('c='). The frequency of this melodic contour varies among languages. Its frequency rank among all the F_0 contours in our speech data is as follows:

- Czech: 15
- German: 16
- Russian: 19
- English: 20
- Finnish: 27

Figures 1–4 show four nearest contours to the '3,0' contour for Czech, Russian, English, German and Finnish languages. The plots include the information about cosine similarity between the input contour and its neighbours.

Note that the presented method enables us to compare the distance between contours of different length, as it compares not the contours but the embedding vectors, which are of the same

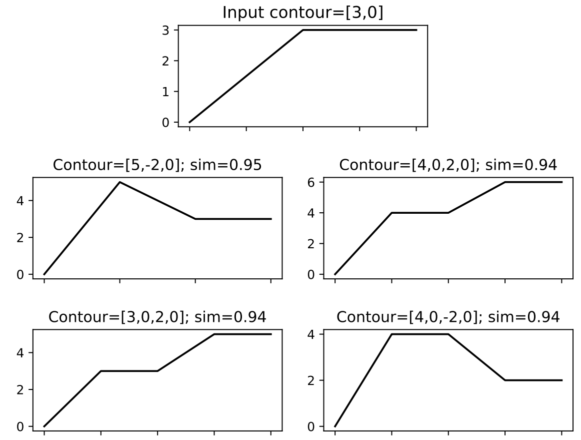

Figure 1: *Czech. Nearest contours to '3,0' contour*

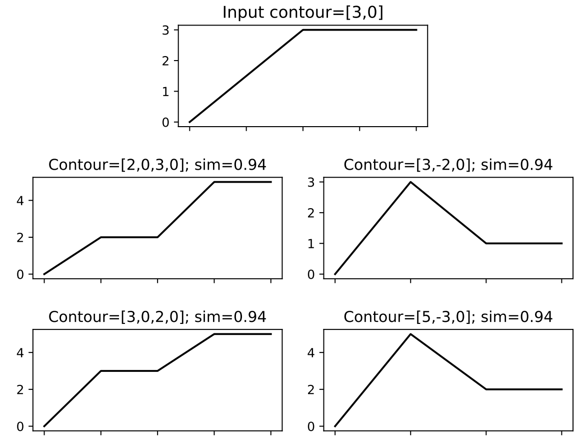

Figure 2: *English. Nearest contours to '3,0' contour*

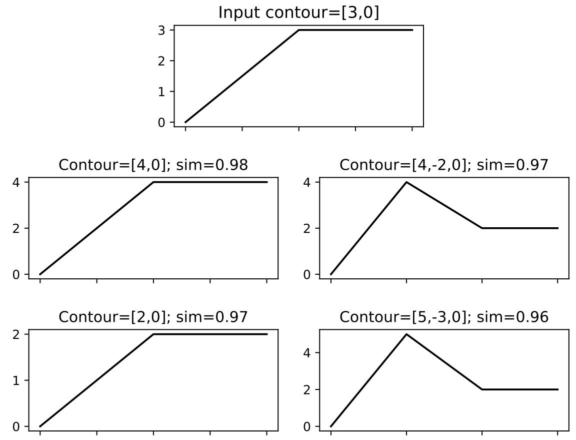

Figure 3: *Russian. Nearest contours to '3,0' contour*

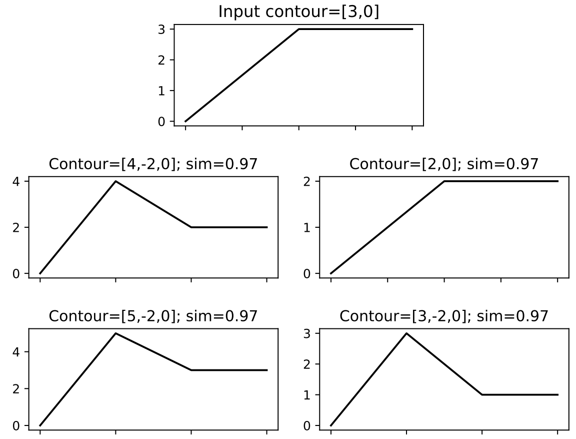

Figure 4: *German. Nearest contours to '3,0' contour*

dimensionality independent of the length of the corresponding contours.

The plots show that all nearest contours start with a rising melody and end with a level melody—as in the input contour. There are three types of middle part geometry: (1) no middle part, the contour has the same 'rise-level' melody with an amplitude close to 3 semitones; (2) a fall with F_0 amplitude smaller than the one of the initial rise; (3) a complex rise consisting of several rising-level steps. The exact nearest contours and their distance to the given '3,0' contour depends on the languages.

Judging by geometry of the nearest contours, the plots for Russian and German look closer to each other than to others, while plots for Czech and English form another cluster. The plot for Finnish is in between these two clusters.

3.2. Comparing Languages by Means of Distributional Melody Model

In the domain of distributional semantics, word embeddings have proved useful for comparing languages [6]. The diversity of distributional melody models among languages allows to assume that it might be reasonable to use them for language comparison.

The general idea is to select a core subset of vocabulary within the units that are the most important in all compared languages. Then the mutual distances among these core units for the given language could characterize this language.

As core units we selected the most frequent contours in the embedding models. The frequency ranks of contours are language-dependent; furthermore, some contours frequent for one language may not be in the model for another language. We calculated mean ranks for all contours that existed in all language-specific models. Then we selected seven top-ranked contours in this list as representing the distributional model for this language. For each language we calculated a vector of mutual distances among these contours. To estimate mutual distance among the given languages, we calculated Euclidean distance between these vectors. Figure 6 shows a dendrogram illustrating the clustering of languages based on the calculated estimates.

4. Conclusions

The proposed melody embedding model has shown reasonable results. First of all, the Finnish language differs much from the other languages. It positively captures the fact that in terms of prosody Finnish differs from other languages more than they differ from each other. The other four languages fall into two clusters: (1) Russian and German, (2) English and Czech. The results are in accordance with the recent findings of M. Vainio and colleagues [1]. They applied the unsupervised method based on wavelet analysis of prosodic data to language com-

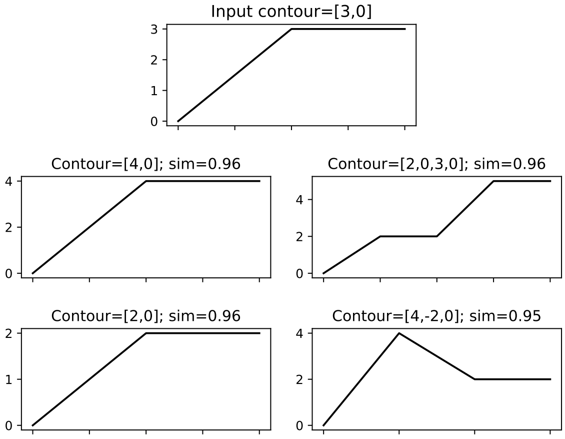

Figure 5: *Finnish. Nearest contours to '3,0' contour*

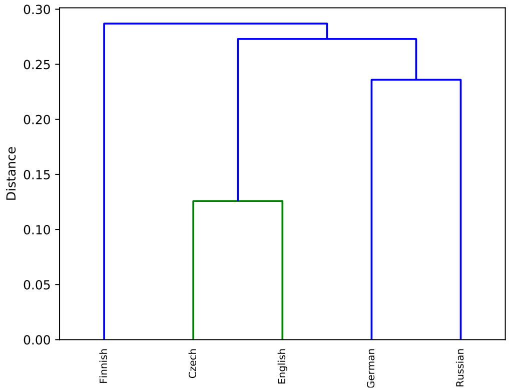

Figure 6: *The dendrogram with language distances calculated by means of melody embeddings*

parison, and clustered the languages in a similar way. Their results also showed that Russian and German are close to each other. While they did not analyze Czech and English, they presented the results for Slovak, which appeared to be very distant from Russian, even both of them are Slavic. In our results, in a similar way, Russian and Czech are seen as rather distant. Our findings, as well as the results presented by M. Vainio and colleagues, show that the use of prosodic features in data-driven approaches may lead to the results that do not fully reflect genetic relations among languages.

Due to a small number of languages, we are yet not able to draw more general conclusions on prosodic typology of European languages. At the same time, adding more language may shed more light on the connections between neighbouring countries and between genealogically related languages in terms of prosody. We assume that the presented method is very promising. We see many possible applications for the proposed unsupervised method of prosody modelling that include: solving tasks in the field of prosody typology; processing the speech of languages that lack annotated speech resources; modelling prosodic variability for text-to-speech synthesis.

5. References

[1] J. Simko, A. Suni, K. Hiovain, and M. Vainio, "Comparing languages using hierarchical prosodic analysis," in *Proceedings of Interspeech*, 2017, pp. 1213–1217.

[2] P. N. Zulu, "Language classification using prosodic features: Comparing intensity and pitch," in *Proceedings of Pan African International Conference on Information Science, Computing and Telecommunications*, 2013, pp. 116–121.

[3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of Neural Information Processing Systems 2013*, 2013.

[4] H. Schütze, "Nonsymbolic text representation," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 1, 2017, pp. 785–796.

[5] N. Kalchbrenner, L. Espeholt, K. Simonyan, A. van den Oord, A. Graves, and K. Kavukcuoglu, "Neural machine translation in linear time," *CoRR*, vol. abs/1610.10099, 2016. [Online]. Available: http://arxiv.org/abs/1610.10099

[6] S. Eger, A. Hoenen, and A. Mehler, "Language classification from bilingual word embedding graphs," in *Proceedings of the 26th International Conference on Computational Linguistics COLING: Technical Papers*, 2016, pp. 3507–3518.

[7] D. Kocharov and A. Menshikova, "Distributed representation of melodic contours," in *Proceedings of Speech Prosody*, 2018.

[8] D. Hirst and A. Di Cristo, Eds., *Intonation Systems. A Survey of Twenty Languages*. Cambridge University Press, 1998.

[9] T. M. Nikolaeva, *Frazovaya intonatsiya slavyanskih yazykov*. Moscow: Nauka, 1977, in Russian.

[10] Z. Palková, *Fonetika a fonologie češtiny*. Praha: Karolinum, 1994.

[11] K. Suomi, J. Toivanen, and R. Ylitalo, *Finnish sound structure. Phonetics, phonology, phonotactics and prosody*. University of Oulu, 2008.

[12] N. Volskaya and T. Kachkovskaia, "Prosodic annotation in the new corpus of Russian spontaneous speech CoRuSS," in *Proceedings of Speech Prosody 2016*, 2016.

[13] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry*, vol. 39, no. 8, pp. 1627–1639, 1964.

[14] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody," in *The Second International Conference on Spoken Language Processing, ICSLP 1992*, 1992, pp. 867–870.

[15] P. Taylor, "The rise/fall/connection model of intonation," *Speech Communication*, vol. 15, no. 1, pp. 169–186, 1994.

[16] D. Hirst, A. Di Cristo, and R. Espesser, "Levels of representation and levels of analysis for the description of intonation systems," in *Prosody: Theory and Experiment*, M. Horne, Ed. Springer Netherlands, 2000, pp. 51–87.

[17] N. Obin, J. Beliao, C. Veaux, and A. Lacheret, "SLAM: Automatic stylization and labelling of speech melody," in *Proceedings of Speech Prosody 2014*, 2014.

[18] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010, pp. 45–50.