

A. A. Batašchikov, E. A. Zolotareva, V. D. Ivanov, A. V. Stolpovskiy
APPLICATION OF DTW ALGORITHM FOR INVESTIGATION OF RUSSIAN PHONEMES
VARIATIONS

FSSE SRI “Specvizavtomatika”
 Russia, 344002 Rostov-on-Don, Gazetnyi lane, 51
 Tel.: (863) 297-50-84; Fax: (863) 297-50-84
 e-mail: asni@asni.rsu.ru

DTW (Dynamic Time Warping) algorithm is usually applied for estimation of likelihood degree for two time successions and is also used for their comparative alignment (building a correspondence between time segments of the first and second sequences). The present paper regards application of DTW for automatic phoneme segmentation of the present Russian recordings classified according to a phrase uttered by the speaker. The applied method involved alignment of each recording from the class against one reference (manually segmented) recording. MFSC (Mel Frequency Spectral Coefficient) u PLP (Perceptual Linear Prediction) feature vector sets were used as vector representation of speech. We also performed quality analysis of the alignment resulting from the present method and obtained vowel variation statistic probability estimate for different speakers based on the received phoneme segmentation.

The goal of the present paper is to check efficiency of DTW (dynamic time warping) algorithm [1] for obtaining automatic phoneme segmentation of speech recordings. DTW algorithm was used for comparative alignment of the same phrases belonging to different speakers followed by statistic analysis of MFSC (Mel Frequency Spectral Coefficient) and PLP (Perceptual Linear Prediction [2]) coefficients for Russian vowels. Within each set of the same phrases recorded by the same speaker (reference recording) the boundaries of the phoneme in question were found by experts, in all the other phrases we obtained the phoneme boundaries automatically by the presented algorithm.

We understand comparative alignment (or just alignment) of two recordings containing the same utterance as finding correspondence between time moments with similar phonetic data. Taking into account the fact that different speakers have different phrase durations (speech tempo) and relative duration of phrase components differ, we should change the coefficient characterizing ratio between speech tempos of two speakers within the whole phrase. That is why we turned to DTW algorithm. We built feature arrays (we used MFSC as feature vectors) for reference and investigated recordings and then applied a version of automatic loudness regulation algorithm to compensate difference on loudness of different speech parts.

The heart of DTW algorithm is finding such a correspondence between vectors of the two sequences under which the sum of distances between the compared aligned vectors is minimal by some metrics, each vector belonging to one of the sequences can have a corresponding element from other sequence whose index is not less than the index of the element compared to the previous vector of the sequence. It means that we receive an opportunity to “slow” and to “speed up” time for one of the sequences against the other one and impossibility of “coming back”.

DTW algorithm is realized by recursion building of matrix A with $n_1 \times n_2$ dimensions, where n_m – the number of vectors in m -th array. Element A_{ij} contains index of element A_{kl} , at that pair of indexes (k, l) is chosen from set $\{(i-1, j), (i-1, j-1), (i, j-1)\}$, so that expression $d_{ij} = \rho_{ij} + d_{kl}$ takes the minimal value, ρ_{ij} – distance between i vector of the first array and j vector of the second array by the chosen method. Here we take Euclidian metrics. Value d_{kl} was calculated at a previous stage. Having built trajectory from $A_{n_1 n_2}$ to A_{11} according the pointers we find correspondences between feature vectors and hence time moments in the given recordings.

In our experiment we used sets of 5 recordings with the same phrases, pronounced by 20 speakers (10 male and 10 female). We chosen 5 phonemes corresponding to accentuated Russian vowels: “a”, “o”, “u”, “e”, “i”. Alignment procedure was performed twice to obtain a better quality control, two different speakers (male and female) being chosen as references. The alignment results

we checked by automatic singling out of phonemes using the received correspondence between time moments in the data and the reference recordings and their listening.

The obtained results lead us to the conclusion that DTW algorithm with MFSC coefficients as features together with automatic loudness regulation permit to gain high accuracy when aligning vowel phoneme boundaries for recordings belonging to different speakers and containing the same utterance if the following conditions are met:

- 1) The recordings don't have loud noises;
- 2) Location of between word pauses corresponds in the recordings of both speakers.

Fig. 1, 2 represent results of statistic estimate of the regarded phoneme variations by the example of normalized MFSC and PLP coefficients. The present coefficients were calculated for a middle segment of the chosen phoneme part. Window duration came to 0,04 sec. for MFSC and 0,064 sec. – for PLP.

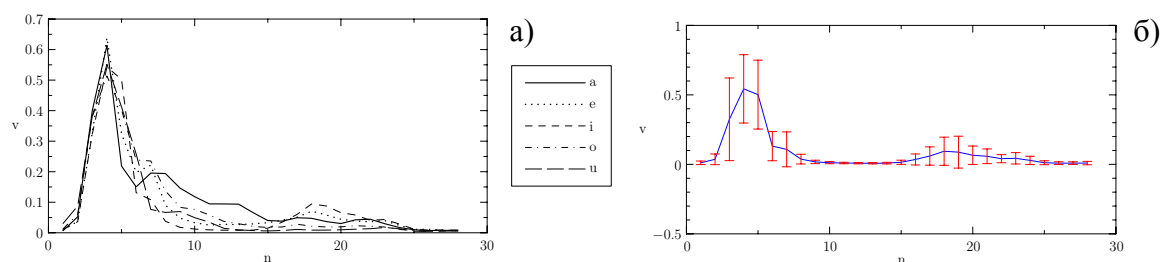


Fig. 1 – Mean values of normalized MFSC coefficients (a) and mean-square deviation for phoneme “i” (b).

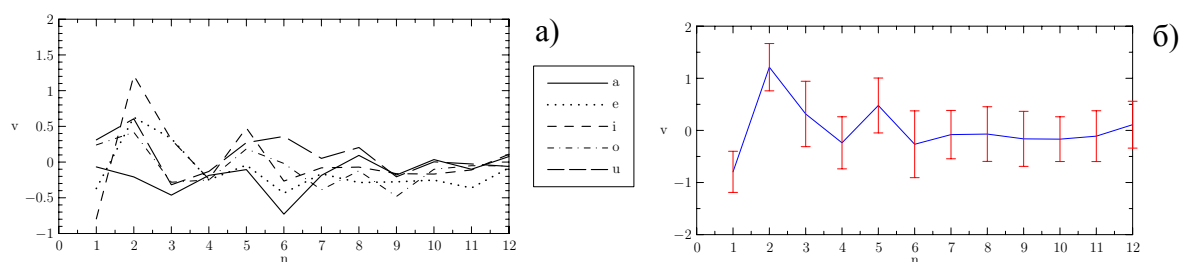


Fig. 2 – Mean values of PLP coefficients (a) and mean-square deviation for phoneme “i” (b).

REFERENCES

1. C. S. Myers, L. R. Rabiner. A comparative study of several dynamic time-warping algorithms for connected word recognition. The Bell System Technical Journal, 60(7):1389-1409, September 1981.
2. H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. The Journal of the Acoustical Society of America – April 1990 – Volume 87, Issue 4, pp. 1738–1752.