Automatic Alignment of Phonetic Transcriptions for Russian

Daniil Kocharov

Department of Phonetics, Saint-Petersburg State University
Universitetskaya Emb., 11, 199034, Saint-Petersburg, Russia
kocharov@phonetics.pu.ru
http://www.phonetics.pu.ru

Abstract. This paper presents automatic alignment of Russian phonetic pronunciations using the information about phonetic nature of speech sounds in the aligned transcription sequences. This approach has been tested on 24 hours of speech data and has shown significant improvement in alignment errors has been obtained in comparison with commonly used Levenstein algorithm: the numbers of error has been reduced from 1.1 % to 0.27 %.

Keywords: automatic alignment, phonetic transcription, Russian.

1 Introductions

The goal of the work described in this paper is to align effectively two sequences of phoneme labels (phonetic transcriptions) that describe the same speech signal. There are two main use cases for aligning phonetic transcriptions and measuring distance between. The first one is a research of how various people read or speak the same text, e.g. dialectic [1] and sociolinguistic [2] studies. The other one is the alignment of speech transcriptions produced by different transcribers, e.g. in automatic speech recognition systems [3] or while annotating speech corpora [4].

The current work has been done as a part of a research on speaker individual characteristics. The aim has been to register and quantitatively measure the deviation of various native speakers of Russian from Standart Russian pronunciation. To make a correct comparison of individual pronunciations these pronunciation has to be well aligned. The nature of phonemes, relations between them, the behaviour of phonemes in fluent speech under different conditions should be considered to perform a perfect alignment of phoneme sequences.

Automatic transcriptions aligners, that used knowledge of phoneme relations, have been done for many languages, including Basque [5], Dutch [6], English [3], Norvegian [1], Spanish [2]. Such an aligner that has been developed for Russian is presented in the paper. It is based on the usage of sets of phonemes that could substitute each other, be inserted into speech or not pronounced in continuous speech.

Section 2 describes the basic ideas of the presented aligner. Section 3 presents the phoneme set that has been considered in the aligner. The achieved results are shown in section 4.

A. Ronzhin et al. (Eds.): SPECOM 2014, LNAI 8773, pp. 123–128, 2014.

2 Transcription Alignment

There are different ways of aligning transcriptions, but using dynamic programming is the most common approach, including Levenstein algorithm [7] and Hirschberg's algorithm [8]. The basic setup is that a cost of any substitution, deletion or insertion is '1', and cost of match is '0'. These algorithms do not distinguish the substitution of similar sounds from substitution of very different sounds and do not take into account that ellision or epenthesis of some sounds is highly probable.

There have been efforts to measure phonetic difference more precisely assuming that a cost of substitution of one phoneme by another should depend on phonetic distance between these phonemes.

Consider a phoneme to be represented as a vector of articulatory features, than the phonetic distance between two phonemes is a sum of absolute differences between feature values of the phonemes [6]. The phonetic distance may be dependent on pointwise mutual information, the number of times phonemes corresponded to each other in aligned transcriptions [9]. In [3] the better results were obtained calculating phonetic distance on the basis of misrecognitions of phones by ASR phone-decoder in comparison with using phonemes perceptual similarity and phonological similarity.

The proposed approach of improving phonetic transcription alignment is based on the idea to define sets of phonemes that are highly probable to substitute each other. The substitution cost for phonemes within a set should be less than substitution cost for phonemes from different sets. This cost reduction is equal for all the sets and for this work is equal to 0.1. Thus, the substitution of phonemes within a set cost 0.9, and the substitution of phonemes across sets cost 1. The cost of probable phoneme deletions and insertions is also reduced to 0.9.

The next section presents all applied phoneme sets.

3 Frequent Phonetic Changes in Russian

The information about phonetic changes in Russian speech may be found in [10] [11]. There are context-dependent and context-independent phonetic changes, elisions or epenthesis in Russian speech. The majority of these speech events are context-dependent and happen due to assimilation (e.g. eventual elision of /f/ in a phoneme sequence /f s/, when labialized /s/ is pronounced instead of /f s/ or consonant devoicing in prepausal position). An example of relatively context-independent phonetic change in Russian is a realization of /y/ instead of /a/ in a post-tonic syllables.

All phonetic changes are treated as context-independent within this work for a purpose of simplicity.

Vowel allophones behave in different manner depending on whether they are stressed or not. In this cases, vowel symbol contained indication of the sounds position regarding stress. Thus, '0' is used for a stressed vowel (e.g. /a0/ is a

stressed /a/), '1' – for an unstressed vowel in a pretonic syllable (e.g. /a1/ is a pre-stressed /a/), '4' – an unstressed one in a post-tonic syllable (e.g. /a4/ is a post-stressed /a/).

In terms of phonetic distance calculation the change is a phoneme substitution, the elision is a phoneme deletion and the epenthesis is a phoneme insertion.

3.1 Phoneme Sets Defining Substitutions

Proposed phoneme sets may intersect, i.e. a phoneme/allophone may be found in different sets. For example, allphones of /a/ appear in both sets of back vowels and front vowels, as they could be pronaunced in a front manner or back manner depending on speaker individual preferences and a context. Sets of phonemes and allophones that are highly probable to substitute each other:

- allophones of phoneme /a/: {a0, a1, a4}
- allophones of phoneme /e/: {e0, e1, e4}
- allophones of phoneme /i/: {i0, i1, i4}
- allophones of phoneme /o/: {o0, o1, o4}
- allophones of phoneme /u/: {u0, u1, u4}
- allophones of phoneme /y/: {y0, y1, y4}
- front unstressed vowels: {a1, a4, e4, e1, i4, i1, y1, y4}
- back unstressed vowels: {a1, a4, o1, o4, u1, u4}
- /j/ and allophones of /i/: {j, i1, i4}
- labial consonants and unstressed rounded vowels: {v, v', o1, o4, u1, u4}
- sibilants: {s, s', š, š':, z, z', ž}
- unvoiced stops and affricates: {t, t', t͡s, t͡š'}
- voiced stops and affricates: {d, d', d͡z, d͡ž'}

Note, that d͡z is used to denote voiced allophone of t͡s, and d͡ž' is used to denote voiced allophone of t͡š'.

There is also a number of phonetic processes in Russian speech which affect almost all Russian consonants. They are listed below with a couple of examples:

- consonant voicing, i.e. /t/ –>/d/ or /s/ –>/z/
- consonant devoicing, i.e. /d/ –>/t/ or /z/ –>/s/
- consonant palatalization, i.e. /t/ –>/t'/ or /s/ –>/s'/
- consonant depalatalization, i.e. /t'/ –>/t/ or /s'/ –>/s/
- affricate and stop spiratization, i.e. /t͡s/ –>/s/

3.2 Phoneme Elision

The elision of /j/ in intervocal position is so often in fluent speech that this type of phoneme deletion was the first phonetic change taken into account to improve transcription alignment. The eventual and context dependent elision of /h/, /h'/, /f/ and /f'/ prior to sibilants was not used otherwise there is need to consider a transformation of phoneme sequences and not single phonemes in transcription.

Table 1 shows an effect of taking into account phoneme sets and a possibility of /j/ elision.

Table 1. Alignment of rule-based and acoustic transcriptions using and not using phoneme classes for a word /b r a1 s a0 j i4 t/ pronounced with a lot of elisions as /b r s e0 t/

Alignment method	Alignment
Rule-based transcription	b r a1 s a0 j i4 t
Alignment not using phoneme sets	b r - s - - e0 t
Alignment using phoneme sets	b r - s e0 - - t

3.3 Phoneme Epenthesis

The only epenthesis taken into account is an epenthesis of a vowel inserted in between plosives and sonants acoustically similar to /e/ or /y/. That means a possible insertion of an element of set {e1, e4, y1, y4}. Table 1 shows an effect of taking into account possibility of phoneme epenthesis.

Table 2. Alignment of rule-based and acoustic transcriptions using and not using phoneme classes for a word /b r a1 s a0 j i4 t/ pronounced with epenthetic vowel /y1/ and elision of /a1/ as /b y1 r s a0 j i4 t/

Alignment method	Alignment
Rule-based transcription	b - r a1 s a0 j i4 t
Alignment not using phoneme sets	b - y1 r s a0 j i4 t
Alignment using phoneme sets	b y1 r - s a0 j i4 t

4 Experimental Results

4.1 Material

There are two Russian speech corpora annotated with several phonetic transcription tiers. The first one is COrpus of Russian Professonally REad Speech [4], which has manual acoustically-based and automatic rule-based text-to-phonemes phonetic transcriptions for 24 hours of speech data. The other one was created within INTAS 00-915 project [12], [13] and has manual acoustical and perceptional phonetic transcriptions and automatic rule-based text-to-phonemes one for 1 hour of speech data. The first one was selected as an experimental material as it contains much more data.

The experiments were carried out on the annotated part of the Corpus of Professionally Read Speech [4], which consists of recordings of read speech made from 8 professional speakers of Standard Russian. The annotated part of the corpus contains about 24 hours of speech with more than 1 million of speech sounds pronounced. There are two pronunciation tiers. The first one was produced automatically by grapheme-to-phoneme transcriber following orthoepic rules of Russian. The second one was produced manually by expect phoneticians during perceptual and acoustic analysis. These transcriptions were automatically aligned with each other and the alignement was manually corrected.

4.2 Results

Two transcripts were automatically aligned within the reported experiments. While the orthoepic transcription was used as a reference transcription, the manually-produced one was used as a hypothesis transcription. The existing alignment available with the corpus was used a "gold standard". Overall different ways of aligning these transcriptions with each other were evaluated.

The simplest way of taking into account acoustic nature of speech sounds is to divide them into two large phoneme sets: consonants and vowels. A more complex way is consider all the sets described in section 3.

Table 3 presents the comparison the alignment efficiency when the information about phonetic changes was not used and when it was used either in a simple or a complex way. Levenstein distance gives an efficiency of almost 99 %. But if we consider speech data with more than 20 hours of speech this leads us to more than 10 000 mistakes.

Vowels and consonants separation already brings an improvement and reduces the error rate by 29 %, see 2^{nd} row. The use of all the phonetic information reduces the erro rate by another 46 %.

Table 3. Comparison of overall alignment efficiency using different setups

Alignment method	Error rate (%)	Total number of errors
Levenstein dist.	1.11	11 899
Levenstein dist. + V \C separation	0.78	8 496
Levenstein dist. + all phonet. classes	0.27	2 905

5 Conclusions

The further improvement is to clarify phoneme sets. The next refinement step is to use information on phonetic changes according to their context-dependency. The further improvement would be to differenciate a cost for different phonetic events according to their probability.

The results of this work is to be used in the development of automatic segmentation of Russian speech into suprasegmental speech units for accurate alignment of automatically produced phonetic sequences along a speech signal considering that many speech sounds could be mispronounced, elised or inserted in continuous speech.

Acknowledgment. The author acknowledges Russian Scientific Foundation for a research grant 14-18-01352 "Automatic segmentation of speech signal into suprasegmental speech units".

References

1. Heeringa, W.J.: Measuring Dialect Pronunciation Differences Using Levenshtein Distance. PhD Thesis, Rijksuniv., Groningen (2004)
2. Valls, E., Wieling, M., Nerbonne, J.: Linguistic Advergence and Divergence in Northwestern Catalan: A Dialectometric Investigation of Dialect Leveling and Border Effects. LLC: Journal of Digital Scholarship in the Humanities 28(1), 119–146 (2013)
3. Álvarez, A., Arzelus, H., Ruiz, P.: Long Audio Alignment for Automatic Subtitling Using Different Phone-Relatedness Measures. In: Proc. of the 2014 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), pp. 6321–6325 (2014)
4. Skrelin, P., Volskaya, N., Kocharov, D., Evgrafova, K., Glotova, O., Evdokimova, V.: CORPRES - Corpus of Russian Professionally Read Speech. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS (LNAI), vol. 6231, pp. 392–399. Springer, Heidelberg (2010)
5. Bordel, G., Nieto, S., Penagarikano, M., Rodríguez-Fuentes, L.J., Varona, A.: A Simple and Efcient Method to Align Very Long Speech Signals to Acoustically Imperfect Transcriptions. In: 13th Annual Conference of the International Speech Communication Association (2012)
6. Elffers, B., Van Bael, C., Strik, H.: ADAPT: Algorithm for Dynamic Alignment of Phonetic Transcriptions. Internal report, Department of Language and Speech, Radboud University Nijmegen, the Netherlands. Electronically (2005), `http://lands.let.ru.nl/literature/elffers.2005.1.pdf`
7. Levenstein, V.: Binary codes capable of correcting deletions, insertions and reversals. Doklady Akademii Nauk SSSR 163, 845–848 (1965) (in Russ.)
8. Hirschberg, D.S.: A Linear Space Algorithm for Computing Maximal Common Subsequence. Communications of the ACM 18(6), 341–343 (1975)
9. Wieling, M., Nerbonne, E.M., Nerbonne, J.: Inducing a Measure of Phonetic Similarity from Pronunciation Variation. Journal of Phonetics 40, 307–314 (2012)
10. Bondarko, L.V.: Phonetics of contemporary Russian language. St. Petersburg (1988) (in Russ.)
11. Phonetics of spontaneous speech. Svetozarova N. D. (ed). Leningrad (1988) (in Russ.)
12. Bondarko, L.V., Volskaya, N.B., Tananiko, S.O., Vasilieva, L.A.: Phonetic Propeties of Russian Spontaneous Speech. In: 15th International Congress of Phonetic Studies (2003)
13. De Silva, V., Iivonen, A., Bondarko, L.V., Pols, L.C.W.: Common and Language Dependent Phonetic Differencies between Read and Spontaneous Speech in Russian, Finnish and Dutch. In: 15th International Congress of Phonetic Studies (2003)