Phoneme Alignment Using the Information on Phonological Processes in Continuous Speech

Daniil Kocharov
Department of Phonetics, Saint Petersburg State University,
7/9 Universitetskaya nab., St. Petersburg, 199034 Russia
kocharov@phonetics.pu.ru

Abstract

The current study focuses on optimization of Levenshtein algorithm for the purpose of computing the optimal alignment between two phoneme transcriptions of spoken utterance containing sequences of phonetic symbols. The alignment is computed with the help of a confusion matrix in which costs for phonetic symbol deletion, insertion and substitution are defined taking into account various phonological processes that occur in fluent speech, such as anticipatory assimilation, phone elision and epenthesis. The corpus containing about 30 hours of Russian read speech was used to evaluate the presented algorithms. The experimental results have shown significant reduction of misalignment rate in comparison with the baseline Levenshtein algorithm: the number of errors has been reduced from 1.1 % to 0.28 %.

Keywords: automatic phoneme alignment, phonological process, confusion matrices, Russian

1. Introduction

The goal of the research presented in this paper is to align effectively two sequences of phonemes (phonetic transcriptions) that describe the same speech signal. This task is important for sociolinguistic and dialectological research on how various people read or speak the same text, e.g. (Heeringa, 2004), (Valls et al., 2013). Another field of application is comparison and alignment of annotations produced by different human or machine transcribers for the same speech data, e.g. (Álvarez et al., 2014).

The current work has been done as a part of a research on inter-speaker variability. There has been a need in prediction of the pronunciation deviation of various Russian native speakers from Standard Russian. A correct comparison of individual pronunciations requires transcriptions to be perfectly aligned. The linguistic approach assumes that for a correct alignment of phoneme sequences one should consider the behavior of phonemes in continuous speech under different conditions.

Automatic aligners using various linguistic approaches have been made for many languages, including English (Álvarez et al., 2014), Spanish (Valls et al., 2013), Dutch (Elffers et al., 2005), Norwegian (Heeringa, 2004), Basque (Bordel et al., 2012). Such an aligner for Russian is presented in the paper. It is based on the usage of phoneme sets defined in such a way that a phoneme is more probable to substitute another phoneme from the same set than from another set. Besides substitutions, the approach suggests taking into account frequent phone insertions and deletions, whereas in previous works only substitution cost was estimated.

Section 2. presents briefly the information on phonological processes in Russian speech that have been considered while developing the aligner. Section 3. describes the modifications that are proposed for basic Levenshtein aligner. The evaluation procedure for estimating alignment efficiency is described in section 4.. The achieved results are shown in section 5..

2. Phonological Processes in Russian Speech

There are both context-dependent and context-independent phone alternations, elisions or epentheses in the Russian speech. The majority of these speech events are either context-dependent assimilation or changes due to position relative to word stress.

Frequent assimilation processes include regressive voicing/devoicing and palatalization/depalatalization. These processes occur both within a word and across word boundaries when word-final obstruents assimilate with initial obstruents. A number of consonants may be vocalized in intervocalic position, e.g. /j/ may be pronounced as /i/.

Position-dependent phonological processes include strong vowel reduction in unstressed position in continuous speech, e.g. /i/ is often pronounced instead of /a/ after palatalized consonants.

Voicing, devoicing, palatalization, depalatalization, sonorant vocalization, and unstressed vowel substitutions were considered as factors influencing phoneme changes in phonetic transcriptions.

There are a number of elision processes in continuous speech that were considered as relevant factors for the rate of deletions in phonetic transcriptions. These processes include elision of /j/ in intervocalic position and possible elision of unstressed vowels, especially in post-stressed syllables. Data analysis showed that /j/ is elided in more than 50 % of cases, and unstressed vowels are elided in about 4 % of cases.

The majority of insertions are epenthetic vowels within consonant clusters, especially in position before sonorants, /v/, and /v^j/ (Evgrafova, 2009). Around 12 % of consonant junctions in Russian read speech contain epenthetic vowels (Skrelin et al., 2010).

The phonological processes described above were considered while developing the efficient algorithm of automatic transcription alignment. All other phonological processes were not taken into account due to their low frequency in Russian.

Table 1: Alignment of reference rule-based transcription for a word /braˈsajit/ and transcription describing pronunciation with large number of elisions as /ˈbrsʲet/. Comparison of simple Levenshtein algorithm and correct alignment.

Alignment method	Alignment							
Reference transcription	b	r	a	s	ˈa	j	i	t
Levenshtein alg.	b	r	–	s	–	–	ˈe	t
Correct alignment	b	r	–	sʲ	ˈe	–	–	t

Table 2: Alignment of reference rule-based transcription for a prepositional phrase /ˈat/ /tʲiˈbʲa/ and transcription describing pronunciation /atʲtʲiˈbʲa/. Comparison of simple Levenshtein algorithm and correct alignment.

Alignment method	Alignment					
Reference transcription	ˈa	t	tʲ	i	bʲ	ˈa
Levenshtein alg.	a	–	tʲ	tʲ	bʲ	ˈa
Correct alignment	a	tʲ	tʲ	–	bʲ	ˈa

The detailed information about phonological processes in Russian speech may be found in works of (Avanesov, 1984), (Bondarko, 1998), and (Svetozarova, 1988).

3. Transcription Alignment

The state-of-the-art solution for alignment of sequences of strings or symbols is a dynamic programming approach. The most common way is to calculate Levenshtein distance between the sequences (Levenshtein, 1965). The Levenshtein distance is measured as a minimum number of edit operations necessary for transformation of one sequence into the other. These edit operations are substitutions, insertions and deletions. Each operation has its cost, usually all the operations cost '1'. A match operation with a cost of '0' is introduced for calculation simplicity.

When comparing phoneme sequences, this approach is not the most efficient because it does not account for similarity of phonemes or frequent phonological processes in speech, such as phone elision, epenthesis, and assimilation. The examples of wrong alignments produced by simple Levenshtein algorithm are presented in tables 1 and 2, where '–' denotes phone deletion. The illustrated alignment errors would lead to unnatural phonological implications. The first example illustrates /ˈe/ pronounced instead of junction /ˈaji/. The wrong alignment would show that a stressed vowel is more likely to be elided than an unstressed one whereas more correct interpretation would be that the stressed vowel changes its quality in this case and there is an elision of post-stress syllable /ji/. The second wrong decision would be that a vowel is likely to be pronounced as a voiceless stop whereas an unstressed vowel is likely to be elided and /t/ is assimilated to the following /tʲ/.

A preliminary experiment showed that basic Levenshtein algorithm produces 1.1 $\%$ of errors, that is almost 12,000 sounds for our 30 hours of experimental speech data. Though it is a very small error rate from the statistical point of view, it may strongly influence linguistic research. It biases probability of infrequent phonological processes and produces unnatural pronunciations, e.g. pronunciation of vowel /i/ as voiceless stop /t/ in table 1.

There have been efforts to measure phonetic difference more precisely assuming that a cost of substitution of one phoneme by another should depend on the phonetic distance between these phonemes. Elffers with colleagues proposed to consider a phoneme as a vector of articulatory features. Then the phonetic distance between two phonemes is a sum of absolute differences between feature values of the phonemes (Elffers et al., 2005). Wieling suggested to estimate the phonetic distance by means of pointwise mutual information, the number of times phonemes corresponded to each other in aligned transcriptions (Wieling et al., 2012). Álvarez and his colleagues tested three different ways of constructing confusion matrix: using confusion matrix of ASR phone-decoder, using phonemes perceptual similarity and using their phonological similarity (Álvarez et al., 2014). The best results were obtained with the help of ASR phone-decoder.

The approach proposed in this paper aimed at improving phonetic transcription alignment is different and is based on the idea of defining sets of phonemes that are highly probable to substitute each other because of certain phonological processes in continuous speech. All phone substitutions, deletions and insertions are treated as context-independent within this research for calculation simplicity. The optimization is done by means of confusion matrices that include information not only on substitutions but on deletions and insertions as well. The edit operation cost for phonemes within the same set should be lower than edit operation cost for phonemes from different sets.

The first modification was to make Levenshtein algorithm VC-sensitive, i.e. reduce costs for substitution of two consonants and substitution of two vowels. The motivation for this step is that vowels are rarely substituted by consonants and vice versa. The speech corpora described in section 4. shows that it happens in no more than 0.02 $\%$ of cases.

The second optimization was to separate consonants into obstruents and sonorants as they also rarely substitute each other.

Phoneme /j/ does not behave as other sonorants because of its frequent elisions and vocalizations. It is elided in more than 50 $\%$ of the cases and vocalized in 8.4 $\%$ of cases. Compare it with 0.1 $\%$ of vocalization rate for other sonorants. The deletion cost for /j/ was reduced significantly and the cost of substitution by /i/, ɨ, and /e/ was reduced by the same value as for all the other phoneme sets.

There were also two setups for taking into account voicing/devoicing and palatalization/depalatalization, because they are essential for Russian speech. The cost of substitutions within the pairs of obstruents differing in these features was reduced. See, for example, the following voiced-unvoiced pairs: $\{p, b\}$, $\{p^{j}, b^{j}\}$, $\{f, v\}$, $\{f^{j}, v^{j}\}$, $\{t, d\}$, $\{t^{j}, d^{j}\}$, $\{s, z\}$, $\{s^{j}, z^{j}\}$, $\{k, g\}$, $\{k^{j}, g^{j}\}$, { ʂ, ʐ}.

The last phonological setup was done to take into account unstressed vowel reduction and frequent vowel epenthesis. The cost of deletion and insertion of unstressed vowels was reduced.

The reduction of operation cost is equal for all the phoneme sets. For example, if the reduction of operation cost equals 0.1, then substitution of phonemes within a given set costs 0.9, and the substitution of phonemes across sets costs 1. If a pair of phonemes is found in two phoneme sets, then the substitution cost for these phonemes is reduced twice.

The last experiment was dedicated to data-driven approach to building confusion matrix. The costs were not predefined according to theoretical knowledge, but were trained from the statistics of phone substitution, deletion and insertion within the speech data. The costs were estimated with following formula: $cost = 1 - P(edit\ operation|phoneme)$. Thus the sum of all edit operations except match equals 1. The cost of match for any phoneme equals 0 regardless of how often a phoneme changes its quality in speech.

The experimental results are presented in section 5.

4. Evaluation

All the experiments were performed using CORPRES—Corpus of Russian Professionally Read Speech (Skrelin et al., 2010), which consists of recordings of read speech produced by eight speakers of Standard Russian. The annotated part of the corpus contains about 30 hours of speech with more than 1.1 mln speech sounds. There are two tiers with phonetic transcription. The first one was produced automatically by grapheme-to-phoneme transcriber following orthoepic rules of the Russian language. The second one was produced manually by expect phoneticians based on perceptual and acoustic analysis. These transcriptions were automatically aligned, and the alignment was manually corrected. The mismatch between two phonetic transcriptions is about 15.4 %. Phrase length in phones varies from 1 to 568 with a mean value of 20 phones per phrase.

CORPRES contains perfectly aligned phonetic tiers. This alignment was used as the 'gold standard'. The orthoepic transcription was used as the reference transcription, the manual transcription of a real speaker's pronunciation was used as a hypothesis transcription. The evaluation procedure is to compare alignments produced by proposed algorithms and the 'gold standard' alignment.

Alignments are produced phrase by phrase. For a given phrase alignment, each pair of aligned phoneme symbols is converted into a single token by joining the symbols with symbol ':' as a delimiter. Thus the alignment of two sequences is converted into a sequence of these tokens. Such sequence is compared with the one produced from 'gold standard' by means of standard Levenshtein distance. Table 3 presents an example of such alignment. As the length of both alignments is the same, only substitutions and matches are possible, and there are no deletions or insertions. Thus the Levenshtein distance is equal to the number of mismatches (substitutions) between these two sequences. Finally, Levenshtein distances for all the phrases are summed up, which provides a number of errors produced by a given alignment algorithm.

Table 3: Alignment of two alignments for a word /ˈks^{j}en^{j}ija/ pronounced as /ˈks^{j}en^{j}ee/. A comparison of simple Levenshtein algorithm and correct alignment.

Correct standard	k:k	s^j:s^j	ˈe:ˈe	n^j:n^j	**i:e**	**j:–**	a:e
Hypothesis alignment	k:k	s^j:s^j	ˈe:ˈe	n^j:n^j	**i:–**	**j:e**	a:e

Table 4: Comparison of overall alignment efficiency showed by algorithms taking into account different phonological processes.

Alignment method	Error rate (%)	Total number of errors
Standard Levenshtein algorithm	1.11	11 899
Data-driven confusion matrix	0.79	8 551
V/C separation	0.70	7 595
V/S/C separation	0.34	3 714
V/C + /j/	0.30	3 246
V/S/C separation + /j/	**0.28**	**3 022**
V/S/C separation + /j/ + consonant voicing	0.28	3 039
V/S/C separation + /j/ + consonant palatalization	0.28	3 045
V/S/C separation + /j/ + unstressed vowels sub. and ins.	0.28	3 036

5. Experimental Results and Discussion

Table 4 presents the efficiency comparison for different alignment algorithms. The efficiency is given in terms of error rate and absolute number of errors produced by the algorithms.

The data-driven estimation of costs for the confusion matrix shows significant improvement against the baseline standard Levenshtein algorithm. There is 28 % error rate reduction.

'V/C separation' refers to the approach when the phonemes are separated into vowels and consonants. This simple optimization of Levenshtein algorithm shows slightly better results than data-driven approach, although the statistical confidence of the difference between the efficiency of these two algorithms is questionable.

'V/S/C separation' refers to the approach when the phonemes are separated into vowels, sonorants, and consonants. This separation shows significant reduction of error rate against V/C separation, which gives an error rate drop for another 40 %.

'/j/' refers to accounting for /j/ elisions and vocalizations. This modification shows significant increase in efficiency against both 'V/C' and 'V/S/C' separation. The relative reduction of error rate against 'V/S/C' is lower than for 'V/C'. This is because sonorant phoneme set in 'V/S/C' includes /j/ as its member, so adding /j/ to this setup increments efficiency less than in case of 'V/C'.

'Consonant voicing' refers to the setup where substitution costs are reduced for consonants that differ only by voicing.

Table 5: Comparison of overall alignment efficiency using different edit operation costs. The setup is: V/S/C separation + /j/ elision.

Alignment method	Error rate (%)
Subs cost = 0.99	0.282
Subs cost = 0.9	0.282
Subs cost = 0.8	0.282
Subs cost = 0.7	0.280
Subs cost = 0.5	0.279
Subs cost = 0.3	0.533
Subs cost = 0.1	0.540

'Consonant palatalization' refers to the setup where substitution costs are reduced for consonants that differ only by palatalization. 'Non-stressed vowels sub. and ins.' refers to the setup where insertion and deletion costs for unstressed vowels are reduced. Accounting for these phonological processes did not add any improvement to so far the best algorithm 'V/S/C separation + /j/'. The reason of this unpredicted result could be that the presented phonological processes, such as palatalization/depalatalization or voicing/devoicing, are consistent and produced by all the speakers. The additional experiments showed that in less than 2 % of cases a speaker does not pronounce the assimilated consonant instead of the original one, or vice versa pronounces it in the context where there is no assimilation process. However, one would expect the increase of such mispronunciations in speech of non-native Russian speakers or in conversational speech where the mismatch rate between orthoepic transcription and real pronunciation is greater.

Thus the 'V/S/C separation + /j/' setup is considered to be the best achieved solution. It showed an error rate reduction of almost 75 % in comparison with the baseline Levenshtein algorithm. This algorithm was further used for experiments aiming to detect whether the value of cost reduction is an important factor influencing the overall efficiency.

It should be noted that in majority of cases each misalignment error leads to two phoneme misalignments, see table 3. One wrong decision on deleting /i/ instead of /j/ has led to two misalignments: 'i:–' instead of 'i:e' and 'j:e' instead of 'j:–'.

The exact value of reduction is a matter of question. Thus a number of experiments were carried out to define the reduction cost that gives the best result. The following values were tested: 0.01, 0.1, 0.2, 0.3, 0.5, 0.7, 0.9. Table 5 presents the efficiency for different alignment algorithms when using different edit operation costs. The table shows that the error rate difference is lower than 0.003 % for all substitution costs above 0.5. The costs below 0.5 work much worse. The value of cost reduction is not important if it is above 0.5, as the efficiency is almost the same for cost values 0.5 and 0.99.

6. Conclusions

The experimental results showed that the simplest way to increase the efficiency of alignment of phonetic transcriptions is to separate phonemes into three classes: vowels, sonorants and obstruents, and to 'forbid' substitutions of elements between these classes. Even this approach shows much better results than a data-driven estimation of costs for edit operations. Further optimization is possible when the information on the most crucial phonological processes is taken into account. Only the elision and vocalization of /j/ dropped the error rate. The others did not add any efficiency but on the other hand they did not drop it either. The achieved efficiency of 0.28 % error rate is almost four times less than a baseline 1.1 % error rate of Levenshtein algorithm.

The experiments were performed using read speech, where the real pronunciation does not differ perceptually from the standard Russian pronunciation and where speakers do not differ much from each other. This may be the reason why accounting for assimilation processes did not play a role. The conversational speech is much more diverse. One would expect the assimilation and vowel reduction to be much more crucial for that kind of speech. The repetition of these experiments on conversation speech might be a goal of some further research.

The confusion matrices and Python code used for the experiments are freely available at the web-site of ISCA Special Interest Group on Russian Speech Analysis within the page dedicated to Resources (http://www.forma.spbu.ru/?q=en).

7. Acknowledgment

The research is supported by the Russian Science Foundation (reserch grant # 14-18-01352 "Automatic segmentation of speech signal into suprasegmental speech units").

8. Bibliographical References

Álvarez, A., Arzelus, H., and Ruiz, P. (2014). Alignment for automatic subtitling using different phone-relatedness measures. In *2014 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, pages 6321–6325, Florence.

Avanesov, R. I. (1984). *Russian standard pronunciation*. Prosveschenije, Moscow. [Russkoe literaturnoe proiznoshenie] (in Russian).

Bondarko, L. V. (1998). *Phonetics of contemporary Russian language*. St. Petersburg. [Fonetika russkogo sovremennogo jazyka] (in Russian).

Bordel, G., Nieto, S., Penagarikano, M., Rodríguez-Fuentes, L. J., and Varona, A. (2012). A simple and efficient method to align very long speech signals to acoustically imperfect transcriptions. In *13th Annual Conference of the International Speech Communication Association*.

Elffers, B., Van Bael, C., and Strik, H. (2005). Adapt: Algorithm for dynamic alignment of phonetic transcriptions. Technical report, Department of Language and Speech, Radboud University Nijmegen, the Netherlands.

Evgrafova, K. (2009). The phonetic characteristics of vowel epenthesis in russian consonant clusters. In *13th International Conference on Speech and Computer SPECOM 2009*, pages 419–422.

Heeringa, W. J. (2004). *Measuring Dialect Pronunciation Differences Using Levenshtein Distance*. Ph.D. thesis, Rijksuniversity, Groningen.

Levenshtein, V. (1965). Binary codes capable of correcting deletions, insertions and reversals. In *Doklady Akademii Nauk SSSR*, volume 163, pages 845–848. (in Russian).

Skrelin, P., Volskaya, N., Kocharov, D., Evgrafova, K., Glotova, O., and Evdokimova, V. (2010). Corpres – corpus of russian professionally read speech. In P. Sojka, et al., editors, *13th International Conference Text, Speech and Dialogue, TSD 2010*, volume 6231 of *LNCS*, pages 392–399. Springer.

N. D. Svetozarova, editor. (1988). *Phonetics of spontaneous speech*. St. Petersburg. [Fonetika spontannoj rechi] (in Russian).

Valls, E., Wieling, M., and Nerbonne, J. (2013). 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.

Wieling, M., Nerbonne, E. M., and Nerbonne, J. (2012). Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics*, 40:307–314.