NGSLT Speech and Speaker Recognition course (Spring 2007)

Phonetic Modeling in ASR (Russian speech) - the impact on performance

Smirnov Valentin

Abstract

In this term paper for NGSLT course "Speech and Speaker Recognition" we will describe the method for pronunciation modeling, which makes part of automatic speech recognition (ASR) technology for Russian, developed by R'n'D department at Vocative, ltd, where the author works. The method comprises both canonical transcription generation and pronunciation variation modeling.

The key idea of the current work is to automate the process of lexicon generation for a given recognition task and to inspect the impact of modeling pronunciation variation on ASR performance. Specific issues concerning pronunciation variation in Russian are discussed and the results on several recognition tasks are presented.

1. Introduction

Modern ASR systems treat speech as a sequence of distinct phonetic units – phonemes. Consequently, lexicon is used to set up a correspondence between words in their orthographic form and their phonetic transcriptions. These lexicons might be prepared manually by an expert, or generated automatically. The latter option is the issue for the present discussion.

The role of a lexicon, which contains canonical transcriptions for the words, is clear; however there has been an emerging interest to model pronunciation variation also, since it has come to mind that a word is never produced the same way and – as a consequence – the performance of ASR system, which uses only canonical transcriptions, might degrade.

In the following paragraphs we say a couple of words about the ASR engine used in the experiments, and outline the reasons for the problem in question. Then we describe the method for pronunciation modeling itself, give some practical examples and, in the end, make suggestions about future research.

2. ASR system overview

Vocative Russian ASR Engine is developed as a commercial product, which is aimed at recognizing Russian speech in telecommunication applications (voice information systems, auto-attendants) and multimedia applications (voice control). The system is based on Hidden Markov Models (HMM) with triphones (context-dependent monophones) as basic phonetic units. The acoustic signal is coded with 13 Mel-Frequency Cepstral Coefficients (MFCC) and their first and second order derivatives. To train acoustic models about 50 hours of prerecorded samples are used (more than 100 speakers as a whole). This corpus was recorded in laboratory conditions, and then wav-files were down-sampled and passed through a telephone filter to obtain telephone-like speech quality. (In near future we also plan to use another training corpus (part of SpeechDat for Russian), which contains telephone speech from several hundred speakers, recorded through PSTN.)

At present fully-continuous density HMMs are used, which are trained in a flat-start fashion with three successive reestimations. Data-driven clustering is used to group triphones into broader classes. Viterbi search is used at the recognition stage. For more detailed information on HMM-based recognizers refer to [1, 2, 3].

3. Problem formulation

There are two main tasks in ASR, which demand pronunciation modeling and pronunciation variation modeling, in particular. First task is transcribing large speech corpora, i.e. converting word-level descriptions of the waveforms into phone-level transcriptions. The second task consists in preparing lexicon automatically for (any) ASR tasks. This term paper concerns the latter task.

The need to supply orthographic form of words with phonetic transcription is obvious and inevitable, since phonemes (triphones in the present case) are basic units for statistical acoustic modeling. The desire to model variations arises from another obvious observation that the words are never pronounced in the same way. There is a vast variety of sources for variation, e.g. Strik in [4] mentions speaking style, degree of formality, interlocutor, environment, speech disability, accent or dialect, socioeconomic factors, anatomical differences, and emotional status. Russian speech is not an exclusion from this universal law. Moreover, being a highly inflected language, Russian presents a particular interest for both canonical pronunciation generation and for variation modeling. When confronted with a task to automatically produce transcriptions for Russian words one should keep in mind rich paradigms for Russian nouns, adjectives and verbs. Significant phonetic differences between the forms are characteristic for these parts of speech (e.g. six cases for nouns, six forms of verb conjugation in present tense, etc.). Lexical stress in Russian being not anchored to a specific syllable, canonical transcription generation is another problem to come across. Moreover, rich variety of forms leads to a large number of homographs; to cope with them one should use automatic POS (part of speech) parser. Otherwise generated transcription will be inadequate, since homographs usually have stress on different syllables. For profound information on Russian pronunciation refer to [5].

When dealing with pronunciation variation, phoneticians usually make distinction between within-

word and cross-word variation [6,7]. While the sources for within-word variation are very diverse (they were cited in the previous paragraph), cross-word variation is usually limited to simplified pronunciations of so-called multi-words for those word sequences which are frequently used together in a language, their phonetic structure being changed as a consequence (e.g., "gonna" instead of "going to" in English) [7]. This kind of variation is of great importance for Russian speech ([tos'] instead of [to jest'] for "that is" conjunction, [mobyt'] instead of [možyt byt'] for "may be"); however we have not yet encountered this type of words in our applications. Once we do, the best decision would be to make a list of such multi-words and assign transcriptions through calling this list.

Another possible source for cross-word variation one may think of is contextual change of phonemes on the word border. E.g., in Russian when a word finishing with an unvoiced consonant is followed by a voiced one the unvoiced final gets voiced ([ko**d** zamka] instead of [ko**t** zamka] – "the code of the locker"). Another example is [i] at the beginning of a word changing to [y], if a previous word finishes with non-palatalized consonant ([b@ris **y**gnat@f] instead of [b@ris **i**gnat@f]). This kind of variation makes part of our automatic transcriber used for transcribing training corpus.

4. Pronunciation modeling method

The primary aim of the method is to automatically generate pronunciation lexicons for speech recognition grammars. It should be noted that our method is designed to generate both (and primarily) canonical transcriptions for an arbitrary set of words and their pronunciation variants through applying knowledge-based rules. At the moment we add the variants manually to the lexicon, automatic generation being planned in near future.

Below we list a number of pronunciation variation rules used in our ASR system. It is crucial to note that all of them are knowledge-based, i.e. they are derived from a prior knowledge about Russian phonetics and morphology. Data-driven rules, i.e. those generated through the direct analysis of actual acoustic data, are left for the future. Refer to [4] for excessive bibliography on both data-driven and knowledge-based methods.

It is well known that knowledge-based rules, being "laboratory" in origin, may happen to be inadequate when confronted with real-life data. However this was our intent to check this critical assumption on our test material. Moreover, during past decade Russian phonetics has undergone a general shift from laboratory speech to fully spontaneous [8,9], and the rules we are aware of at the moment are based also on research concerning spontaneous speech, i.e. are supposed to be quite close to real-life.

The rules we use are divided into two main groups. The first contains substitution, deletion and insertion rules, which apply to (automatically generated) canonical phonetic transcriptions. Some examples of such rules are listed in table 1.

The second group of rules makes use of both morphological and orthographical level of linguistic representation. Hence, this is not correction to canonical transcriptions (phone-to-phone rules), but a separate set of letter-to-phone rules. Some examples are presented in Table 2.

Table 1. Substitution, deletion and insertion rules

Rule	Example
[@] ("schwa"), followed by a consonant, is deleted in the unstressed position after stressed syllable;	[kol@k@l] -> [kol@kl] "колокол" (bell)
[f] is deleted from consonant sequence [fs] + (any) unvoiced consonant;	[fs'akij] -> [s'ak'ij] "всякий" (any)
affricates [c] and [t͡ʃ'] are substituted by fricatives [s] and [ʃ'] respectively (sign ' denotes that a consonant is palatalized);	[pr@kt'i t͡ʃ'isk'i] -> [pr@kt'i ʃ'isk'i] "практически" (practically)
sonorant [j] is deleted before unstressed vowel at the beginning of words;	[jida] -> [ida] "еда" (food)
noise stops (e.g. [p], [t], [p'], [t']) are deleted in the final position after vowels due to implosive pronunciation (i.e. without burst following articulators closure).	[lop] -> [lo] "лоб" (forehead) [m'es't'] -> [mes'] "месть" (revenge)

Table 2. Letter-to-phone rules, generating pronunciation variations

Rule	Example
[@j@], [uju] in unstressed inflections of adjectives "–ая", "–ую" are changed to [@e] and [u] respectively	[krasn@j@] -> [krasn@e] "красная" (red)
[@v@], [yv@], [iv@] in unstressed noun and adjective inflections "–ого", "–его" is changed to [@@], or [@:], [y@], [i@]	[na ʃyv@] -> [na ʃy@] "нашего" (ours)
[@t] in verb inflections "–ат" is changed to [yt]	[usly ʃ@t] -> [usly ʃyt] "услышат" (will hear)

For very frequent words we also added another set of rules, which generate simplified pronunciation, which is common to informal spontaneous speech. These include [d'] and [v] deletion in intervocalic position, [s't'] changing to [s'], etc.

Below we outline main steps of our method of automatic lexicon generation for a given recognition task. The scheme is shown at Figure 1.

1) The list of words is extracted from a given recognition grammar;
2) Hand-made lexicon is accessed, which contains transcriptions for words created by an expert;
3) If the word is not in this lexicon, automatic transcriber is called which makes use of a dictionary of the Russian language, containing several hundreds of thousands

words with morphological information, and a part-of-speech (POS) detection procedure. This dictionary also contains proper names and non-native words. As a result of this stage the stress is assigned to the right syllable of the word;

4) Canonical transcription is generated, context-dependent letter-to-phone rules being applied;
5) If desired by a researcher, at the previous stage pronunciation variation might be executed through applying a separate set of letter-to-phone rules.
6) Knowledge-based rules are applied to transcription generated at the previous step (either canonical or morphology-based variations) in order to generate knowledge-based pronunciation variants.

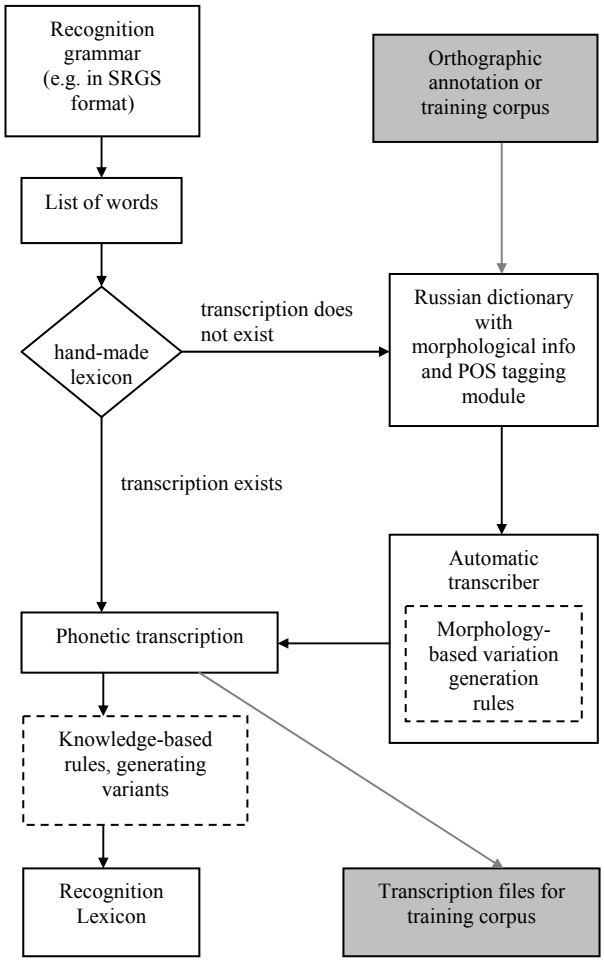

Figure 1. Method for automatic generation of recognition lexicon. Dashed blocks denote steps performed manually. Grey blocks denote applications planned in near future

One of the most important parts of automatic transcriber is morphological description of words, given in the dictionary. The structure of the dictionary derives from the ideas shared in [10]. Each word in a dictionary is assigned its POS, grammatical markers (gender and inclination type for nouns, conjugation type for verbs, etc.). The idea to use morphological dictionary for Russian ASR applications is also expressed in [11, 12].

5. Results

We tested the performance of our recognizer on a test corpus, containing "real-life" examples of spoken prompts to various imaginary telecom applications. The speech signal parameters are: PCM, 16 bit, 8000 Hz, telephone-filtered. The results referred below should be treated as preliminary since the development of our ASR engine continues at the moment (the acoustic model is being tuned, additional spectral features are added, etc.). However even with this prototype we will try to show the impact of pronunciation modeling on Russian ASR performance.

Several studies were conducted in order to measure the extent to which pronunciation modeling might be of help for improving speech recognizer performance. Thus, McAllaster in [13] reports that adding pronunciation variants to a lexicon led to a significant improvement (5% of WER compared to 40% without any variants) Wester in [6] reports but a slight improvement: 0,68% when modeling within-word pronunciation variants and 1,12% for cross-word. "Cheating" experiments, described in [14], show that when the lexicon contains "real" transcriptions (extracted from the test corpus through preliminary phone-recognition of this corpus) the result can improve significantly – WER drops 43% (relative). Thus we see: the degree of correspondence between acoustic models and transcriptions being artificially set higher, the improvement gained is significant.

Table 1 contains the results of several trials of ASR engine. Three very simple recognition tasks are tested, each comprising rather small lexicons ("Movie-names" – 40 words, "Subway stations" – 70 words, "Dates of birth – 30 words"). Average grammar perplexity is 10. All recognition parameters in terms of HMM type, feature vector and search algorithm are kept the same. The only difference is the kind of lexicon used in each trial. "Baseline" stands for a lexicon comprising canonical transcriptions only (automatically generated). In the second set of experiments this lexicon was enlarged with knowledge-based pronunciation variants. The last column represents the results obtained through adding to the lexicon an excessive number of variants , rare deletion and rare substitution rules being considered, thus resulting in doubling the number of pronunciation variants for the words (on average).

Our results show that adding numerous pronunciations is sometimes prohibitive, since the confusability increases and, consequently, recognition performance decreases. However, relevant ratio being found, the results do improve significantly. We should also note that canonical transcriptions tend to give worse results for all three tasks, our test corpora being produced by untrained, non-professional speakers in a spontaneous fashion. Carefully adding variants leads to improvement and we consider this as encouragement for continuing the development of our method, namely to automate rules described earlier in this paper.

Table 1. Various kinds of pronunciation modeling. The impact on ASR performance

ASR task	Baseline (automatic)	Knowledge -based variants added	Excessive number of variants added
movie-names	90.56	93.70	92.25
subway stations	89.90	92.93	91.41
dates of birth	89.02	91.46	87.80

6. Conclusion

We have pointed out that pronunciation modeling in ASR can be used for two purposes, mainly for improving recognition results incorporating pronunciation variants in a lexicon and for generating transcriptions for the training corpus. This term paper describes the first task, i.e. building a system which automatically generates lexicon for a given recognition task calling a dictionary with built-in POS tagging, with knowledge-based rules used for producing pronunciation variants. Our future work will focus on using our method to enhance training, since to date we have used only canonical transcriptions generated in semi-automatic fashion (automatic transcriber with subsequent corrections made by an expert). Pronunciation modeling at this step will be of help to resolve discrepancies between transcriptions and the actual acoustic data, which in fact leads to less adequate acoustic model estimation. Applying forced Viterbi alignment to training corpus supplied with more exact transcriptions may lead to better estimates of statistical parameters of the acoustic model.

Another possible direction for future research is to try data-driven approach to pronunciation modeling, when pronunciation variation rules are extracted automatically from a large annotated speech corpus [cf. 15].

References

[1] X. Huang, A. Acero, H.-W. Hon (2001). Spoken language processing: a guide to theory, algorithm, and system development, Prentice Hall

[2] L. Rabiner and B.-H. Juang (1993). Fundamentals of Speech Recognition. Prentice Hall.

[3] S. Young (2001). Statistical modeling in continuous speech recognition (CSR). In UAI '01: Proc. of the 17th International Conference on Uncertainty in Artificial Intelligence, Seattle, WA.

[4] H. Strik (2001). Pronunciation adaptation at the lexical level. In Proc. of the ITRW Adaptation Methods for Speech Recognition, Sophia-Antipolis, pp. 123–130.

[5] L. Bondarko, L. Verbickaya, M. Gordina, L. Zinder, V. Kasevich (1975). Stili proiznosheniya i tipy proizneseneija. (Speaking styles and pronunciation types). In Proc. of 5th USSR conference on psycholinguistics and communication theory. Moscow, pp. 15-17 (in Russian).

[6] M. Wester (2002), Pronunciation Variation Modeling for Dutch Automatic Speech Recognition, thesis, Wageningen.

[7] M. Wester, J. M. Kessens, and H. Strik (1998). "Improving the Performance of a Dutch CSR by Modeling Pronunciation Variation," In Proc. of the Workshop Modeling Pronunciation Variation for Automatic Speech Recognition. Kerkrade, pp. 145-150

[8] Bondarko L.V., Iivonen A., Pols Lous C.W., de Silva Viola (2003). Common and Language Dependent Phonetic Differences Between Read and Spontaneous Speech in Russian, Finnish and Dutch. In Proc. of the 15th ICPhS. Barcelona, p. 2977.

[9] Bondarko L.V., Volskaya N.B., Tananaiko S.O., Vasilieva L.A. (2003). Phonetic Properties of Russian Spontaneous Speech. In Proc. of the 15th ICPhS. Barcelona, p. 2973.

[10] Zaliznyak, A. A. (1977). Grammaticheskii slovar russkogo yazyka. (The Grammatical Dictionary of Russian). – Moscow (in Russian).

[11] L. Bondarko, V. Velichko, N. Zagoruyko (1982). Slovoobrazovatelnyj slovar I ego ispolzovanie dlja avtomaticheskogo raspoznavanija rechi (Morpholgical dictionary and its deployment for automatic speech recogmition). In Proc. ARSO (Automatic recognition of speech signals), Kiev, pp. 442-445 (in Russian)

[12] A. Asinovskij, S. Bogdanov, L. Bondarko (1978) Ob ispol'zovanii morfologicheskoj informacii v celjah raspoznavanija nepreryvnogo rechevogo potoka. (Morphological information used for the recognition of continuous speech). In Proc. ARSO (Automatic recognition of speech signals), Tbilisi, pp. 70-71 (in Russian).

[13] D. McAllaster, L. Gillick, F. Scattone, and M. Newman (1998). Fabricating conversational speech data with acoustic models: a program to examine model-data mismatch. In Proc. of ICSLP '98, Sydney, pp. 1847–1850.

[14] M. Saraçlar and S. Khudanpur (2000). Pronunciation ambiguity vs. pronunciation variability in speech recognition. In Proc. ICASSP '00, Istanbul, pp. 1679–1682.

[15] J. E. Fosler-Lussier (1999), Dynamic Pronunciation Models for Automatic Speech Recognition. PhD thesis, Berkeley, California