IMPROVING AUTOMATIC SPEECH RECOGNITION ACCURACY BY MEANS OF PRONUNCIATION VARIATION MODELING

Vladimir Chuchupal, Anton Korenchikov

Abstract: *We explore the properties of the pronunciation variation (PV) models as an approach for an automatic speech recognition accuracy improvement. The PV model is formulated as well as the methods to find out PV parameters and include the model into the search procedures. We show that utilizing of the PV models could substantionally increase the accuracy of automatic recognition of natural speech.*

Keywords: *automatic speech recognition, acoustic modeling, speech pronunciation variation modeling, explicit models, hidden markov models.*

ACM Classification Keywords*: I.2.7 Natural Language Processing - Speech recognition and synthesis*

Introduction

The pronunciation of a word in a speech recognition system (ASR) usially is determined by its pronunciation or phoneme transcription. As a rule most of the words has a single pronunciation transcription, namely the basic or canonical one.

In spontaneous speech a pronunciation may substantionally differs from the canonical and this is one of the most important sources of the errors of speech recognizers.

There are currently two approaches to pronunciation variation (PV) modeling for ASR [Wester, 2003; Fosler, 1999]. The explicit modeling describes all probable pronunciation variations in terms of explicit changes of the basic word transcriptions. In the other words, in explicit modeling the given word pronunciation could be defined as a set of the most probable word transcriptions. The implicit modeling [Saraclar, 2004] describes the variations in pronunciation by means of changes in the structure of the allophone hidden Markov models of the basic transcription.

The both approaches do not eliminate the need to use the basic transcriptions.

The correct implementation of pronunciation variation models may have a great impact on the accuracy of ASR. Such a conclusion is follows from the heuristic analysis of the errors done by ASR as well as the oracle-style experiments. It had been showed that using of the adequate phonemic transcriptions reduce word error rate (WER) as much as nearly a twice [Saraclar at al., 2000].

At the same time the reported improvements in WER obtained with the use of pronunciation variation models still are far from the expected ones. For example on Dutch corpora VIOS the WER decrease obtained was about 0.8% (from initial level of 10.7% to 9.9%) at the significant number of the pronunciation variants: around 4.9 per a vocabulary word [Wester, 2003]. On Switchboard corpora the implementation of the implicit pronunciation models led to decrease in WER at 1.7% (from 39.4% to 37.7%) [Saraclar, 2004]. On NIST–2000 Hub-5 data the use of pronunciaton variation models improved WER at 2.2%: from 54.6% down to 52.4% [Zheng, 2003].

In this study we implemented the pronunciation variaton model in existing Russian ASR. We follow the explicit approach to pronunciation variation modeling in that all changes in pronunciation could be adequately described in terms of deletions, substitutions and insertions of the phonemes.

For implementation of this approach we have to address the following issues:

- Define a pronunciation variation models,
- Find the most probable phone transcriptions for the words,
- Find a method for estimating the parameters of the pronunciation model,
- Find a use of implementation of pronunciation models in the search procedure.

Pronunciation variation model

$$W = \underset{W}{arg\,max}\, P(W|X) = \underset{W}{arg\,max}\, \frac{P(X|W)P(W)}{P(X)}. \tag{1}$$

Note that the important distinction between words and phone transcriptions is that a word defines or relates to the meaning whereas a transcription defines the acoustic image of the word. Such a difference could be accounted in the framework of known statistical approach to speech recognition [Bahl, 1983].

Let $X = \{x_t\}, t = 1, \ldots, T$ be a sequence of the vector parameters of the observed speech signal, $W = \{w_i\}, i = 1, \ldots, N$ be a sequence of the vocabulary words. The result of recognition of X, the most probable word sequence W^*, can be obtained from the equation [Jelinek, 1997]

$$W^* = \underset{W}{arg\,max}\, P(W|X) = \underset{W}{arg\,max}\, \frac{P(X|W)P(W)}{P(X)}. \tag{2}$$

The first factor $P(X|W)$ in the numerator (2) denotes the data likelihood conditioned the given word sequence and could be obtained with the help of the acoustic phone models. The value of second factor $P(W)$ is estimated with the help of the language model.

We use t^w to denote the phonemic transcription (pronounciation model) of a word w. The set of phonemic transcriptions of the given word w is denoted as T^w. The pronuncicaion model for the word sequence W is denoted T^W. The designation t^W will be used as a notation for the arbitarry sequence of word phonemic trancsriptions from T^W.

The conventional speech decoding and recognition procedures as a rule defines the best sequence of acoustical models (phonemic transcriptions), not the best sequence of words, that is instead of (2) de facto is used:

$$t^{W^*} = \underset{t^W}{arg\,max}\, \frac{P(X|t^W)P(t^W)}{P(X)}. \tag{3}$$

Then the most probable word sequence could be obtained by mapping of each pronunciation model to the corresponding word, i.e.:

$$t^{W^*} \rightarrow W^*. \tag{4}$$

If for all words there is a single pronunciation per word in vocabulary the methods (2) and (3) are equivalent.

Using the identity $P(t^W) = P(t^W|W)P(W)$ the expression (3) could be written as:

$$W^* = \underset{t^W}{arg\,max}\, \frac{P(X|t^W)P(t^W|W)P(W)}{P(X)}. \tag{5}$$

The expression (5) differs from the one of (3) in that it contains the factor $P(t^W|W)$ that accounts the pronunciaton variation. The set of probabilities $P(T^W|W) = \{P(t^W|W), t^W \in T^W\}$ is considered as parameters of the PV model.

Estimation of parameters of pronunciation variation modell

In order to utilize (5) we need to know the parameters of the three types of models: acoustic models, language model and the pronunciation model.

The language model parameters for estimation $P(W)$ usially considered as an independent of the acoustic models. Therefore the estimation of language model parameters could be performed in the independent manner exactly as it is done in conventional (9) approach.

The pronunciation model parameters $P(T^W|W)$ are dependent on the acoustic training data, therefore the independent (of acoustic one) estimation of $P(T^W|W)$ is not correct.

Consider the maximum likelihood estimate of the pronunciation model parameters.

Suppose that the traninig corpora X is such that for all its utterances we know the sequence of words $w_1w_2 \ldots w_N$ as well as a sequence of the phonemic transcriptions $t_1^wt_2^w \ldots t_N^w$. In such a case the most probable estimate of the parameters $p(t^w|w)$ will be obtained by solving:

$$p(t^w|w) = \underset{w,t^w}{arg\,max} \prod_{w,t^w} p(t^w|w). \tag{6}$$

This frequency estimate is similar to the estimate for the n–gram language model[Young, 1997]:

$$p(t^w|w) = \frac{\#\{t^w\}}{\#\{w\}}, \tag{7}$$

where $\#$ denotes the number of events in curly braces, encountered in the traning data. Therefore the most probable estimate for the given transcription will be the relative frequency of that transcription it in the traninig corpora.

Since the independent estimation of the acoustic and pronunciation model parameters is not correct consider the algorithm consisting of two-step iterations.

Suppose that there is a traninig speech corpora along with the vocabulary. Suppose that for each vocabulary word we know all of the pronunciation variants and consider for beginning the variants as equally probable.

On the first step the maximum likelihood estimates of the acoustic model parameters are obtained. The conventional training methods based on forward-backward and Baum-Welch algorithms can be used.

Then make the co-called «restricted» recognition of all utterances in the speech corpora. Term «restricted» means that the purpose is to find out the most probable sequences of phones given the true word sequences.

On the second step using (7) the maximum likelihood estimations of PV model parameters is obtained. It is done with the help of the co-called «restricted» recognition of all utterances in the speech corpora. Term «restricted» means that the true word sequence is known in advance and the target is to find out the most probable sequences of phones or, another words, sequence of transcriptions.

Then repeat the step 1 and retrain the acoustic models using the obtained on the step 2 the most probable sequencies of phones.

There steps can be repeated for the fixed number of times or until some stopping criteria will be reached.

The embedding of the pronunciation variation modeling into speech decoder

A conventional way to use the several pronunciation transcriptions per word in speech decoder consists of inclusion of each transcription to the pronunciaton vocabulary and treating this transcription in an independent manner as if it is a transcription of a new word. This approach implies no changes in the search algorithms (3)-(4).

It is not the optimal solution though.

Rewrite the expression (2):

$$P(W|X) = \frac{P(W,X)}{P(X)} = \frac{\sum_{t^W \in T^W} P(X,t^W)}{P(X)} = \frac{\sum_{t^W \in T^W} P(X|t^W)P(t^W)}{P(X)}. \tag{8}$$

From (5) and (8) it follows that the most probable sequence of words W^* should satisfy

$$W^* = \underset{W}{arg\,max} \sum_{t^W \in T^W} P(t^W|X)P(t^W). \tag{9}$$

Solution (9) let us define the most probable word sequence (not a most probable phone or transcription sequence) that is exactly what we need from the speech recognition system.

The algorithm (5) differs from the one of (3)-(4) in that we need to take into account the relative frequencies of word phone transcriptions and make the final decision using the weighted sum of the transcription likelihoods.

To implement (9) we need to make an additional, as compared to (3)-(4) calculations accordingly to (9).

Since for every word w:

$$P(w) = \sum_{t^w \in T^w} P(t^w|X), \tag{10}$$

then if, for example, a prefix tree lexicon representation is used, in every tree leaf the word likelihood should be estimated accordingly to (10).

Comparing with the conventional (3)-(4) approach it is also necessary to make some trivial changes in the search data structures and the memory allocaiton.

The practical implementation of the (9)-(10) is associated with the difficulty because of tree pruning [Young, 1997]. Suppose that some leafs of the prefix tree have been pruned because of the relatively small likelihoods. In such a case the likelihoods of these leafs are not known and the corresponding transcriptions could not be used in (10).

To overcome that difficulty consider the following version of (10):

$$W^* = \underset{W,t^W}{arg\,max}\, P(t^W|X)P(t^W). \tag{11}$$

Here the weighted sum of the likelihoods is replaced with the likilihood of the most probable transcription penalized with $P(t^W|X)$.

Numerical experiments

The performance of the considered PV models have been compared on the speech corpora ISABASE–2 [Bogdanov, 2004] and TeCoRus [Chuchupal, 2005]. The training data of the first test consisted of speech utterances of 200 speakers of ISABASE–2 (40K utterances) Đÿ 50 speakers from TeCoRus (3K utterances). The test material consisted of the 776 utterances that contained the connected digit strings (3147 digits). The vocabulary has been limited to the digits. The reason to use numbers was that the numbers and numerals could provide a lot of examples of pronunciation variations.

No language models has been used.

The recognition results in terms of word error rate (WER) values are presented in Tabl. 1. The column «Basic» contains the results for the case when the basic transcription is used only. The column «Convent.» correspondes the method (2 - 4). The column «Optim.» containes the results for the method (9). Column «SubOptim.» contains results for the method (11). The row «Variability» contains the mean number of transcriptions per vocabulary word.

The results depicted above could be interpreted as an evidence of lack of pronunciation variability in the test corpora. It can be true because the speakers of TeCoRus belong to the same high-edicated profeccisonal group and were born and living in Moscow region. The test material contained a read, carefully articulated speech.

Table 1: Word Error Rate for some pronunciaion variation models (TeCoRus data only)

Method	Basic	Convent.	Optim.	SubOptim.
WER	1.62	5.78	2.00	3.17
Variativity	1.0	1.9	1.9	1.9

The lack of the PV in the first test could explain the observed behavior of the training algoritm: on the TeCoRus data with the increasing number of iterations the mean number of transcriptions per word converged to one.

To obtain recognition results for the data with actual pronunciation variabiliry the second recognition experiment had been fullfilled. The training set of the second test was the same as in the first test. The test set consisted of 867 utterances of 11 test TeCoRus speakers. These data mostly consists of the sequences of numbers and numerals. The vocabulary of the test set consisted of 129 words. Test utterances also contain an additive and casual types of office noise as well as amount of the speech disfluencies that often led to the speech recognition errors.

The pronunciation vocabulary contains 129 numerals.

Talbe 2 shows the WER result for the second test. The table column «Convent.» shows the WER value for the case when the basic pronunciations were used only.

Table 2: WER value for some pronunciaion variation models with the TeCoRus extended data

Method	Basic	Convent.	Optim.	SubOptim.
WER	7.78	7.57	7.38	7.44
Variativity	1.0	1.3	1.3	1.3

The results drawn in (2) could be considered as more relevant to the expected. The best approach appears to be the one that corresponds to the frequency weighting of the pronunciation variants(9). The approach with the inclusion the rival transcription to the pronunciation vocabulary (2 - 4) appears to be less effective both the (9) as well as algotithm (11). In all cases the inclusion of the pronunciation variations appears to be more effective than the use of the basic transcriptions only.

The WER improvements in the second test were not so substantial as it could be expected though. On the one hand it could be because of the type of test material. At the same time the WER improvements observed might be because of the speech corpora TeCoRus and Plantronics had been collected in different conditions. TeCoRus had been recorded with the Senheiser professional microphone while ISABASE–2 corpora had been recorded with the cheap Plantronics microphones.

To clarify these issues the third recognition test had been performed on the speech corpora that contained the natural spontaneous speech that had been extracted from the radio interviews. We used the interviews downloaded from the radiostation «Echo Moscow» [Echo Moscow].

The initial set of rival pronunciation transcriptions for the numerals as well as their relative frequency were the same as in the previous test.

The inteviews have been automatically segmented. The utterances with the numerals have been found and extracted to the separate speech files. The test set consisted of 200 speech utterances of 2–4 words each, with total vocabulary of 91 words.

No language models were used during recognition.

Table 3 presents the results for this test. The table column «Equal.» contains the WER values for the method (9) in the case when the equal relative frequencies for all rival transcriptions were used.

The substantionally higher WER values obtained because of the lack of language model, mismatch between training and testing conditions for acoustic models, and noisy environment during an interviews.

Table 3: WER values for the differnet types of pronunciation variation models on the natural fluent spoken speech

Method	Basic	Convent.	Optimal.	SubOpt.	Equal
WER	69.3	57.44	59.7	60.0	59.5

In third test the observed relative improvements in WER was from 13,4% to more than 17,1% comparing to 5% relative improvement in previous test.

It is shown therefore that for fluently spoken numerals the use of PV models can lead to the substantional improve the accuracy of speech recognition.

Note that there are the other (besides of pronunciation changes) possible reasons of improvements the accuracy of recognition in third test are exists. There is a significant mismatch in the traning and testing data. The test data coded in MPEGt. Howeve if it was the case then the similar WER improvements were to take place in the second test. It had not happened though.

The observed absence of improvement in WER (compared with the other methods) for the methods with weighting of rival transcriptions can be explained from the point of the language modeling. The transcription weighting as well as using the number of rival word transcriptions for numerals has an effect that is similar the using of the unigram languge model. In the test material the relative numeral frequencies were much higher than for the other. The use of conventional method has an effect of using the bigger unigram weights for numerals that was relevant to the data of the test corpora.

Conclusion

The research of the methods for improving the automatic speech recognition accuracy through the use of pronunciation variation models is fulfilled. The probabilistic pronunciation variation model is formulated and well as the ways to estimate the model parameters. The numerical experiments shows that the implementation of the pronunciation variation models is an effective way to improve an accuracy of spontaneous speech recognition.

Acknowledgements

The work was supported with Russian Fund of Basic Research, project 14-01-00607

Bibliography

[Jelinek, 1997] Jelinek F. Statistical Methods for Speech Recognition. Cambridge, Massachusetts: The MIT Press, 1997.

[Wester, 2003] Wester M. Pronunciation modeling for ASR âĂŞ knowledge-based and data-derived methods // Computer Speech and Language. 2003. Vol. 17, P. 69-85.

[Fosler, 1999] Fosler-Lussier E. Dynamic pronunciation models for automatic speech recognition. Ph.D. thesis. University of California, Berkley, CA, 1999.

[Saraclar, 2004] Saraclar M., Khudanpur S. Pronunciation change in conversational speech and its implications for automatic speech recognition // Computer Speech and Language. 2004. Vol. 18(4). P. 375-395.

[Saraclar at al., 2000] Saraclar M., Nock H., Khudanpur S. Pronunciation modeling by sharing Gaussian densities across phonetic models // Computer Speech and Language. 2000. Vol. 14(4). P. 137-160.

[Zheng, 2003] Zheng J., Franco H., Stolcke A. Modeling word-level rate-of-speech variation in large vocabulary conversational speech recognition // Speech Communication. 2003. Vol. 41. P. 273âĂŞ285.

[Bahl, 1983] Bahl . R., Jelinek F., Mercer R. L. A maximum likilihood approach to continuous speech recognition // IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol 5. 1983. P. 179-190.

[Schwartz, 1990] Chow Y.-L., Schwartz R. The N-Best Algorithm: Efficient Procedure for Finding Top N Sentence Hypotheses // Proceedings of the International Conference on Acoustic, Speech and Signal Processing, ICASSP. 1990, P. 199-202.

[Young, 1997] Young S., Bloothooft G. (Eds.). Corpus-based methods in language and speech processing. Dordrecht: Kluwer Academic Publishers (Text, Speech and Language Technology series, Vol. 2), 1997.

[Bogdanov, 2004] Bogdanov D. S., Krivnova O. F., Podrabinovitch A. J., Arlazarov V;L. Creation of Russian Speech Databases: Design, Processing, Development Tools // Proceedings of the International Conference on Speech and Computers, SPECOM. Moscow, 2004.

[Chuchupal, 2005] Chuchupal V.J., Makovkin K.A., Chichagov A.V., Kuszetsov V.B., Ogarysyev V.F. Speech corpora TeCoRus. Data base registration sertificate 2005620205, 2005.

[Echo Moscow] http://www.echo.msk.ru.

Authors' Information

Vladimir Chuchupal *- Leading scientist, Dorodnicyn Computing Centre Russian Academy of Sciences, P.O. Box: 111333, Moscow, Vavilov Str, 40, Russia; e-mail: v.chuchupal@gmail.com*
Major Fields of Scientific Research: Speech recognition, signal processing,

Anton Korenchikov *- Student, Lomonosov Moscow State University, P.O. Box: 111333, Moscow, Vavilov Str, 40, Russia; e-mail:*
Major Fields of Scientific Research: Speech recognition