Accuracy Analysis of Generalized Pronunciation Variant Selection in ASR Systems

Václav Hanžl and Petr Pollák

Dept. of Circuit Theory, Czech Technical University, Prague
{hanzl,pollak}@fel.cvut.cz

Abstract. Automated speech recognition systems work typically with pronunciation dictionary for generating expected phonetic content of particular words in recognized utterance. But the pronunciation can vary in many situations. Besides the cases with more possible pronunciation variants specified manually in the dictionary there are typically many other possible changes in the pronunciation depending on word context or speaking style, very typical for our case of Czech language. In this paper we have studied the accuracy of proper selection of automatically predicted pronunciation variants in Czech HMM ASR based systems. We have analyzed correctness of pronunciation variant selection in forced alignment of known utterances used as an ASR training data. Using the proper pronunciation variant, more exact transcriptions of utterances were created for further purposes, mainly for the more accurate training of acoustic HMM models. Finally, as the target and the most important application are LVCSR systems, the accuracy of LVCSR results using different levels of automated pronunciation generation were tested.

1 Introduction

Development in the field of speech technology during several recent years, together with increasing power of computers, has allowed the application of Large Vocabulary Continuous Speech Recognition (LVCSR). It represents one of the most challenging application of speech technology today. Current LVCSR systems can reach high accuracy, especially for English. Also for Czech the acceptable results of LVCSR are available [1] or [2] and this is also our current main task of our research activities.

LVCSR represents very complex system composed of several principal modules and for all of them very high accuracy is required to be able to achieve acceptable accuracy, or Word Error Rate (WER) of such system. One of the key inputs for recognizing of the utterance is the pronunciation of particular words [1] or [2]. The basic form of this information comes standardly from pronunciation lexicon but real pronunciation may vary in many situations. Consequently, if we do not have the proper variant in our pronunciation dictionary the final accuracy of LVCSR is worse. Pronunciation variability is known and natural phenomena in speech communication and for Czech we can express the following reasons why these changes can appear:

A. Esposito and R. Vích (Eds.): Cross-Modal Analysis, LNAI 5641, pp. 399–408, 2009.

- *general context dependency*,
- *speed of uttered speech* - and speaking style in general,
- *emotions*, as a particular aspect of the speaking style,
- *dialect*,
- *different meanings* of the word resolved by the pronunciation.

All these problems are not unique for Czech only but they appear in particular languages [3] on different level and we would like to summarize it for Czech in this work and present the study of application of automatically generated pronunciation variants both during the training phase and during the application within ASR system testing. We will describe several typical changes which can be met in Czech language commonly with the analysis of the accuracy of proper pronunciation variant choice which is very important. It is supposed that it can bring the increasing accuracy at the level of trained acoustic HMM models of particular speech elements (phones or triphones), same as more precise decoding of recognized speech at the output of LVCSR system. Within the experimental part of this work we present automated tests together with several manual checks of proper pronunciation variant selection.

2 State of the Art

The basic and standard source of discussed phonetic content of words required by ASR is pronunciation lexicon. This approach is standardly used mainly for English as the language with the most developed level of speech technology applications. For Czech this approach is also adopted but it is more difficult due to increasing size of such lexicon due to higher number of inflected word forms. Moreover, Czech is the language with relatively strong rule between regular phonetic contents (orthoepic form) and written (orthographic) form of speech. The application of rule based conversion between orthographic and orthoepic word form can be an alternative to lexicon based selection of word pronunciations.

Both above mentioned approaches can have some advantages and disadvantages from the point of view of possible pronunciation variants. Within the lexicon approach we can generate automatically possible variant based on systematic replacement of particular phones by other ones. The disadvantage is mainly in the fact that we work with words independently without any context and consequently all possible pronunciation variants must be included in the lexicon in this case. Rule based system may implement well interword context of the pronunciation, on the other hand the stronger irregularities cannot be implemented in this system.

Experiments with pronunciation variants extension and selection based on real acoustic data were reported for the German language [4] [5].

The research in this field is logically part of our activities dealing with applications of ASR recognition. We have already created basic support for ASR in the form of pronunciation lexica which were created within different database collection, as the last one the lexicon of LC-StarII has been created. As the basis for pronunciation lexica generation we use our tool *transc* implementing conversion

rules for generation of pronunciation from orthographic form of the word (sentence) [6]. We have also realized experiments with automated re-annotation of irregularly appearing glottal stop within available databases [7] and our current work follows and generalizes these activities.

Main targets of this work could be split to the solution of following problems.

- *Summary of possible changes in pronunciation* - We would like to analyze precisely all possible changes in pronunciation due to different reasons as context, fluent speech style (spontaneous speech), emotional speech, etc.
- *Extension of the lexica* - Having the list of possible changes we would like to extend our lexica by these pronunciation variants of particular lexical entries.
- *Database re-annotation* - HMM forced alignment should be applied on databases to choose proper pronunciation variants and to obtain more precise transcription of available speech utterances.
- *Experimental part* - Within our experiments we analyzed the accuracy of pronunciation variant selection automatically by retraining of HMM models with more precise pronunciation and check of LVCSR WER. The rates of pronunciation changes were also studied. Finally, the proper selection of pronunciation variants was checked manually on small amount of data.

3 Variability of Czech Pronunciation

3.1 Phonetic Inventory

As this paper deals with the language which can be unknown for some readers, we would like to mention some brief introduction about phonetic inventory of Czech. The basic information can be available from SAMPA Alphabet WEB-page where standard set of Czech phonemes is available [8]. We work with 46 phoneme set containing 10 vowels, 3 diphtongs, 29 consonants (8 plosives, 4 affricates, 11 fricatives, 2 liquids, 4 nasals) completed by 2 special allophones, glottal stop, and schwa. We do not use three additional syllabic consonants as they are from the acoustic point of view the same as the non-syllabic versions of these phonemes.

3.2 Basic Pronunciation

The principle theoretical description of Czech phonetics and phonology is available in [9]. On the basis of this background we have created the basic form of the tool *transc* for conversion between orthographic and orthoepic form of the utterance [6]. This tool is continuously updated and more precise and its usage is possible for Czech words with regular pronunciation. It is used as the pronunciation predictor for the words which are not included in the pronunciation lexicon.

Basic and standard source of phonetic contents of the word are pronunciation lexica. We have created or we have participated in several projects within which large pronunciation lexica have been created. We have started with lexica

of collected speech databases as SpeechDat, SPEECON, or Car speech collection. Within our last project we have created Czech version of LC-Star lexicon containing more than 132 000 entries where approx. 84 000 of them represent general common words.

The pronunciations in these lexica were obtained by our rule based tool followed by manual correction of irregularities. Some of them have been created during annotation of speech DBs so they are based on real pronunciations by particular speakers. Other lexica are available also in Czech broadcast news databases from LDC. These lexica are sources for our further generalization of pronunciation exceptions.

3.3 Studied Changes in Pronunciation

Glottal-stop prediction
The results of this study were published in [10]. As our current work extends this study and as it uses similar methodology we are presenting the basic summary of these experiments.

The following *rules for glottal-stop prediction* were used: Firstly, the glottal stop was inserted at the *beginning of each word* starting by a vowel. Secondly, the glottal stop in *inner word position* was placed

- after word prefixes ("do-, na-, vy-, pod-, nad-, ... ") followed by vowels,
- in word composites starting with words ("pseudo-, spolu-, samo-, ... ") again followed by vowels,
- in numeral composites ("jedno-, dvoj-, ... ") also followed by vowels.

When the lexicon was extended by these variants of words with glottal stop, forced alignment was performed for whole Czech SPEECON database. It can be presented as recognition of present glottal stop and achieved results were analyzed. Basically following conclusions of this experiment were stated:

- presented glottal stop was usually localized very precisely,
- higher error rate was in missing glottal stop recognition,,
- i.e. the presence of glottal stop was slightly preferred by our models.

General changes in pronunciation
Studied regular changes in pronunciation are listed bellow. They represent the most important changes which can appear regularly or irregularly in different speaking styles of Czech language.

1. *Changes of voicing character of ending or starting consonant* - It represents context dependent change very frequent in Czech. Our grapheme to phoneme conversion rule works also with this context dependency. When lexicon is used both variants must be contained.

   Ex: "*nad pecí*" vs. "*nad botníkem*" :
   "`n a t  p e t_s i:`" vs. "`n a d  b o t J i: k e m`"

2. *Back propagation of soft characters* - It is already in-word change.
   Ex: "*botník*" : "`b o t J i: k`" vs. "`b o c J i: k`"
   Some Czech databases include this type of assimilation variants in a second extended version of a lexicon [11].
3. *Pronunciation of diphthongs "e_u" ("a_u")* - This represent very difficult problem as the boundary between pronunciation of "e_u" and "e u" is rather soft. Moreover, however there are exactly defined rules for pronouncing "e_u" or "e u", people exchange irregularly these variants many times to both sides.
   Ex: "*neuměl*" : "`n e u m J e l`" vs. "`n e_u m J e l`"
4. *Manually given pronunciation variants* - Such strongly different pronunciations appear especially in words of foreign origin without stabilized pronunciation or in Czech words with different meanings. These variants must be included in pronunciation lexicon only manually.
   Ex: "*email*" : "`i: m e j l`" vs. "`e m a j l`"
   "*panický*" : "`p a n i t_s k i:`" vs. "`p a J i t_s k i:`"

4 Experiments

We have tested the influence of accuracy of ASR on proper pronunciation variant selection. The main target of realized experiments was to analyze possible improvement of target ASR accuracy and particular pronunciation variants selection. We did not want to perform too many manual checks so we tried to substitute them by following automated analysis.

4.1 Results with ASR

Basic setup
Firstly, basic speech recognition was performed using the model trained after precision of the pronunciation by forced alignment. The basic setup of our recognizer was as follows:

- The experiments were realized with speech sampled at 16 kHz. The features were composed of 16 MFCC coefficient plus signal energy, commonly with Δ and $\Delta\Delta$ parameters, i.e. MFCC_E_D_A in HTK notation.
- Acoustic HMM models are based on 3-state monophones (i.e. 3 emitting states) with 1 mixture.
- Acoustic HMM models were trained on 170 hours of speech from database Czech SPEECON [12], including rather clean utterances from office, entertainment, and car environment with low background noise level.
- Test setup. Recognizer was constructed on the grammar without any language modelling. The loop of equally probable 11782 words without any out-of-vocabulary word.
- This very simple ASR without any further support was used as a tool for the most clear analysis of pronunciation variant selection which should give the information about the contribution of proper pronunciation variant selection.

Multiple alignment cycles
The alignment of trained acoustic models on training data with the purpose of selection proper pronunciation variant when more than one is available is standardly used procedure. Within our experiment we realized:

1. the application of this procedure to above mentioned general changes of pronunciation which was previously automatically generated,
2. the alignment was applied iteratively more than once.

In the table 1 or figure 1 we can find the results of above mentioned experiments with iteratively applied forced alignment. Recognition accuracy (Acc) is defined as the ratio of the sum of substituted, inserted, and deleted words to the number of whole words in recognized text. Basically we can summarize these experiments in the following points:

- It is possible to obtain Acc=15.36% using *baseline system* with random selection of variant (using 1st variant is unusable, we might get no samples for "G" in "*abych byl*" etc.)
- After the first alignment and 3 retraining cycles (Align 1) the accuracy has been increased to Acc=17.74%. This is standardly used procedure during the training of ASR.
- When forced alignment is applied iteratively always after 3 retraining cycles, the accuracy Acc=27.37% can be achieved after the re-alignment and re-training at the 7-th step. It represents 78% relative improvement of the accuracy with respect to baseline system, and still 54% relative improvement above the single one re-alignment and re-training.
- Interesting comparison can be done with more retrainings after single one re-alignment. In this situation the accuracy is saturated at the value Acc=23%

Table 1. Achieved accuracy of pure recognition without language model

Training step	**Baseline**	Align 1	Align 2	Align 3
Acc [%]	**15.36**	17.74	21.58	23.29
Training step	Align 4	Align 5	Align 6	**Align 7**
Acc [%]	24.84	25.67	26.40	**27.37**

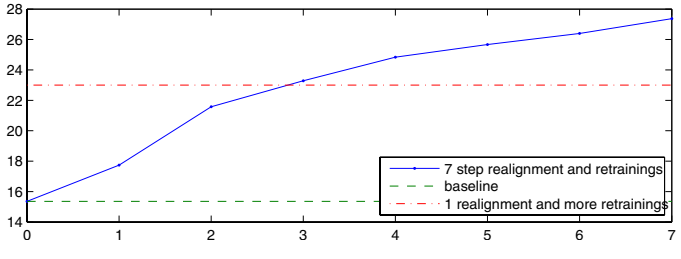

Fig. 1. Trends of increasing accuracy of pure recognition without language model

(dash-dotted red line in the figure 1). The best result achieved with 7 step re-alignment and re-training gives still approx. 20% relative improvement with respect to this value.

– For the comparison, when triphone multimixture models were used in the baseline system, 70-75% word accuracy was reached in this test.

4.2 Analysis of Phone Change Rates

Together with the above mentioned analysis of recognition accuracy we have also analyzed the amount of exchanges between particular phones in word pronunciations after each forced alignment. Within this experiment the rates of changes among 3 million phones in training data were computed. It should give the answer if this iterative re-alignment converges to more stable solution. The results are in the following table 2 or figure 2.

Table 2. Changes after re-alignments among 3 million phones

	1-2	2-3	3-4	4-5	5-6	6-7
d → t	3776	3239	3066	2606	2119	1641
t → d	1099	1025	614	470	445	673
z d → s t	2774	876	346	220	182	119
z → s	1524	564	311	184	132	77
S → t_S	1247	507	219	126	66	64
i: → i	901	424	160	48	32	28
G → x	592	363	253	201	179	140
g → k	547	268	173	124	107	120
e u → e_u	436	137	62	24	13	7
a: → a	436	121	25	7	5	0
s → z	227	212	104	55	29	32
k → g	281	172	108	49	28	36
d_Z → t_S	125	147	177	126	48	23
x → G	254	123	65	67	36	24
d_z d → c t	137	117	73	66	48	56
J/ → d	245	68	48	38	47	54
f → v	161	80	48	36	34	37

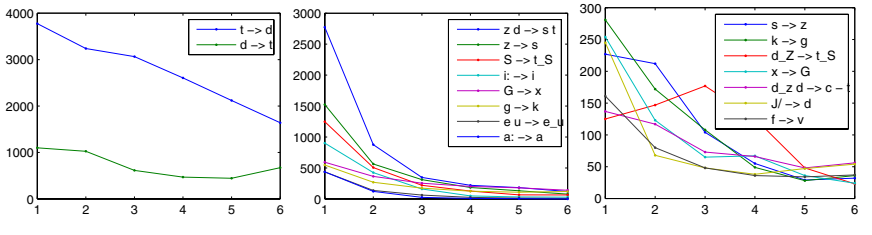

Fig. 2. Trends of changes after re-alignments among 3 million phones

This table is a result of rather huge computational effort, far exceeding usual pronunciation variant selection phase in LVCSR systems construction. Usual one phase alignment would correspond to the first column only. We can however see that the complex interaction between pronunciation variant selection and acoustic HMM training is far from settled even in the rightmost column. Selected changes presented in the table give some idea about the nature of this convergence in the complex search space.

4.3 Results of Manual Checks

Finally, also manual check was performed on small amount of data. We have selected the manual analysis of the most frequent changes appearing above in the table 2 or in the figure 2, i.e. choice between the pronunciation of "`d`" vs "`t`" at the and of the word.

For this purpose we have checked 226 randomly selected sentences with possible "`d`" vs "`t`" exchange. It means approximately 20 minutes of speech. The following most important observations have been done:

1. Only 45% human/computer agreement was observed after first alignment and 53% after 7th re-alignment. It means rather small correlation but the improvement after iterative re-alignment was reasonable.
2. 98% of above mentioned mismatch was in human preference of "`t`" vs. computer preference of "d". On repeated closer examination of data we concluded that not only the computer decision contained errors but also the human decision was quite often wrong, as described below.
3. Our mind does not like cases like "`p j e d h o d i n`" – with voiced assimilated pronunciation – when orthography suggests otherwise: "*pět hodin*". This type of errors have to be expected in all transcriptions made by annotators with mostly technical education and only marginal background in phonetics. (This is in sharp contrast with the well known opposite case of voiced final phone assimilated to voiceless.)
4. Automatic choice based on acoustic data sometimes strongly prefers variants which are theoretically impossible or at least plain wrong, like "`o p j e d s l a v i: t r i u m f`" for "*opět slaví triumf*". On closer examination, some strange cases like this really happen but quite often we found yet more complex assimilation to yet other phone sequences. Here the automatic procedure had to choose between the two variants but neither of them was the real pronunciation.

5 Conclusions

The paper presented detailed analysis of importance of proper pronunciation variant selection for accurate speech recognition. The most important contributions can be summarized in following points.

– We have summarized the most important changes in pronunciation due to context dependency and different speaking styles which appear frequently in pronounced Czech speech.

- It was proved that already very small percentage of wrong pronunciation variants in the training material severely degrades ASR performance so at least one forced alignment procedure is necessary. This represents standard training procedure.
- It was found that more than one forced alignment phase followed by several retraining cycles can bring further improvement (compared to one variant-selection alignment described in classical tutorials [13]). It was possible to observe reasonable increasing of the accuracy of our testing ASR.
- The decreasing number of changes between particular iterative alignments proved that this iterative re-alignment yields to a stable solution, however the correlation between human and automated variant selection was not too high.
- Systematic variant bias needs many iterations to be eliminated.
- Relative improvement of WER: 12% in early stages.
- As a future work we plan mainly to analyze the influence of multi-step re-alignment on models with full complexity, i.e. using triphones and multi-mixture structure. We suppose 70-75% accuracy reached after 1st alignment should be improved by realignment steps 2 - 7.
- Altogether, we found that it probably makes sense to devote an order of magnitude more computational effort to good automatic selection of pronunciation variants than is usual in preparation of LVCSR systems. Moreover, using hand-labelled bootstrap data may be no good substitute for these expansive iterative procedures based on automatic processing of the acoustic data: We found human bias to be far too strong to allow human annotators to serve as a reliable etalon.
- These effects were studied on somewhat artificial LVCSR system strongly adapted to the purpose of intended experiments with acoustic properties of phones. Influence of higher levels (dictionary, syntax etc.) was minimized as much as possible to get clearer picture of phone changes. There is however no doubt that for the purpose of LVCSR itself, similar more complex study including all these higher level interactions would be valuable, though it might result in changes in data which will be rather hard to attribute to the pronunciation selection only.

Acknowledgements

The research was supported by grants GAČR 102/08/0707 “Speech Recognition under Real-World Conditions” and by research activity MSM 6840770014 “Perspective Informative and Communications Technicalities Research”.

References

[1] Psutka, J., Ircing, P., Psutka, J.V., Hajič, J., Byrne, W.J., Mírovský, J.: Automatic transcription of Czech, Russian, and Slovak spontaneous speech in the MALACH project. In: Proc. Interspeech 2005, Lisbon, Portugal, pp. 1349–1352 (2005)

[2] Nouza, J., Ždánský, J., David, P., Červa, P., Kolorenč, J., Nejedlová, D.: Fully automated system for Czech spoken broadcast transcription with very large (300K+) lexicon. In: Proc. Interspeech 2005, Lisbon, Portugal, pp. 1681–1684 (2005)
[3] Dupont, S., Ris, C., Couvreur, L., Boite, J.-M.: A study if implicit and explicit modeling of coarticulation and pronunciation variation. In: Proc. Interspeech 2005, Lisbon, Portugal, pp. 1353–1356 (2005)
[4] Wolff, M.: On representation and training of pronunciation dictionaries. In: 8th Czech-German Workshop 'Speech Processing', Prague, Czech Republic (1998)
[5] Wolff, M., Eichner, M., Hoffmann, R.: Evaluation of automatically trained pronunciation dictionaries. In: Proc. Czech-German WS on Speech Processing, Prague, Czech Republic (2002)
[6] Pollák, P., Hanžl, V.: Tool for Czech pronunciation generation combining fixed rules with pronunciation lexicon and lexicon management tool. In: Proc. of LREC 2002, Third International Conference on Language Resources and Evaluation, Las Palmas, Spain (May 2002)
[7] Pollák, P., Volí, J., Skarnitzl, R.: Influence of hmm's parameters on the accuracy of phone segmentation - evaluation baseline. In: ESSP 2005, Electronic Speech Signal Processing, Prague (September 2005)
[8] Wells, J.C., et al.: Czech SAMPA home page (2003), http://www.phon.ucl.ac.uk/home/sampa/czech-uni.htm
[9] Palková, Z.: Czech phonetics and phonology. In: Czech language - Fonetika a fonologie češtiny, Charles University. Karolinum (1994)
[10] Pollák, P., Volí, J., Skarnitzl, R.: Analysis of glottal stop presence in large speech corpus and influence of its modelling on segmentation accuracy. In: 16th Czech-German Workshop on Speech Processing, Prague (September 2006)
[11] Psutka, J., Müller, L., Matoušek, J., Radová, V.: Mluvíme s počítačem česky (Talking to the Computer in Czech). Academia, Prague (2006)
[12] Pollák, P., Černocký, J.: Czech SPEECON adult database (November 2003), http://www.speechdat.org/speecon
[13] Young, S., et al.: The HTK Book, Version 3.3, Cambridge (2005)