Non-native Pronunciation Modeling in a Command & Control Recognition Task: A Comparison between Acoustic and Lexical Modeling

Judith Kessens

TNO Human Factors, Soesterberg, The Netherlands
judith.kessens@tno.nl

Abstract

In order to improve automatic recognition of English commands spoken by non-native speakers, we have modeled non-native pronunciation variation of Dutch, French and Italian. The results of lexical and acoustical modeling appeared to be source language and speaker dependent. Lexical modeling only resulted in a substantial improvement (of 35%) for the French speakers. Acoustic model adaptation halved the word error rates for the Italian speakers, whereas no improvements were found by lexical modeling of frequently observed Italian-accented non-native pronunciation variants. The performance for the Dutch speakers only slightly improved by lexical and acoustic modeling.

1. Introduction

Within the EU project Safesound we studied the possibilities of improving safety for ground and flight operations by the application of enhanced audio functions in the cockpit of an airplane. One of these enhanced audio functions is Direct Voice Input, achieved by means of an automatic speech recognition (ASR) system. In aviation, English is the lingua franca but it is not the native language of most of the pilots. Therefore, the speech recognition system should be able to cope with non-native pronunciations of speakers from a wide variety of language backgrounds. This paper describes a number of experiments in which we tried to improve recognition performence by modeling the accents of the non-native pilots.

2. Non-native pronunciation modeling in ASR

In [1], [2] extensive overviews are given of approaches of modeling (mainly native) pronunciation variation. A distinction can be made between *data-driven* and *knowledge-based* approaches. In data-driven studies the information on the pronuniciation variation is obtained from the data, whereas in knowledge-based studies, the information is obtained from sources that are already available, like pronunciation dictionaries and linguistic studies. In contrast to native pronunciation variation, non-native pronunciation is mainly modeled using data-driven approaches. The main reasons are probably that non-native pronunciation is extremely speaker-dependent and is affected both by the mother tongue of the non-native speaker (=source language) and by the language that the non-native is trying to speak (=target language).

In [1], [2] also a distinction is made between modeling pronunciation at the level of the acoustic models, the lexicon and the language model. In the current study, we investigate the effectiveness of modeling non-native variation at the level of the acoustic and lexical level. Furthermore, we investigate whether the use of variants-specific priors in the the language model (grammar) can further improve the lexical modeling.

2.1. Acoustic modeling of non-native accents

The most obvious way of modeling non-native accents is to train acoustic models from scratch using non-native data [3] or to adapt the native models with non-native accented speech [4]. However, non-native speech material will not always be available. Therefore, it has been tried to adapt the acoustic models with speech from either the source or target language:

- Acoustic models from the source language are used [5]
- A mixture of models from both the source and target language are used [6],
- The native acoustic models are adapted with speech from the source language[7]
- Acoustic models are interpolated with adequate native models [8] or native and non-native models [9].

2.2. Lexical modeling of non-native pronunciations

To a certain extent segmental non-native variation can be modeled by training context-dependent HMM models. However, it has been shown in [10] that some variation such as syllable deletion can not be captured in this way. The non-native pronunciation variation can be obtained in a similar way as it is obtained for native pronunciation variation (see [1, 2]):

- *Knowledge-based*: By using phonological rules which describes how the non-native pronunciation deviates from the native pronunciation for a source-target language pair (see e.g. [6], [11]).
- *Data-driven*: By using data-driven rules [11] or manual phonetic transcriptions [12] of non-native accented speech, or by hypothesis testing, e.g. testing of vocalic substitution hypotheses [13]. Goronzky et al. [14] obtain hypotheses on non-native pronuniation variation based on non-native speech.

In this study, a data-driven approach for modeling the non-native pronunciation variation is applied.

3. Method

3.1. Recognition task and speech recognizer

The speech recognition is part of a dialogue system in which the simulated avionics instruments react on the voice input. The recognition task is a small vocabulary (240 words) command and control task. Commands includes for instance:

page selections, display changes and parameter settings. Examples of commands are:

"HF1 two nine three five"
"FMS flight plan departure page"
"set cost index one hundred"

A rule-based grammar (Java Speech Grammar Format [15]) is used to model the command structure. Loquendo ASR 6.0.2 [16] software is used to perform the recognition experiments. Loquendo ASR uses a hybrid Hidden Markow Model (HMM) and Artificial Neural Network (ANN) recognition system where each phonetic unit is described in terms of single or double state left-to-right automaton with self-loops. The acoustic models are based on a set of stationary context-independent phones and diphone-transition coarticulation models [17]. Each 10 ms telephone bandwidth features including deratives are calculated; for more details, see [18]. The US English acoustic models are trained using the US English Macrophone database from LDC [19], consisting of 200,00 utterances of 5,000 speakers from all regions of the US. Phonetic transcriptions are automatically obtained using a rule based language-dependent phonetic grapheme to phoneme converter.

3.2. Recognition experiments

3.2.1. Acoustic modeling

For the baseline recognition experiments we measured recognition performance using the standard US and UK English models. Next, the US English acoustic models were adapted with non-native accented speech. The Italian accented speech from db2 was used in order to adapt the US English models to the Italian accent. To adapt the US English models to the Dutch accent, the Dutch speakers of db1b were used. A multi-speaker adaptation technique was used which is an enhanced version of Linear Input Network for Neural Networks (see [20]).

3.2.2. Lexical modeling

In order to obtain non-native pronunciation variants, manual phonetic transcriptions were made of db1b. The transcriptions were made by an experienced phonetician with a Dutch mother tongue. The task was to map the pronunciations to the sequence of US-English phones that closest matches the non-native pronunciation. In earlier research on modeling of native pronunciation variation [21], several measures that might predict improvements in recognition results due to pronunciation modeling were investigated. It appeared that there exists a strong correlation between the *absolute* frequency of occurrence of a pronunciation rule and its contribution to the net improvement in recognition performance. Therefore, the absolute variant frequency (F_{abs}) is used as a criterion for variant selection. F_{abs} is defined as the variant count divided by total number of words in db1b. The threshold for F_{abs} was varied from 1% to 0.02%. Two testing conditions were applied:

1) All variants have equal probability,
2) Variants are assigned priors estimated from the frequency of occurrence in db1b. In order to obtain reliable prior estimates, only priors were used for words with a frequency of ten or more.

3.2.3. Speech Material

Our speech material was recorded in two different sessions. Each speaker had to read a number of commands in English. The commands were randomly generated with two different versions of the syntax, resulting in two databases: database 1 (db1) and database 2 (db2)

Db1 is divided in two independent test sets; db1a and db1b. Db1a is used for performing the recognition experiments, whereas db1b is used to obtain the non-native pronunciation variants. For acoustic model adaptation with the Dutch non-native accent, the speech of the Dutch speakers from db1b are used. The Italian speakers of database 2 (db2) are used for acoustic model adaptation of the Italian non-native accent. No overlap in speakers exists between the three databases. The statistics of the two databases are given in Table 1.

Table 1: Statistics of non-native speech databases

	Native language	#speakers	#comm./ speaker	#words/ comm.
db1a	Italian	4	100	6
	French	6	100	6
	Dutch	7	100	6
db1b	Italian	4	100	6
	French	6	100	6
	Dutch	7	100	6
db2	Italian	11	226	4

4. Results

Recognition performance is measured in terms of Word Error Rate (WER), which is defined as the total percentage of word substitutions, deletions and insertions. For each group of non-native speakers, mean WERs are reported. For all the experiments a recognition lexicon is used containing all possible words that can be used in the command syntax (240 words).

4.1. Baseline performance

Baseline recognition performance was measured using the US and UK English acoustic models. The results, which are summarized in Table 2, show that for all native languages optimal performance is obtained using the US English models. Furthermore, the WERs indicate that the amount of non-native pronunciation variation is largest for the French and Italian speakers.

Table 2: Baseline WERs (db1a)

Native language	US	UK
Italian	7.1%	9.2%
French	6.5%	8.0%
Dutch	1.4%	2.4%

4.2. Acoustic model adaptation

The results in Table 3 show that the error rates are improved by acoustic model adaptation. The improvement for the Dutch speakers is small, whereas a significant improvement is

obtained for the Italian speakers: the WERs are halved by using adapted models. However, the improvements due to acoustic model adaptation are speaker dependent; the relative WER reductions per speaker varies from 27-71%. These results are in line with the improvements found when HMM recognizers are adapted with non-native accented speech [5].

Table 3: WERs using baseline and adapted models

Native language	baseline	adapted
Italian	7.1%	3.3%
Dutch	1.4%	1.2%

4.3. Lexical modeling

In Table 4 the Phone Mismatch Rates (PMR) for the various source languages are reported. The PMR is defined as the percentage of substituted (S), inserted (I) and deleted (D) phones in the manually obtained phonetic transcriptions compared to the automatically derived phonetic transcriptions. The PMRs seems to be determined by a large number of substitutions.

Table 4: PMRs for the manual vs. automatic transcriptions of db1b

Native language	S	D	I	PMR
Italian	12.1%	2.6%	2.0%	16.6%
French	13.3%	1.8%	1.3%	16.4%
Dutch	13.4%	2.8%	0.4%	16.5%

The most frequent non-native pronunciation mismatches as a percentage of the total PMR are shown in Table 5 (IPA notation is used [22]). In Table 5 , only percentages ≥ 3.0% are given; "-" means that the percentage is smaller than 3.0%.

Table 5: Most frequent non-native pronunciation mismatches

Rule	Dutch	Italian	French
/ɪ/ → /i/	6.3%	7.6%	19.0%
/e/ → /ə/	5.1%	3.4%	6.7%
/d͡ʒ/ → /t͡ʃ/	7.6%	7.1%	4.6%
/ɑː/ → /ɔː/	4.1%	3.0%	4.2%
/ə/-deletion	5.4%	3.8%	-
/æ/ → /e/	11.3%	-	-
/v/ → /f/	5.1%	-	-
/d/ → /t/	4.2%	-	-
/ɑː/ → /ʌ/	3.6%	-	-
/t/-deletion	-	7.0%	-
/æ/ → /ɑː/	-	3.6%	-
/ə/ → /ɑː/	-	3.4%	-
/ʌ/ → /ɑː/	-	3.3%	-
/æ/ → /ʌ/	-	3.0%	-
/æ/ → /ə/	-	-	4.1%

It can be seen that for all of the non-native accents some common native variations were found (e.g. /e/ → /ə/). Other variations are typically non-native variations, like vowel-lengthening (e.g. /ɪ/ → /i/). Some of the non-native variations are typically dependent on the native language, e.g. the voiced-unvoiced confusions (/v/ → /f/ and /d/ → /t/) for the Dutch speakers.

4.3.1. *Number of variants in decoder output*

In Figure 1, the percentage of variants counted in the decoder output is displayed as a function of F_{abs}. A variant is a pronunciation that is different than the automatically generated pronunciation. In going from left to right in Figure 1, F_{abs} becomes smaller, meaning that the number of added variants becomes higher. The percentages of variants are given as a percentage of the total number of correctly and incorrectly recognized words. It can be seen that in general the inclusion of more variants in the lexicon results in a higher percentage of variants that are recognized. For F_{abs}= 0.025%, the percentages of variants for the incorrect words are higher than for the correct words, which might indicate the number of errors that are solved becomes higher than the number of errors that are introduced.

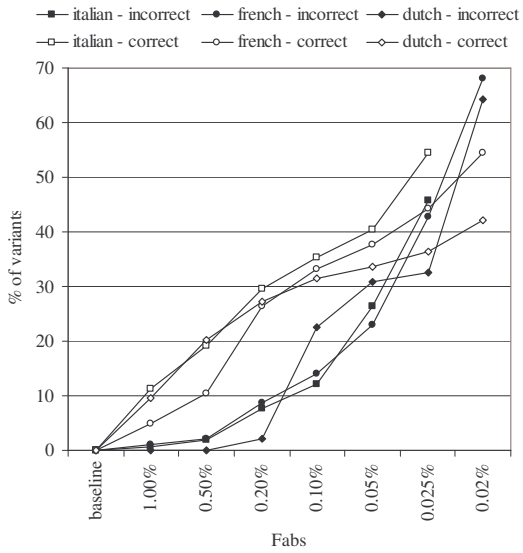

Figure 1: Percentage of alternative variants counted in the decoder output as a function of F_{abs}

4.3.2. *Lexical modeling results per native language*

Figure 2 shows the mean WERs per source language as a function of the number of variants that has been added (the lower Fabs, the larger the number of added variants). For all native languages, the same general pattern is found: the more variants are added, the larger the improvement. However, when a large amount of variants is included, performance can actually deteriorate. This pattern has been found in other studies as well (e.g. [23]).

The amount of improvement found due to lexical modeling are dependent on the native language: For the Italian and Dutch speakers, only small improvements are found. For the French speakers a large, significant reduction in WER of 2.3% (35% relative improvement) is obtained.

Figure 2 also shows that using variant-specific priors only slightly influence recognition performance compared to using equal weights.

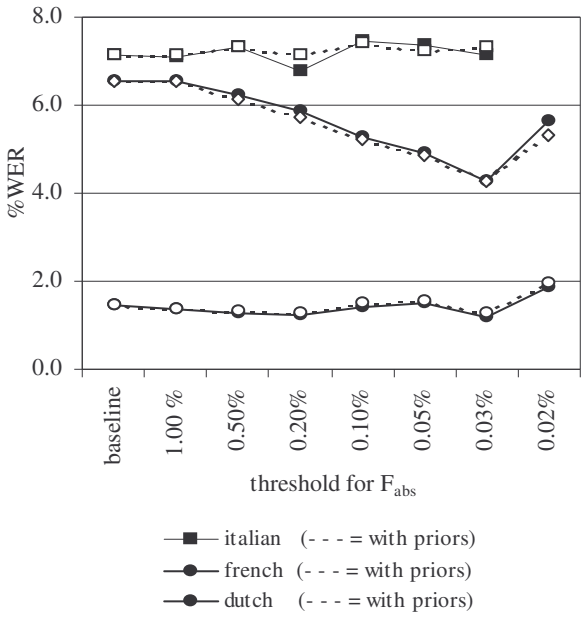

Figure 2: Lexical modeling results per native language

4.3.3. Speaker dependent results

Closer inspection of the recognition rates reveals that the results are speaker-dependent. For the Dutch speakers, only small changes in WERs are found. For the Italian speakers, the changes in mean WERs are mainly caused by the worst performing speaker. Figure 3 shows the speaker-dependent results for the French speakers. Figure 3 shows that the mean WER reduction pattern is reflected in the individual patterns of all French speakers, but the pattern is most prominent for the four worst performing speakers. For these speakers, WERs reductions of 30-45% relative are found.

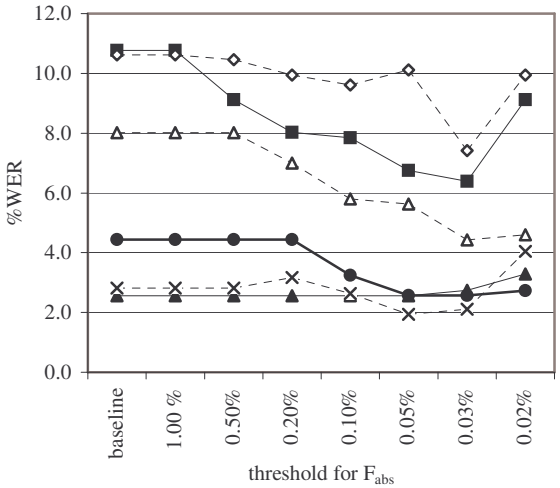

Figure 3: Lexical modeling results for the individual French speakers. Each curve represents results for one speaker.

5. Discussion

An interesting question that arises from this study is why the French speakers benefit from lexical modeling, whereas this is not the case for the Italian and Dutch speakers. For the Dutch speakers, the result can be explained by the fact that less non-native variation need to be modeled as the WERs are quit low. The results of acoustic modeling might indicate that for the Italian speakers, the acoustic differences deviate more from the US English pronunciation and can only be accurately modeled by acoustic model adaptation. Other factors that might play a role will be discussed below.

Closely related to speaker-dependency is the generalizability of the non-native pronunciation variation. Is the variation found for a set of non-native speakers representative for other speakers with the same native language? For the results analyzed in this study, the answer is yes. The general rules as reported in Table 5 seem to be consistent among speakers. Furthermore, a self test (db1b) shows similar results as for the independent test set (db1a).

Another explanation why the lexical modeling in this study is not always successful is related to the acoustic variability of the non-native variation. The PMRs in Table 4 show that the number of substitutions is much higher than the number of deletions and insertions. Modeling substitutions at the level of the lexicon is probably less adequate since the modeling units still deviates from the non-native pronunciation. Another factor that plays an important role is the complexity of the acoustic models. He et al. [24] showed that for modeling of non-native variation with existing models of the target language, highly complex models are less adequate than models with a moderate complexity. The explanation for this finding can be that there is a mismatch between the commonly used triphones used by natives and the triphones used by non-natives. In this study, coarticulation is modeled in the diphone-transition models. The non-native pronunciation variants that are added to the lexicon contain diphone contexts that are rarely used by the non-natives or that do not match with the native coarticulations.

Yet another, probably less important factor is that both native as non-native pronunciation processes are modeled in this study. For instance, the vowel reduction process /e/ → /ə/ occurs often in native US English and was already implicitly modeled in the acoustic models. In this case, explicit modeling of the vowel reduction process can be counter productive.

Other authors (e.g. [23] pp.118, [25], [26]) have pointed out that variant-specific priors need to be used in order to ensure recognition improvements for lexical modeling. The reason why the use of variant priors is not beneficial in our experiments is probably the recognition task: For the command & control application a final state grammar is used. The use of a strict grammar reduces lexical confusability compared to, for instance, a connected speech large vocabulary recognition task.

In future work, we will combine lexical modeling and acoustic model adaptation. Goronzy [12] showed that lexical modeling of non-native pronunciation can further improve acoustic models that have been adapted with non-native accented speech. Furthermore, we plan to perform acoustic model adaptation with lower complexity models.

6. Conclusions

The results from this study show that both lexical and acoustic model adaptation with non-native accented can substantially improve recognition performance. However, the results are source language and speaker dependent. Using variant-specific priors in combination with lexical modeling does not further improve recognition performance.

7. Acknowledgements

The research presented in this paper was carried out within the framework of the EU SafeSound project, contract G4RD-CT-2002-00640. We would like to thank Loquendo for providing us with the ASR software and providing us with special script for supervised acoustic model adaptation. Special thanks goes to Luciano Fissore for all his help and for the useful conversations about the topic of non-native pronunciation modeling.

8. References

[1] Strik, H., Cucchiarini, C., "Modeling pronunciation variation for ASR: A survey of the literature", Speech Communication 29, pp. 225-246, 1999.

[2] Strik, H., "Pronunciation adaptation at the lexical level", proc. of ISCA workshop on Adaptation medhods for Speech Recognition, Sophia-Antipolis, France, Vol. 1, pp. 123-130, 2001.

[3] Mayfield Tomokiyo, L., Waibel, T., “Adaptation methods for non-native speech in LVCSR”, proc. of ICSLP’00, pp. 1619-1622, 2000.

[4] Zavagliakos, G., “ Maximum A Posteriori Adaptation For Large Scale HMM Recognizers”, proc. of ICASSP’96, pp. 725-728, 1996.

[5] van Leeuwen, D., Orr, R., “Speech recognition of non-native speech using native and non-native acoustic models”, proc. of RTO workshop MIST, RTO-MP-28 AC/323(IST)TP/4, Neuilly-sur-Seine, pp. 27-32, 2000.

[6] Bartkova, K., Jouvet, D., “ Language based phone model combination for ASR adaptation to foreign accent”, ICPhS’99, San Francisco, USA, pp. 1725-1728, 1999.

[7] Kat, L.W., Fung, P., “MLLR-based accent model adaptation without accented data”, proc. of ICSLP’00, Beijing, Vol. 3, pp. 738-741, 2000.

[8] Steidl, S., Stemmer, G., Hacker, C., Nöth, E., “Adaptation in the Pronunciation Space for Non-Native Speech Recognition”, proc. of ICSLP’04, 2004.

[9] Wang, Z., Schultz, T., Waibel, A., “ Comparison of acoustic model adaptation techniques on non-native speech”, proc. of ICASSP’03, pp. 540-543, 2003.

[10] Jurafsky, D., Ward, W., Jianping, Z., Herold, K., Ziuyang, Y., Sen, Z., “ What kind of pronunciation variation is hard for triphones to model?”, proc. of ICASSP’01, pp. 577-580, 2001.

[11] Amdall, I., " Learning pronunciation variation: A data-driven approach to rule-based lexicon adaptation for automatic speech recognition", Ph.D. thesis, NTNU, Norway, 2002.

[12] Goronzy, S., Eisele, K., "Automatic pronunciation modeling for multiple non-native accents", proc. of ASRU'03, pp. 123-128, 2003.

[13] Raux, A., "Automated Lexical Adaptation and Speaker Clustering based on Pronunciation Habits for Non-Native Speech Recognition", proc. of ICSLP'04, 2004.

[14] Goronzy, S., Rapp, S., Kompe, R., “Generating non-native pronunciation variants for lexicon adaptation”, Speech Communication 42, pp. 109-123, 2004.

[15] http://java.sun.com/products/java-media/speech

[16] http://www.loquendo.com

[17] Fissore, L., Laface, P., Ravera, F., “Acoustic-Phonetic Modeling for Flexible Vocabulary Speech Recognition”, proc. of Eurospeech’95, pp. 799-802, 1995.

[18] Albesano, D., Gemello, R., Mana, F., “Hybrid HMM-NN Modelling of Stationary-Transitional Units for Continuous Speech Recognition”, Int. Conf. On Neural Information Processing, pp. 1112-1115, 1997.

[19] http://www.ldc.upenn.edu

[20] Gemello R., Mana F., Scanzio S., Laface P. and De Mori, R. "Adaptation of hybrid ANN/HMM models using linear hidden transformations and conservative training", accepted at ICASSP’06

[21] Kessens, J.M., Cucchiarini, C., Strik, H., "A data-driven method for modeling pronunciation variation", Speech Communication 40(4): 517-534, 2003.

[22] http://www2.arts.gla.ac.uk/IPA/index.html

[23] Kessens, J.M., ‘Making a difference: On automatic transcription and modelling of Dutch pronunciation for automatic speech recognition”, Ph. D. thesis, University of Nijmegen, The Netherlands, 2002.

[24] He, X., Zhao, Y, "Model complexity optimization for nonnative English speakers", proc. of Eurospeech'01, vol. 2, pp. 1461-1464.

[25] Saraçlar, M. “Pronunciation Modeling for Conversational Speech Recognition”, Ph. D. Thesis, John Hopkins University, Baltimore, Maryland, 2000.

[26] Yang, Q., Martens, J.P., ‘On the importance of exception and cross-word rules for the data-driven creation of lexica for ASR’, proc. of ProRisc. Workshop, Veldhoven, The Netherlands, 589-593, 2000.