IMPROVED PHONEME SEGMENTATION OF GERMAN-ACCENTED ENGLISH BY LEXICON AND ACOUSTIC MODEL ADAPTATION

Maria Paola Bissiri, Ivan Kraljevski and Rüdiger Hoffmann

TU Dresden, Chair for System Theory and Speech Technology, Dresden, Germany
{maria_paola.bissiri, ivan.kraljevski, ruediger.hoffmann}@tu-dresden.de

Abstract: In the present study, a German ASR system was used to perform phoneme segmentation of German-accented English speech. The phoneme models were created on German training data and the used lexicon consisted of English words whose pronunciation was represented by means of the German phoneme inventory. The production of accurate segmentation is significantly affected by the language mismatch between the German training data and the German-accented English test data. In order to reduce this mismatch, enhancement of the lexicon and of the phoneme models was performed. The lexicon was enhanced by means of pronunciation rules for German-accented English and according to recognition results analysis. Acoustic model adaptation was carried out to reduce mismatch regarding language and recording differences between training and test data. Lexicon enhancement and acoustic model adaptation improved recognition accuracy providing a reliable phoneme and word segmentation framework.

1 Introduction

Automatic pre-segmentation of phonemes is often employed in basic phonetic research since it helps to save time and effort compared to manual segmentation. Automatic speech recognition (ASR) systems are extensively used for initial phoneme and word segmentation of non-transcribed speech. Successful and reliable speech segmentation depends on the similarity between training and test speech data as well as on the diversity of the used lexicon. However, speech recognizers are mostly trained on native speech, so that their performance is worse on non-native data due to the mismatch between native language training data and non-native test data. Therefore, compensation of different pronunciation characteristics between native and non-native speakers is crucial for improving the ASR performance of non-native speech.

The effects of non-native speech recognition have been much investigated in the literature. E.g. it has been found that German-accented English words are recognized with significantly higher error rates [1]. To reduce the error, English-German phoneme mapping and parallel usage of two acoustic models have been employed. In [2], the recognition performance of an English ASR system used by German speakers has been assessed. Word error rate for the German speakers was 49.3% compared to 16.2% for the native English speakers. Furthermore, in [3] a trained German ASR system provided word recognition rate of 18.5% for German speakers and 34.0% for English speakers.

The straightforward approach to solve the problem is to create a system exclusively on speech data spoken by non-native speakers. However, that is not always possible because of the lack of large amount of non-native training data. Hence, non-native speech recognition can be improved using an existing native ASR system and adapting separately or in combination (hybrid approach) the pronunciation dictionary, the acoustic and the language model [4]. The pronunciation dictionary can be optimized by adding pronunciation variants that could be found in non-native speech. These variants can be created with knowledge-based or data-driven methods [3]. The disadvantage of this approach is that the inclusion of too many pronunciation variants, since the speakers' accent is not known in advance, increases word confusion. Traditional acoustic model adaptation techniques like maximum a posteriori (MAP) [5] or maximum likelihood linear regression (MLLR) [6] can be applied on speaker-independent models to adjust them to the characteristics of a non-native accent. Finally, the language modeling approach targets the grammatical effects of the non-native speaking style [7] and the hybrid approach combines the fore-mentioned methods for further improvement of ASR performance [8].

In the present study, an existing German ASR system was used to perform phoneme segmentation of German-accented English speech. Phoneme models were created from German training data, and the lexicon consisted of English words with German phoneme transcription. In order to reduce this mismatch, lexicon and phoneme model enhancement must be performed. The lexicon was enhanced by means of pronunciation rules selected according to the literature on German-accented English and by phoneme recognition results analysis. Furthermore, acoustic model adaptation was performed to decrease the mismatch between the training and test data regarding language and recording differences.

2 Lexicon enhancement

2.1 Phoneme mapping

For non-native speech acoustic modeling, phoneme similarities across languages have to be investigated in order to create phone mapping tables and consequently lexicons with adapted pronunciations.

Two basic methods for phoneme mapping are commonly used: a) *knowledge-based*, which uses a priori knowledge, e.g. from phonetic studies or dictionaries, to create a mapping between the native and the non-native phoneme system, and b) *data-driven analysis,* which derives the phoneme mapping from a database and can be an option when training data in the non-native language is available [9].

2.1.1 Knowledge-based phoneme mapping

As a first step, an English to German phoneme mapping was determined by human experts, choosing the most appropriate phoneme in the German inventory to match each English phoneme. This context-free phoneme mapping however, introduces uncertainty regarding the best correspondent phoneme to the speakers' actual productions. Besides, there can be significant size differences between the source and the target phonemic inventory [10] and therefore many-to-one or one-to-many matching in the case of phonemes without their corresponding peer. For example, the affricate (i.e. a plosive followed by its homorganic fricative) /dZ/ is rarely found in German but very frequent in English, which renders proper acoustic modeling difficult. Thus, this phoneme sequence could be mapped to /d/+/z/, /tS/, /S/ and /ts/ [1, 11, 12] (phoneme symbols in this paper are given in SAMPA [13]).

Table 1 shows the phoneme mapping at the base of the lexicon used for the present study. The pronunciation of English words was realized by means of the German phoneme inventory with one pronunciation variant per word. This lexicon was employed to force align non-native speech to the transcriptions using trained German phoneme models.

In German, glottal stops are often produced in front of word-initial vowels [14]. In the lexicon of German-accented English glottal stops were represented in front of all word-initial vowels, despite the fact that they might not occur.

Table 1. English-German phoneme mapping table used for the initial lexicon

Symbol	Word (e.g.)	English transcription [12]	German mapping
/dZ/	gin	dZIn	/d/+/S/
/T/	thin	TIn	/s/
/D/	this	DIs	/s/
/Z/	measure	meZ@	/z/
/w/	wasp	wQsp	/v/
/e/	pet	pet	/E/
/{/	pat	p{t	/E/
/Q/	pot	pQt	/O/
/V/	cut	kVt	/a/
/eI/	raise	reIz	/E/+/I/
/@U/	nose	n@Uz	/o/+/U/
/3:/	furs	f3:z	/9/+/6/
/A:/	stars	stA:z	/a:/
/O:/	cause	kO:z	/O/
/I@/	fears	fI@z	/i/+/@/
/e@/	stairs	ste@z	/E:/+/6/
/U@/	cures	kjU@z	/u:/+/6/

2.1.2 Data-driven analysis

After non-native speech had been automatically aligned and labeled, a supervised method was used to pinpoint problematic mapping entries. The confusion matrix between native (English) and the hypothesized non-native (German) phoneme models represents the likelihood of the confusion between two phonemes and therefore was used for phoneme similarities evaluation. For the data-driven analysis, a confusion matrix with normalized summed frequency values of the hypothesized phoneme occurrences was employed.

2.2 Pronunciation modeling

The presence of only one pronunciation per word in the lexicon is not appropriate since there can be several non-native pronunciation variations. ASR systems use lexicons to estimate expected phonemes of particular words in recognized utterances, however, there can be inter- and intraspeaker pronunciation variations. This is especially emphasized in the case on non-native speech since non-natives usually transfer their native language pronunciation habits to their second language productions. Therefore, the initial lexicon had to be enhanced by means of pronunciation rules for German-accented English. The rules were derived from linguistic knowledge of pronunciation variations of German-accented English [15-16] (knowledge-based approach), as well as from the analysis of the ASR confusion matrix (data-driven approach).

Table 2. Lexicon enhancement with pronunciation rules for German-accented English

SAMPA symbol	Position	Rule
/b/	word-final	/b/ or /p/
/d/	word-final	/d/ or /t/
/g/	word-final	/g/ or /k/
/s/	word-final	/s/ or /z/
/v/	word-final	/v/ or /f/
[?] (glottal stop)	word-initial	optional [?] before vowels

The rules applied refer to word-final obstruent devoicing and glottal stop insertion before word-initial vowels (s. Table 2), which are characteristics of German and likely to be transferred to German-accented English.

The pronunciation rules were implemented by adding alternative phonemes in the specified contexts. In this way, during the forced alignment labeling process, the phonemes with the highest confidence are chosen among the available alternatives. Pronunciation rules should be carefully chosen to optimize recognition accuracy: on one side few rules cannot foresee all aspects of non-native pronunciation variations, on the other side too many rules can increase the confusion in the lexicon and reduce recognition performance.

3 Experimental setup

3.1 Speech database

The speech materials were English BBC news bulletins read by 4 male and 3 female German native speakers, producing a speech database of 3 hours and 13 minutes duration (418 recorded prompts). The corpus is intended to be used for basic phonetic research of German-accented English speech and it consists of total 3094 words divided in 79 recording prompts.

The speech was studio recorded with 44.1 KHz PCM quality and later downsampled to 16 KHz and 16 bit resolution. The speech database was randomly divided into test and adaptation set for the acoustic model adaptation procedure. The adaptation set was large enough (10% of the database) to adapt to recording conditions and language mismatch, but small enough to avoid speaker dependency.

3.2 Baseline ASR system

The UASR (Unified Automatic Speech Recognition and Synthesis) system [17] was used for speech recognition. The system uses arc-emission HMMs with one single Gaussian density per arc and an arbitrary topology. The Gaussians are assigned to arcs and create one incoming and one loop arc for every Gaussian at the state for that distribution. The mixture weights are represented through different transition probabilities for each Gaussian distribution. The structure is built iteratively during the training process by state splitting from an initial HMM model.

For training and baseline performance evaluation, a subset of the German Verbmobil Database was used: 32337 of total 41512 turns (1022 speakers, 96 hours) [18]. The acoustic model structure consists of 42 monophonic HMMs plus one pause and one garbage model derived after the third HMM states split.

3.3 Model adaptation

Phoneme mapping and lexicon enhancement can provide appropriate modeling of simple phoneme substitutions and acceptable segmentation even without acoustic model modification. However, non-native speakers tend to pronounce sounds differently than native speakers for several reasons, such as distinct sound characteristics and phonotactics (i.e. permissible combinations of phonemes) in the native and the non-native language.

In order to compensate for the differences in language and recording conditions, and to further improve the recognition accuracy, the conventional adaptation method maximum a posteriori (MAP) [5] was used on the baseline acoustic model. Adaptation is the natural choice over model re-training due to the limited amount of recorded speech data.

3.4 Automatic generation of transcriptions

Phoneme segmentation and labeling were performed automatically by means of the German speech recognizer in the UASR system [17]. Viterbi forced alignment was successfully performed and a phoneme aligned labeled data set was produced. Figure 1 presents a labeled speech sequence showing that a reasonable accuracy for pre-segmentation can be achieved even with non-adapted acoustic models.

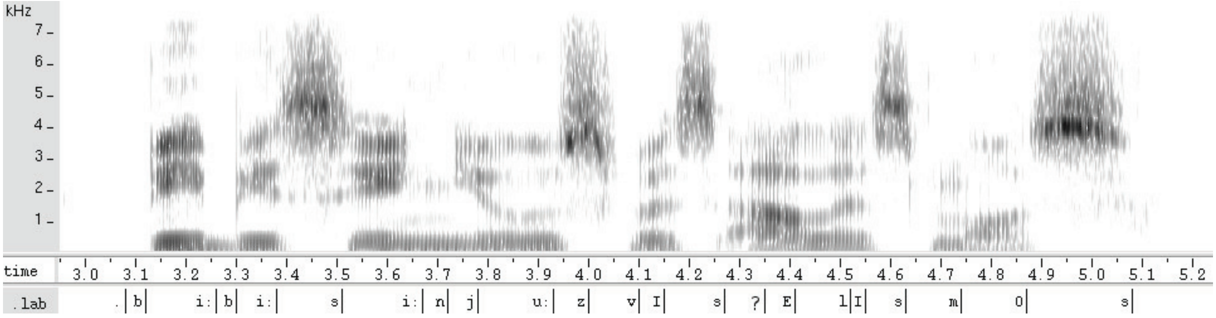

Figure 1. Automatic labeling of speech sequence "BBC news with Alice Moss" by a female German speaker.

4 Results and discussion

Using the UASR system with orthographic transcriptions, a lexicon of English words represented with German phonemes according to an English to German phoneme mapping (s. Table 1) and German acoustic models, the speech recordings were accordingly transcribed and labeled by Viterbi forced alignment.

Recognition results are presented in Table 3, where the parameters accuracy of the recognized label (phoneme) sequence (LSA), label (LSC) and frame correctness (FSC), and the lattice density (Latt) are calculated over the number of all phonemes in the reference sequence N^{all}, of deleted phonemes N^{del}, substituted phonemes N^{sub} and over the number of inserted phonemes N^{ins}, with sequence alignment:

$$LSC(\%) = (N^{all} - N^{del} - N^{\mathrm{sub}}) / N^{all} \cdot 100 \tag{1}$$

$$LSA(\%) = (N^{all} - N^{del} - N^{ins} - N^{\mathrm{sub}}) / N^{all} \cdot 100 \tag{2}$$

$$Latt = (N^{all} - N^{del} + N^{ins}) / N^{all} \tag{3}$$

First, the recorded speech was labeled using the lexicon created by means of the phoneme mapping table (Lex. 1), the labeled sentences were recognized and the observed accuracy for phoneme recognition (LSA - Label Sequence Accuracy: 34.0 ± 1.2%) was, as expected, lower than the system's baseline accuracy (47.7 ± 1.0%), due to speaker, recordings, as well as language mismatch between the training and test data. An empirical analysis of randomly selected labeled speech sequences showed, as predicted, that a lexicon, which does not contain pronunciation variations, is not appropriate for modeling German-accented English.

Table 3. Results of the speech recognition experiments

	FSC (%)	±	LSC (%)	±	LSA (%)	±	Latt.	±
Baseline	54.90	0.70	60.20	0.70	47.70	1.00	1.023	0.011
Lex. 1	58.70	1.30	49.50	0.90	34.00	1.20	1.033	0.007
Lex. 2	61.00	1.30	52.80	0.90	36.80	1.10	1.048	0.007
Lex. 2+Adp.	70.00	1.10	63.40	1.00	52.50	1.20	1.006	0.007

Some notable phoneme confusions (e.g. /z/ with /s/, /v/ with /f/, /d/ with /t/, /g/ with /k/) after the speech recognition experiment with Lexicon 1 can be seen from Figure 2 (representing the confusion for phonemes whose occurrence is more than 0.1% over the total count). The phoneme mapping is generally appropriate and accurate since there are almost no phonemes that are regularly confused with others.

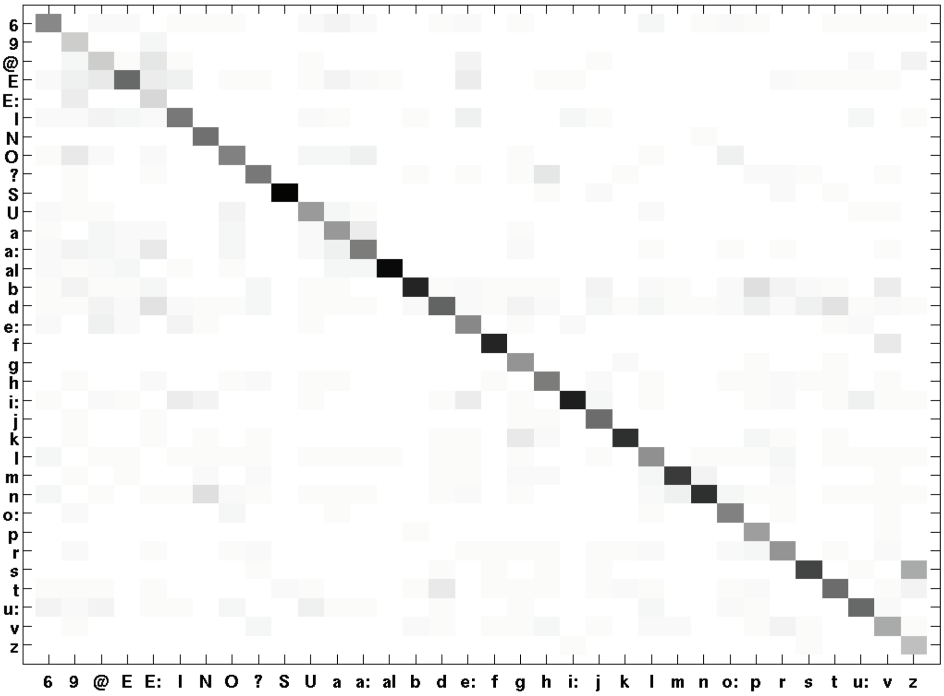

Figure 2. Confusion matrix derived from speech recognition with Lexicon 1.

Lexicon 2 was created by enhancing Lexicon 1 by means of pronunciation rules derived from German-accented English pronunciation variations (s. Section 2.2 and Table 2). During forced alignment, phonemes with the highest confidence were chosen among the available alternatives. This enhanced version of the lexicon was used in the next forced alignment iteration and the phoneme accuracy improved to 36.8 ± 1.1% (s. Table 3, Lex. 2). An improvement in the phoneme confusions scores was observed in terms of reduction of the mismatch frequency and the segmentation obtained was sufficiently reliable.

In order to further improve recognition by compensating the differences in language and recording conditions, a supervised MAP adaptation procedure (Lex. 2 + Adp.) was performed using the known transcriptions and the enhanced lexicon. The combination of adaptation on all three levels, i.e. phoneme, lexicon and acoustic models, further improved the phoneme accuracy (52.5 ± 1.2 %, higher than the baseline). Phoneme confusions further improved, except for voiced vs. unvoiced plosives, which is a common problem in speech recognition due to their spectral characteristics.

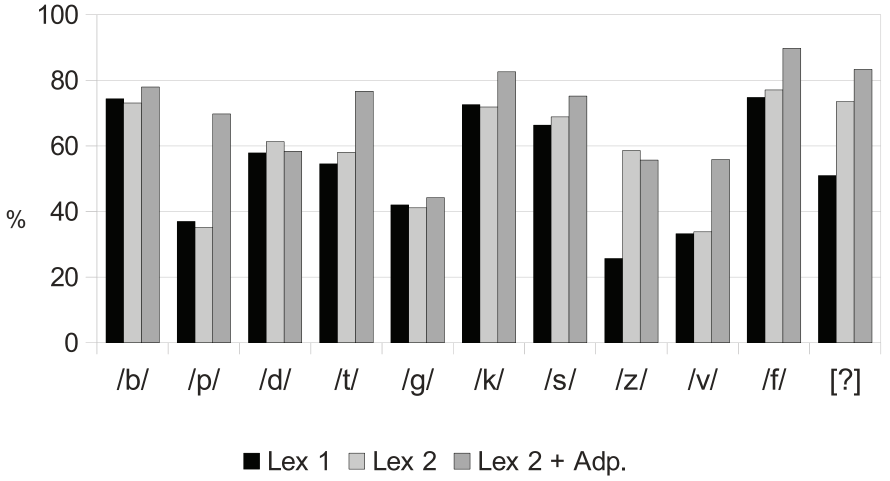

Figure 3. Correct recognition for phonemes affected by the pronunciation rules.

Figure 3 shows the correct recognition according to the confusion matrix for the phonemes affected by the pronunciation rules and for the glottal stop [?]. The optional glottal stop in front of word-initial vowels obviously improved its recognition from 51% to 73%, adaptation brought additional improvement to 83%. Recognition of /z/ greatly improved after lexicon enhancement from 26% to 56%. In other cases we observed slight or no improvement. For /p/, /t/, /k/, /v/ and /f/ remarkable improvement in the recognition was brought about by adaptation. In total, the recognition of the phonemes affected by the rules and of the glottal stop improved from 55% to 62% after lexicon enhancement and to 74% after adaptation.

5 Conclusions

In order to perform phoneme segmentation of German-accented English speech, an ASR system was used with German phoneme models. The lexicon consisted of English words whose pronunciations were represented by means of German phonemes. As expected, due to language mismatch, the first lexicon with no pronunciation variation produced lower phoneme recognition accuracy than the baseline system. Lexicon enhancement by means of pronunciation rules introducing optional glottal stops before word-initial vowels and word-final obstruent devoicing improved phoneme accuracy and phoneme confusion scores. Finally, acoustic model adaptation improved recognition accuracy providing a reliable phoneme and word segmentation framework. Ideally, for non-native speech recognition an ASR system with acoustic models trained on speech data in that specific language and spoken by that specific group of non-natives should be used. However, such a system is seldom available, since ASR are normally trained on native speech data, and it would be necessary to create an ASR system for each language pair. To obtain a reliable phoneme pre-segmentation for phonetic analysis of German-accented English, a German ASR system could be successfully used after lexicon enhancement and acoustic model adaptation.

Literature

[1] Ochs, S., Wölfel, M., Stüker, S., 2008. Verbesserung der automatischen Transkription von englischen Wörtern in deutschen Vorlesungen. In: Proc. of ESSV, Frankfurt, pp. 157-164.

[2] Wang, Z., Schultz, T, Waibel, A., 2003. Comparison of acoustic model adaptation techniques on non-native speech. In: Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Hong Kong, China, pp. 540-543.

[3] Steidl, S., Stemmer, G., Hacker, C., Nöth, E., 2004. Adaptation in the pronunciation space for non-native speech recognition. In: Proc. of Interspeech 2004 ICSLP, Jeju Island, Korea, pp. 318-321.

[4] Tan, Tien-Ping, Besacier, Laurent, 2008. Improving pronunciation modeling for non-native speech recognition. In: Proc. of Interspeech 2008, Brisbane, pp. 1801-1804.

[5] Gauvain, J., Lee, C., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains, IEEE Transactions on Speech and Audio Processing 2(2), 291-298.

[6] Legetter, C. J., Woodland, P. C., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer, Speech and Language 9, 171-185.

[7] Bellegarda, J., 2001. An overview of statistical language model adaptation. Invited lecture. In: Proc. of ISCA Adaptation 2001, Sophia-Antipolis, France, pp. 165-174.

[8] Bouselmi, G., Fohr, D., Illina, I., 2007. Combined acoustic and pronunciation modelling for non-native speech recognition. In: Proc. of Interspeech, Antwerp, Belgium, pp. 1449-1452.

[9] Goronzy, S., Rapp, S., Kompe, R, 2004. Generating non-native pronunciation variants for lexicon adaptation. Speech Communication 42 (1), 109-123.

[10] Schultz, T., Waibel, A., 2001. Experiments on cross-language acoustic modeling. In: Proc. of Eurospeech 2001, Aalborg, Denmark, pp. 2721-2724.

[11] Samsudin, N., Lee, M., 2012. Building text-to-speech systems for resource poor languages. In: Proc. of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, pp. 3327-3334.

[12] http://www.phonetik.uni-muenchen.de/Bas/BasGermanPronunciation/ (02.04.2003)

[13] SAMPA - computer readable phonetic alphabet. Available from http://www.phon.ucl.ac.uk/home/sampa/home.htm (25.10.2005)

[14] Kohler, K. J., 1994. Glottal stops and glottalization in German. Data and theory of connected speech processes. Phonetica 51, 38-51.

[15] Biersack, S., 2002. Systematische Aussprachefehler deutscher Muttersprachler im Englischen – Eine phonetisch-phonologische Bestandsaufnahme. In: Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München (FIPKM) 39, pp. 37-130.

[16] Goronzy, S., Sahakyan, M., Wokurek, W., 2001. Is non-native pronunciation modelling necessary? In: Proc. of Eurospeech 2001, Aalborg, Denmark, pp. 309-312.

[17] Hoffmann, R., Eichner, M., Wolff, M., 2007. Analysis of verbal and nonverbal acoustic signals with the Dresden UASR system. In: A. Esposito et al. (eds.), Verbal and nonverbal communication behaviours. Lecture Notes in Artificial Intelligence, vol. 4775, Berlin, Heidelberg: Springer, pp. 200-218.

[18] Bub, T., Schwinn, J., 1996. VERBMOBIL: The evolution of a complex large speech-to-speech translation system. In: Proc. of Int. Conf. on Spoken Language Processing (ICSLP 96), vol. 4, Philadelphia, pp. 2371-2374.