Acoustic Modeling of Foreign Words in a German Speech Recognition System

Georg Stemmer, Elmar Nöth, Heinrich Niemann

University of Erlangen-Nürnberg, Chair for Pattern Recognition,
Erlangen, Germany
stemmer@informatik.uni-erlangen.de

Abstract

The paper deals with the development of acoustic models of foreign words for a German speech recognizer. The recognition quality of foreign words is crucial for the overall performance of a system in application fields like spoken dialogue systems, when foreign words occur as proper names. One of the main problems in the modeling of foreign words is the limitation of training data, which must contain samples of the non-native pronunciation of the foreign sounds. In order to obtain robust acoustic models, which are still precise enough, we compare several methods to map or to merge the models of phonemes, which are pronounced in a similar way by German speakers. We utilize an entropy-based distance measure between sets of phoneme models. The best approach yields a reduction of 16.5% word error rate, when compared to a baseline system.

1. Introduction

1.1. The Problem

The paper deals with the development of acoustic models of foreign words for a speech recognizer in a German dialogue system. Foreign words, especially English words, are used frequently by German speakers in colloquial speech. They occur also as proper nouns in certain application fields of spoken dialogue systems. Examples are names of companies that have to be recognized by stock information systems, and the names of airlines, hotels or cities in the domain of travel information and reservation. The dialogue system which is described in this paper gives information about movie showtimes. Usually, more than 50 percent of all movie titles in the lexicon contain foreign words, because foreign movie titles are not translated, or original and translated title can be used equivalently. Misrecognitions of the foreign words have negative impact on the overall performance of a dialogue system, because the proper names carry important information, which has to be understood correctly. The recognition quality of the foreign words depends mostly on their acoustic models, because for most proper nouns the language model does not deliver a lot of discriminating information. For example, in utterances like *"when can I see Star Wars"*, the number of movie titles with a similar language model score is very large.

1.2. Previous Work

The appropriate treatment of foreign words in speech recognition and synthesis in general has been described as a multidimensional problem in [1]. In this paper, we only consider the difficulties which arise when we try to integrate the foreign words in a speech recognizer following a standard approach, which is also used for non-foreign words. That means, we start with a phonetic transcription for each foreign word and train acoustic models based on this transcription. The problem in doing so is that usually the amount of data is very restricted: In order to keep the mismatch between training and test low, we need samples of the foreign language, pronounced by German speakers. To our knowledge, currently only a few databases of non-native speech are available. The limitation of training data relates our work to those approaches for multilingual speech recognition, which improve the robustness of acoustic models by utilizing similarities between languages. In [2] a multilingual phoneme set is developed. The work in [3] uses a set of several source languages to build a speech recognizer for a target language. In contrast to multilingual speech recognition, the acoustic models of foreign words have to integrate typical pronunciation errors of the German speaker.

1.3. Approach

We start with a phonetic transcription of each foreign word with the phonetic symbols of its original language. Because we do not have sufficient training data for robust models of all foreign speech units, we intend to share acoustic models of similar sounds across the two languages. Therefore we have to map foreign phonemes to the German ones or to merge German and foreign phonemes. This approach is justified for a subset of the foreign phonemes, because speakers tend to do the same, when they pronounce certain foreign phonemes [4]. We have to determine, which of the foreign phonemes need their own model and which of them fall together with German phonemes.

In the rest of the paper, we compare different approaches to the generation of an appropriate set of phoneme models, that enables us to model German and English words as robust and precise as possible.

2. Short Description of the System

The speech recognizer is part of the German spoken dialogue system FRÄNKI, which can be accessed via the public phone network (+49 9131 16287). FRÄNKI answers questions concerning movie showtimes and where a movie will be shown. Users can also ask for a certain movie theatre, city or a time interval and get a list of movies that fit to the request.

The speaker independent continuous speech recognizer is based on polyphones, which are represented by semi-continuous HMMs. The output pdfs of the HMMs are full-covariance Gaussian densities. For real time decoding, a dialogue–state dependent bigram language model is used.

3. Data

In our experiments, we used the data from three different sources: Acoustic models of German speech are trained on the EVAR set. It consists of 20678 utterances, which have been

recorded by phone with our conversational train timetable information system. A detailed description of this system can be found in [5]. Nearly all utterances are in German language. The total amount of data is 23 hours. 16767 utterances have been selected randomly for training and validation, the rest of 3911 utterances is available for testing.

As there was no other collection of English speech sounds pronounced by German speakers available, we decided to use DENGLISH, a small subset of the data, which has been collected in the Verbmobile project [6]. This database is just a compromise for our purpose, because it has been recorded with microphone and the recording scenario is based on human-human-interaction. In order to adapt the dataset to our application, we processed it with a bandpass filter. The bandpass filter is designed to model the acoustic characteristics of a phone recording. The data has a total length of 1.5 hours and consists of 609 utterances. 505 have been selected for training and validation, 104 utterances will be used for testing.

For evaluation, we use the dataset FRÄNKI. It consists of 842 requests for information on movies, read by 19 different German speakers (12 male, 7 female). None of the speakers is included in training data. Each utterance is in German language and contains at least one English word as part of a movie title. 2148 of the 5888 words in the data set are English. None of the movie titles is included in the training data.

4. Acoustic Modeling of English Words

4.1. Mapping English Phonemes to German Phonemes

In order to get baseline results, we evaluate the performance of a speech recognizer, which has been trained on the German speech samples of the dataset EVAR only. English phonemes in the transcription of the foreign words in the lexicon get mapped to German phonemes. The mapping is done according to the recommendations of an expert group [7]. We have to mention, that only a preliminary version of those guidelines was available when our evaluations were conducted. In general, the mapping makes use of the fact, that similar phonemes across languages share the same phonetic symbol in the inventory of the International Phonetic Association (IPA). Phonetic symbols of the English transcription, which do also exist for the German language, are not changed. For all other symbols, mapping rules are defined, which do also depend on the position in the word. For example, the original and converted transcription of the movie title *"Blues Brothers"*, using SAMPA (http://www.phon.ucl.ac.uk/home/sampa/home.htm) is

$$\text{blu:z brVD@z} \rightarrow \text{blu:s bras@s}$$

A major advantage of the mapping approach is that we don't need any samples of non-native English speech. Any speech recognizer for German can recognize foreign words without any retraining, and the phoneme models are always estimated robustly.

4.2. Language-Dependent Phoneme Models

Our next step is to evaluate a recognizer based on phoneme models, which have been trained only on data of a single language. For both languages a full set of phoneme models is created. The result is a multilingual speech recognizer. We have to evaluate, if the size of the training data set is sufficient for a robust estimation of all acoustic model parameters. Similarities between German sounds and the non-native pronunciation of English phonemes are not utilized.

4.3. Knowledge-Based Merging of Phoneme Models

A straightforward way to increase the robustness of the language-dependent phoneme models is to represent several phonemes by a single model. Phonemes which have a similar pronunciation by German speakers get merged. In analogy to the mapping rules described above, all phonemes which have the same IPA symbol share one acoustic model. Only those English phonemes, which do not have a representation in the symbol set for the German language, e.g. { (as in *p<u>a</u>t*) or V (as in *c<u>u</u>t*), get their own model. An important disadvantage of the approach is, that only phonemes are tied together, which have a similar pronunciation across native speakers. For our application, only similarities between the native pronunciation of German phonemes and the non-native pronunciation of English phonemes should guide the grouping of the phonemes.

4.4. A Distance Measure Between Sets of Phoneme Models

In order to evaluate the similarities between the pronunciation of the phonemes from different languages, we need a suitable distance measure. We can also utilize such a distance measure for a data driven merging or mapping of phoneme models, which will be described in the following section. The computed distance should help us to determine, if two phonemes can be represented by a single HMM oder if there is a need for a separation. In general, we compare two sets of phoneme models $\mathcal{L}_a, \mathcal{L}_b$:

$$\mathcal{L}_a = \{\lambda_1, ..\lambda_m\} \tag{1}$$

$$\mathcal{L}_b = \{\lambda'_1, ...\lambda'_n\} \tag{2}$$

The HMMs λ'_j in $\mathcal{L}_b$ result from some transformation of the models λ_i in $\mathcal{L}_a$, e.g. the merging of λ_k and λ_l. We have to decide, which one of $\mathcal{L}_a$ and $\mathcal{L}_b$ is better suited to represent the data.

For the computation of distances between acoustic models, measures based on estimations of entropy are well known from literature, e.g. [8]. In [2] such a measure is used to generate multilingual phoneme set by clustering of similar phonemes. In contrast to [2], we compute the distance not between single phonemes but between sets of phoneme models. In our approach, it is not necessary, to have a time alignment of the data on the phoneme level. In order to define the distance D between the sets $\mathcal{L}_a$ and $\mathcal{L}_b$, we start with a phonetic transcription $L = l_1, l_2, .., l_s$, which gives us a sequence of phoneme symbols l_i for the data. Each symbol l is represented by an HMM λ_l or λ'_l. The distance between $\mathcal{L}_a$ and $\mathcal{L}_b$ is then

$$D(\mathcal{L}_a, \mathcal{L}_b) = \frac{1}{s}[\log P(\mathbf{X}|\lambda_{l_1}, .., \lambda_{l_s}) - \log P(\mathbf{X}|\lambda'_{l_1}, .., \lambda'_{l_s})] \tag{3}$$

$\mathbf{X}$ stands for the acoustic data. In order to use this measure to compute a distance $d(\lambda_e, \lambda_g)$ between a German and an English phoneme model, we first train English and German phoneme models. $\mathcal{L}_a$ is the set of English phoneme models, $\mathcal{L}_b$ results from $\mathcal{L}_a$, when we replace the single English phoneme model λ_e by the German model λ_g. Setting

$$d(\lambda_e, \lambda_g) = D(\mathcal{L}_a, \mathcal{L}_b) \tag{4}$$

gives us an estimation of the resulting loss in information. For small values of $d(\lambda_e, \lambda_g)$ we can expect, that the represented phonemes are pronounced very similar and that a substitution will not degrade the performance of the recognizer.

Table 1 shows the best matching German models and the computed distances for some English phonemes. The values have been computed using a part of the DENGLISH training sample, which contains 8734 words and 27113 phonemes. As one can see, the results from the distance computation fit to the expectation, for example, the English S is mapped to the German S, the same is true for the fricative T, which is pronounced as a ts by German speakers. In some cases, phonemes of the same phonetic class get confused, e.g. the plosive p has a very small distance to the phoneme combination tr, which also contains a plosive. Some individual mappings conflict with our expectations. For example, the third best matching sound for the English phoneme @U is the phoneme combination dr, this may be due to an unusual high frequency of a particular context of the phoneme in our data. For a few English phonemes, the distance is even negative, i.e. the German phoneme model gives higher score on the English data than the English model.

Table 1: *Distance between models of English and German phonemes and phoneme combinations.*

English phoneme	best matching German models + distance					
p	tr	2.56	ts	2.70	t	2.70
d	t	12.64	z	13.22	ts	14.89
dZ	ts	0.49	tS	0.50	S	0.61
S	S	0.14	Z	0.15	tS	0.52
T	ts	0.82	z	2.06	s	2.11
D	b	10.77	dj	11.58	dr	11.83
N	N	2.60	y6	2.74	m	3.31
m	m	4.49	N	6.29	n	6.51
{	E6	2.96	e6	5.00	E	6.26
V	e6	7.51	6	7.55	E6	8.21
3:	6	2.53	96	2.95	Y6	2.95
@U	aU	4.70	a	8.99	dr	12.14

4.5. Data-Driven Merging of the Phoneme Sets

One possible application of our distance measure, which has been described in the previous section, is to map English phonemes to the German phoneme models which have the lowest distance. This approach should help to replace only English phoneme models, for which a similar German model does exist. English phonemes, which cannot be represented by a German model are left unchanged. We can hope to achieve an optimal compromise between robustness and precision of the models. The speech recognizer is generated by the following steps:

1. train a speech recognizer with models for all English and German phonemes
2. for each English model λ_e and German model λ_g compute $d(\lambda_e, \lambda_g)$
3. predetermine a threshold Θ
4. replace all English models λ_e with $d(\lambda_e, \lambda_g) < \Theta$ by the best matching German model λ_g
5. the rest of the English phoneme models are not changed

We evaluate the results for several different values of Θ.

5. Experimental Results

5.1. Configuration of the Recognizer

For evaluation on the test set FRÄNKI, which represents the application data, we used the realtime configuration of our recognizer. The bigram model is a mixture of the dialogue-state dependent language models which are normally used in the recognizer. It has a relative high perplexity of 57.7 on the sentences of the test set FRÄNKI. 312 different movie titles can be recognized. All movie titles are in one category of the language model, i.e. the language model score does not deliver any information that can be used during the decoding phase to discriminate between the movie titles. 192 (62%) of the movie titles contain English words. The lexicon of the recognizer contains 956 words, 200 (21%) are in English.

5.2. Evaluation

In this section, the performance of the different approaches to the integration of foreign words are evaluated.

Table 2 gives an overview on the word error rates of the different recognizers, computed on the test subsets of the German speech database EVAR and the collection of non-native English speech samples DENGLISH. For the experiments of Table 2 a 4-gram language model is used, which has been estimated on the training sentences. The recognizer for these evaluations has a lexicon of 3548 words, 908 (26%) of them are in English. Of course, the best results on the German speech data are achieved by the baseline recognizer, which contains only models for German sounds. The loss in recognition performance for the language-dependent phoneme models is 2.6% absolute, and for the knowledge-based merging of phoneme models it is 4.5%. The results on the DENGLISH data are much worse, this is mainly due to the very bad performance of the language model, it has a perplexity of 157.1 on the DENGLISH data. For comparison, its perplexity on the EVAR test data set is 19.9. The language-dependent phoneme models perform 2 percent points better on the DENGLISH data set than the merged phoneme models because of their higher precision.

Table 2: *Word error rates on the test subsets of* EVAR *and* DENGLISH.

experiment	EVAR	DENGLISH
baseline German recognizer	28.78	-
language-dependent models	31.39	79.75
knowledge-based merging	33.23	81.77

The word error rates on the application test set FRÄNKI are given in Table 3 and Figure 1. The best results are achieved with the knowledge-based merging of phoneme models. The improvement achieved by the knowledge-based merging approach is 16.5 % (absolute) when compared to the knowledge-based mapping. The language-dependent models are not as robust as the merged models: Even though they perform better on the DENGLISH data, they are significantly worse on the application data. As can be seen from Figure 1, the data-driven merging of the phoneme models across the languages does only improve word error rate, when the distance of the merged phonemes is negative or zero. When the threshold Θ is positive, the word error rate is heavily dependent its value. The best word error rate for the data-driven merging is 54.2%, which is still significantly higher (4.4%) than the knowledge-based merging approach.

Table 3: *Word error rates on the test set* FRÄNKI.

experiment	WER [%]
knowledge-based mapping	66.32
language-dependent models	59.71
knowledge-based merging	49.80

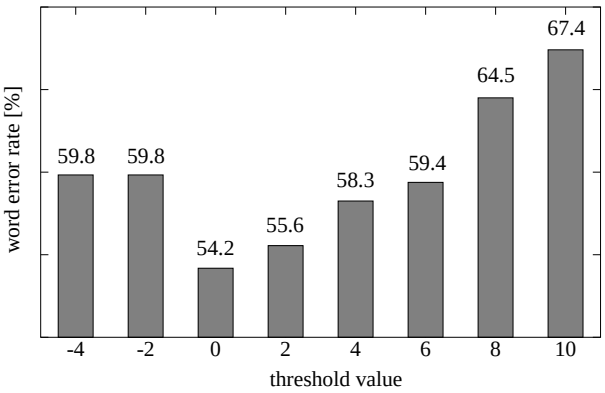

Figure 1: *Results for data-driven merging of phoneme sets on the test set* FRÄNKI *for different threshold values.* Θ *is varied between -4 and 10.*

5.3. Discussion

As can be seen from the results, the use of additional, non-native speech data for the modeling of foreign phonemes can greatly improve the recognition rates, even when the data set is small and the mismatch between training and testing data is relative high. It has been shown, that the lowest word error rates can be achieved with acoustic models which represent a compromise between precision and robustness.

It must be mentioned, however, that due to our application scenario, the English words in the FRÄNKI testing data are to a large part proper names, e.g. in movie titles like *"Being John Malkovich"*. It can be expected, that the mapping rules of the knowledge-based phoneme mapping perform much better on normal English words. The high portion of proper names in our data does also intensify the sensitivity of the robustness of the models, because we can assume, that a large number of different contexts occur only in the testing data and not in the training data. We do therefore expect, that the language-dependent phoneme modeling does perform better on testing data, which has a lower portion of proper names.

6. Conclusion and Outlook

Foreign words occur frequently in colloquial German speech. The good recognition of foreign words is crucial for some application fields of speech recognition. We compared several approaches to a suitable modeling of English words in a German speech recognizer. It is important to find the best compromise between robustness and precision of the acoustic models. Compared to the word error rate of a baseline system, which utilizes knowledge-based mapping rules in order to recognize the foreign words, our best recognizer achieves a reduction of 16.5% (absolute) word error rate. The approach is based on a knowledge-based merging of foreign and German phoneme models.

Further experiments will investigate into the modeling of variations in pronunciation. For some foreign languages, only a fraction of the speakers is able to find the correct pronunciation, others tend to experiment with different variants. For example, we observed at least three different pronunciation variants of the Italian word *"Cinecitta"*, which is the name of a movie theatre:

tSi:n@tSIta
zi:n@sIta
zIn@sIta

We also plan to evaluate, how additional speech data from native speakers can be utilized to increase the robustness of the phoneme models.

7. References

[1] R. Eklund and A. Lindström, "Pronunciation in an internationalized society: a multi-dimensional problem considered", FONETIK 96, Swedish Phonetics Conference, TMH-QPSR 2/1996, pp. 123-126, Nässlingen, 1996.

[2] J. Köhler, "Multi-lingual Phoneme Recognition Exploiting Acoustic-phonetic Similarities of Sounds", Proc. Int. Conf. on Spoken Language Processing, pp. 2195–2198, Philadelphia, 1996.

[3] P. Beyerlein and W. Byrne and J. Huerta and S. Khudanpur and B. Marthi and J. Morgan and N. Peterek and J. Picone and W. Wang, "Towards Language Independent Acoustic Modeling", IEEE Workshop on Automatic Speech Recognition and Understanding, Keystone, 1999.

[4] A. Lindström and R. Eklund, "How Foreign are "Foreign" Speech Sounds? Implications For Speech Recognition and Speech Synthesis", Proc. of the Workshop on Multi-Lingual Interoperability in Speech Technology, Leusden, 1999.

[5] F. Gallwitz and M. Aretoulaki and M. Boros and J. Haas and S. Harbeck and R. Huber and H. Niemann and E. Nöth, "The Erlangen Spoken Dialogue System EVAR: A State–of–the–Art Information Retrieval System", Proc. of 1998 International Symposium on Spoken Dialogue, pp. 19–26, Sydney, 1998.

[6] W. Wahlster, "Verbmobil: Foundations of Speech-to-Speech Translation", Springer, New York, Berlin, 2000.

[7] A. Batliner and B. Möbius and A. Schweitzer and S. Goronzy and P. Regel-Brietzmann, "Guidelines for 'Text-to-Phone' (TTP) conversion", Smartkom Technical Document, to appear.

[8] V. Digalakis and A. Sankar and F. Beaufays, "Training Data Clustering For Improved Speech Recognition", Proc. European Conf. on Speech Communication and Technology, pp. 503–506, Madrid, 1995.