RECOGNITION OF NON-NATIVE GERMAN SPEECH WITH MULTILINGUAL RECOGNIZERS

Ulla Uebler, Manuela Boros

Bavarian Research Center for Knowledge Based Systems (FORWISS)
Research Group for Knowledge Processing
Am Weichselgarten 7
D-91058 Erlangen, Germany
e-mail: {uebler,boros}@forwiss.de

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

10.21437/Eurospeech.1999-222

ABSTRACT

In this study we present different approaches to the recognition of non-natives. With a corpus in German spoken by speakers with 56 different first languages, the *Strange Corpus*, we perform recognition experiments with monolingual and multilingual recognizers. Among other, we compared two German recognizers, one that was trained in addition with non-native (Italian) speech and the other trained with German speakers only. We found that best performance is achieved with the recognizer trained with German including non-native speech, followed by a bilingual recognizer and an Italian recognizer trained with German and Italian natives.

1. INTRODUCTION

Over the years, speech recognition has become more robust for applications in noisy environments and other issues like speaker independency and spontaneous speech. One problem, however, is to understand speakers that vary from the speakers the system is trained with. This difference may be regional dialects where people training a system have the same dialect, but someone else coming from another region still has to be understood reliably [4]. This can also be variations due to an accent of a speaker, when someone speaks to the recognizer in his second language. This issue may be important for information systems used by foreign travelers or in bilingual regions like South Tyrol in Italy where both Italian and German are spoken by the inhabitants but only one of them as first language [9].

For this study we have chosen a corpus with native speakers of different countries with some knowledge ofGerman who all speak one German phrase. We compare the performance of several recognizers that were either trained with German or the native language of a speaker as well as bi- and trilingual recognizers. For the German recognizers, we have data with German natives and non-natives.

In order to reduce confusion in the naming of languages, we will speak of the language to be recognized as the *goal language*, the first or native language is that one the speaker learnt first in childhood. With non-native speech, we mean speech uttered in a language that is different from the first language of the speaker.

In the following, we will shortly describe the corpus we used for these experiments. In Section 3 we will describe the system architecture we used for our approach and describe the data that were available for the training of mono- and multilingual recognizers. Then, the experiments and results will be given in the following section, followed by a conclusion.

2. THE STRANGE CORPUS

The Strange Corpus was recorded at the Bavarian Archive For Speech Signals (BAS) of the Institute for Phonetics at the University of Munich, Germany [7]. It contains a sentence spoken by people with 56 first languages from all over the world entitled *Nordwind und Sonne* (North Wind and Sun) with 125 words, thereof 72 distinct words:

> Nordwind und Sonne. Einst stritten sich Nordwind und Sonne, wer von ihnen beiden wohl der Stärkere wäre, als ein Wanderer, der in einen warmen Mantel gehüllt war, des Weges daherkam. Sie wurden einig, daß derjenige für den Stärkeren gelten sollte, der den Wanderer zwingen würde, seinen Mantel abzunehmen. Der Nordwind blies mit aller Macht, aber je mehr er blies, desto fester hüllte sich der Wanderer in seinen Mantel ein. Endlich gab der Nordwind den Kampf auf. Nun erwärmte die Sonne die Luft mit ihren freundlichen Strahlen, und schon nach wenigen Augenblicken zog der Wanderer seinen Mantel aus. Da mußte der Nordwind zugeben, daß die Sonne von ihnen beiden der Stärkere war.

The words of this sentence are modeled with 61 German phonemes. The corpus is spoken by 88 speakers, including 16 German natives. The sentences are recorded with a Sennheiser microphone with 16 kHz sampling rate.

3. RECOGNITION SYSTEMS

For these experiments, we use the ISADORA recognition system that has already been employed among others for a train information system [3] and within a data-entry system [1].

3.1. System Architecture

The ISADORA system processes feature extraction of 12 cepstral features, with their first derivates, resulting in 24 feature at each frame. We use semi-continuous HMMs, and no language model for recognition.

The vocabulary of the Strange Corpus is inserted to the recognizers with their pronunciation in monophones. For the German recognizers, also most words are unknown, since this text has a vocabulary that does usually not occur in the applications like appointment scheduling or in the land register domain.

3.2. Multilingual Recognizers

As acoustic units we chose to use the 61 phonemes as acoustic units, since the use of e. g. polyphones would make the approach more complicated for experiments with other languages, which do not have the same phones, and even less the same polyphones. Also, the substitution process would get more complicated, if not only a certain percentage of these 61 phonemes would have to be replaced but also the up to 2000 polyphones.

When multilingual recognizers are trained with data from different languages, many phones like /a/ appear in most languages. A decision must be made, if e. g. the phones $/a/_{lg1}$ and $/a/_{lg2}$ of two languages are assumed identical and thus, in the training process, the occurrence of each of these phones trains the same multilingual phone or if they are assumed as different phones, and thus the number of phones of a multilingual recognizer is up to n times higher when training n languages. Another problem when using distinct acoustic units for each language, is to decide which of the e. g. $/a/_{lg1}$ to $/a/_{lgN}$ will be used for the recognition in the Strange Corpus.

We use both approaches, but for the evaluation of this approach, we will only give results for the first approach with common acoustic units, i. e. if they have the same IPA or SAMPA notation.

Another problem when recognizing one language with a recognizer that was trained with other languages is that some phones of the goal language (here German) have not been trained at all since these phones do not occur in one of the training languages. We have training material in German, Italian, Slovak, Slovenian, Czech, English and Japanese, and out of these languages, the phone /y/ does only occur in German. Thus, for recognition the needed phone must in order to be able to recognize the most similar must be substituted by the most similar trained phone, or some interpolation between two similar phones has to be done.

There are mainly three ways to estimate the most similar phone, which have been applied in different approaches to multilingual speech recognition, e. g. [2, 6, 8]:

1. Substitute phones na(t)ively, depending on how non-natives performs the substitution of a for him *unpronounceable* phone. This approach needs to have data from non-natives and some experience in the classification, i. e. which phonetic differences are systematic.
2. Substitute or cluster phones according to their phonological characteristics like *voiced plosive*. Substitution uses the most similar phone instead, whereas clustering takes several different phones and interpolates the parameters in some way, e. g. a mid vowel could be interpolated by the front and back vowel if both occur in the training language. The problem here is to find out which of the classification criteria (e. g. for consonants), like manner or place of production is more important, and which of the criteria should be in accordance.
3. Data-driven according to a similarity measure in the acoustic space, in the codebook etc. Here, without any regard to the process of the production of the phone, the most similar phone is chosen and substituted. In many cases, the chosen phone corresponds to the phonologically chosen one, but in some cases, others are chosen [2]. The problem is that at least some data of the new language has to be available in order to find out which phones are the most similar ones.

Here we use the na(t)ive approach for the trained languages, since we have speaking examples of these speakers, and in general a certain knowledge on non-native accent in our first language.

4. EXPERIMENTAL RESULTS

As already mentioned, we have training data of seven languages: for Slovak (Sa), Slovenian (Se) and Czech (Cz) from an information requiry system for train/flight connection (SQEL, [5]). For English (En), Japanese (Jp) and German (G2), we have data from the VERBMOBIL appointment scheduling task. Both applications contain spontaneous speech. For Italian (It) and German (G1), we have data from the SPEEDATA project which developed a natural speech data-entry system for Italian and German in the land register domain. The data are spoken in Italian or German with natives and dialect speakers of both languages. Thus, half of the German language data are spoken by Italian natives with a certain degree of accent. In the following, we will call this German Corpus G1, while we call the German data from VERBMOBIL G2. The amount of acoustic data (in hours) can be seen in Table 1.

Language	G1	It	Sa	Se
Data/hours	8.6	7.6	5.1	6.1
	Cz	Jp	En	G2
	7.2	27.4	9.6	28.5

Table 1. Acoustic data for each language

We have trained 8 monolingual recognizers for all 8 corpora (7 languages). Furthermore we have trained an Italian-German-G1 bilingual recognizer (which showed best performance for bilingual recognition), as well as one for Slovak and Slovenian. Furthermore, we have trained a trilingual recognizer for the Slavic languages (Sa-Se-Cz). Finally, we have trained a bilingual recognizer for English and German-G2.

Table 2 shows the word accuracy without language model for these recognizers. The second column shows the average word accuracy for the complete corpus of native and non-native speakers, while column three and four show the performance evaluated on the native language of the speaker.

Best performance is achieved for the G1 German recognizer with 60 % word accuracy. The other German

Recognizer	total	native	non-native
G1	60.6	62.3	60.2
Italian	40.1	37.0	40.8
Slovak	16.5	19.7	15.8
Slovenian	12.7	19.7	11.2
Czech	18.7	20.7	18.3
English	13.4	20.9	11.8
Japanese	18.1	18.7	18.0
G2	56.6	70.3	53.7
G1 + Italian	59.9	62.3	59.4
Slovak + Slovenian	20.0	24.6	19.0
G2 + English	45.0	57.2	42.4
Sa + Se + Cz	24.0	28.8	23.0

Table 2. Word accuracy for the Strange Corpus with different recognizers

recognizer (G2) achieves a performance of only 56 %. Although there is three times more data available for the German G2 language, recognition is worse. One explication for the difference in performance can be seen in the speakers of the training data, i. e. that the G1 speakers are non-native, and therefore the acoustic units are modeled less sharp, but more robust for the recognition of non-natives.

The bilingual Italian-G1 recognizer has the second best performance and even outperforms the G2 recognizer. Previously, we also found out that this bilingual recognizer performs better than the G1 recognizer for its application task, thus we can see, that in this case there is no degradation caused by speech in another language and no bad influence on the acoustic units which are estimated also with Italian phones.

The Italian recognizer shows with 40 % word accuracy the best performance of the monolingual non-German recognizers with the others being in the range of 12 to 18 % word accuracy. This difference can be explained by the fact that for the training of the Italian recognizer German natives were involved. Another reason may be the number of phone substitutions: for the Italian recognizer, no substitutions were necessary since all German phones were represented due to the bilingual application, whereas for the other languages several phones had to be replaced.

These results show clearly the influence of the native/non-native training material on the performance of non-natives. It is best to have non-natives already included in the training data. If there are no non-natives included in the training material, it may also be helpful to have data of native Germans speaking another language, since some acoustic structures of German remain also when Germans speak a foreign language (which can be perceived by humans as the accent).

Looking at bilingual recognizers, we can see the same effect as for monolingual recognizers: performance is better, if the German language is involved in training. The performance decreases compared to the monolingual German recognizers, more for the English-G2 recognizer than for the Italian-G1 recognizer, since in the Italian-G1 recognizer information about the German language comes also from the Italian part of the data. For the recognizer without German training data, an improvement is observed compared to the corresponding monolingual recognizers from 12–16 % to almost 20 %. We assume that on the one hand the higher amount of training material causes the difference, but also the higher robustness of the acoustic units which are now estimated by different languages and therefore may converge towards multilingual acoustic units where the German phones are only one special representation.

Looking at trilingual recognizers, we can find more improvement for the recognizer where German is not involved in training from 20 % to 24 % for the trilingual Slavic recognizer. This supports the hypothesis that the more languages are involved the better recognition gets, although it is far from the performance from a recognizer trained with suitable (native+non-native German) data.

Comparing the performance of native vs. non-native speakers, we can find an astonishing gap for the German recognizers: while the performance for the G1 recognizer is almost the same for both speaker groups (62 vs. 60 %), the G2 recognizer has a large gap in the performance with 70 % for natives and 53 % for non-natives. Again, we can suppose that the acoustic units are modeled more precisely (also according to the larger amount of training data) for native Germans, but have problems in the recognition of non-natives.

Performance is better for natives besides for the Italian recognizer which was trained with German natives speaking Italian. We could guess that some kind of non-native characteristics may be learnt by the recognizer, or simply, that most of the non-natives in the Strange Corpus speak with a structure similar to that of Italian like a lower consonant/vowel ratio (also for Japanese with CV structured words, the difference between the speaker groups is low.

Among the speakers are also some Swiss people who speak one out of a set of regional dialects as their first language as well as standard German influenced by the respective dialects. In this corpus, they speak dialect influenced standard German. Swiss dialects belong to the group of Southern German dialects like the one spoken in the region of South Tyrol, and we can thus find a performance for the G1 recognizer of 77.0 % for the Swiss compared to 62 % for Germans in Germany. For native Italians, we also found a big difference for of 79 % vs. 60 % compared to non-Italian non-natives with the G1 recognizer. Differences in the same direction can also be found for the other recognizers, but to a much smaller extent.

When we look at language families, we can find a relatively better performance for Slavic natives when a recognizer trained with Slavic data is involved: while for most recognizers, the performance of Slavic speakers is around the average performance for non-natives, the performance for Slavic speakers almost doubles with Slavic recognizers compared to other non-natives. Still, the performance is some points worse for the trilingual Slavic recognizer than for the G1 recognizer with Slavic speakers.

We also evaluated the performance with respect to language families and found a performance for Ger-

manic languages slightly higher than average (61.9 vs. 60.2 %), with Romanic languages having an even higher performance of 64.8 % word accuracy for the G1 recognizer. The G2 recognizer, however has a recognition for Germanic languages of 65 % and 59 % for Romanic languages. Thus, we can see that our assumptions on languages might also be transferred to language families: if a Romanic language is included into training, the performance gets better for Romanic natives speaking German.
An evaluation on continents in order to find out if some geographical vicinity may have an influence on the acoustic structure of languages (like the influence on dialects at language borders) showed the best performance for Europe (with or without German natives), followed by South America and Asia for both German recognizers. Poor performance was achieved for African speakers. Instead of a similarity due to the vicinity the difference may be caused by the vicinity of the language families, since in all over Southern America Indoeuropean languages are spoken, while most of the African languages do not belong to Indoeuropean languages.

5. CONCLUSION

In this study on the recognition of non-native speech we were able to compare the performance of two recognizers trained with German speech. One had a part of the data spoken by non-natives, the other had purely natives and a three times higher amount of training data. The performance on natives is better with the approach with pure German data, but may also be due to the higher amount of training data. For non-natives, however, the performance is higher for the recognizer that was trained with non-natives. If non-natives shall be recognized reliably it is necessary to include non-native speech into the training material. The origin of the non-natives is not as important as their non-native speech, which possibly makes the acoustic units more robust than the pure German language acoustic units.
If there is no non-native speech available, it also improves performance to include non-native speech of people with the first language to be recognized, in this case German natives speaking another language with accent. This may probably help since there is a certain similarity to the goal language, but still some variation to make the parameters more robust in contrast to taking completely different training data like native speech in another language which might confuse the system's parameters.
Thirdly we found that, if the goal language is not involved in training, multilingual recognizers perform better than monolingual ones. This may be due to the higher amount of training material, but also a way of finding a robust but meaningful representation of phones.
For an evaluation on geographic vicinity of language and therefore an influence on the performance, we found no evidence.
Thus, for the understanding of non-natives, nothing is more important than having non-native speech of the goal language in the training material or at least non-native speech of native speakers of the goal language.

REFERENCES

[1] U. Ackermann, F. Brugnara, M. Federico, and H. Niemann. Application of Speech Technology in the Multilingual SpeeData project. In *3rd Crim-Forwiss Workshop*, Montréal, 1996.

[2] P. Dalsgaard, O. Andersen, and W. Barry. Cross-Language Merged Speech Units And Their Descriptive Phonetic Correlates. In *Proc. Int. Conf. on Spoken Language Processing*, volume 6, pages 2627–2630, Sydney, December 1998.

[3] W. Eckert, T. Kuhn, H. Niemann, S. Rieck, A. Scheuer, and E. Schukat-Talamazzini. A spoken dialogue system for German intercity train timetable inquiries. In *Proc. European Conf. on Speech Communication and Technology*, volume 2, pages 1871–1874, Berlin, September 1993.

[4] V. Fischer, Y. Gao, and E. Janke. Speaker-Independent Upfront Dialect Adaptation In A Large Vocabulary Continuous Speech Recognizer. In *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 787–790, Sydney, December 1998.

[5] S. Harbeck, E. Nöth, and H. Niemann. Multilingual Speech Recognition. In *Proc. of the 2nd SQEL Workshop on Multi-Lingual Information Retrieval Dialog s*, Plzeň, 1997.

[6] J. Köhler. Multi-lingual Phoneme Recognition Exploiting Acoustic-phonetic Similarities of Sounds. In *Proc. ICSLP'96*, Philadelphia, USA, 1996.

[7] F. Schiel. Speech and Speech-Related Resources at BAS. In *First International Conference on Language Resources and Evaluation*, Granada, Spain, 1998.

[8] T. Schultz and A. Waibel. Language Independent and Language Adaptive Large Vocabulary Speech Recognition. In *Proc. Int. Conf. on Spoken Language Processing*, volume 5, pages 1819–1822, Sydney, December 1998.

[9] U. Uebler, M. Schüßler, and H. Niemann. Bilingual and Dialectal Adaptation and Retraining. In *Proc. Int. Conf. on Spoken Language Processing*, volume 5, pages 1815–1818, Sydney, December 1998.