ELSEVIER

Speech Communication 35 (2001) 53–69

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Multilingual speech recognition in seven languages

Ulla Uebler [*]

Bavarian Research Center for Knowledge Based Systems (FORWISS), Research Group for Knowledge Processing, Am Weichselgarten 7, 91058 Erlangen, Germany

Abstract

In this study we present approaches to multilingual speech recognition. We first define different approaches, namely portation, cross-lingual and simultaneous multilingual speech recognition. We will show some experiments performed in the fields of multilingual speech recognition. In recent years we have ported our recognizer to other languages than German (Italian, Slovak, Slovenian, Czech, English, Japanese). We found that some languages achieve a higher recognition performance with comparable tasks, and are thus easier for automatic speech recognition than others. Furthermore, we present experiments which show the performance of cross-lingual speech recognition of an untrained language with a recognizer trained with other languages. The substitution of phones is important for cross-lingual and simultaneous multilingual recognition. We compared results in cross-lingual recognition for different baseline systems and found that the number of shared acoustic units is very important for the performance. With simultaneous multilingual recognition, performance usually decreases compared to monolingual recognition. In few cases, like in the case of non-native speech, however, the recognition can be improved. © 2001 Elsevier Science B.V. All rights reserved.

Zusammenfassung

Diese Untersuchung zeigt Ansätze zur multilingualen Sprachverarbeitung. Zunächst werden verschiedene Ansätze vorgestellt: Portierung, transsprachliche Erkennung über Sprachgrenzen hinweg und simultane multilinguale Sprachverarbeitung. In diesem Beitrag werden wir unsere Ergebnisse in den Feldern der multilingualen Sprachverarbeitung vorstellen. In den letzten Jahren haben wir Erkenner in andere Sprachen als deutsch portiert (italienisch, slovakisch, slowenisch, tschechisch, englisch und japanisch). Unsere Experimente zeigen, dass einige Sprachen, bei vergleichbarer Komplexität, bessere Erkennungsraten zeigen, also leichter zu erkennen sind als andere. Die Ersetzung der Laute der jeweiligen Sprachen ist ein wichtiger Punkt bei der translingualen und simultanen multilingualen Sprachverarbeitung. Wir vergleichen Ergebnisse bei der translingualen Erkennung für unterschiedliche Basissysteme und fanden, dass die Anzahl der gemeinsamen Laute der Sprachen ein ausschlaggebendes Kriterium für die Erkennungsleistung ist. Bei der simultanen multilingualen Sprachverarbeitung sinkt gewöhnlich die Erkennungsrate im Vergleich zu monolingualer Sprachverarbeitung. In wenigen Fällen, wie bei Nichtmuttersprachlern, kann sich die Erkennung allerdings auch verbessern. © 2001 Elsevier Science B.V. All rights reserved.

Résumé

Cet article décrit notre travail dans le domaine de la reconnaissance multilingue de parole. D'abord nous présenterons les différentes stratégies: portation, reconnaissance à travers plusieurs langues et la reconnaissance simultane des plusieurs langues. Puis nous présentons les résultats obtenus. Ces dernières années nous avons porté notre

[*] Tel.: +49-9131-691-258.
E-mail address: ulla.uebler@eed.ericsson.se (U. Uebler).

0167-6393/01/$ - see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S0167-6393(00)00095-9

système de reconnaissance en différentes langues (Italien, Slovaque, Slovène, Tchèque, Anglais et Japonais). Nos expériences montrent que certaines langues sont plus facile a reconnaître, pour la même complexité du domaine. La substitution des sons des langues inclues est de grand intérêt pour la reconnaissance à travers plusieurs langues et la reconnaissance simultane des plusieurs langues. Nous comparons les résultats pour les différents systèmes de base pour la reconnaissance à travers des langues. Nous avons trouvé que la nombre de sons communs est un critère principale pour la qualité de la reconnaissance. Pour la reconnaissance multilingue, la reconnaissance diminue en générale comparé à la reconnaissance monolingue. Dans peu de cas, par exemple, quand il est parlé dans une autre langue que la langue maternelle, la reconnaissance peut être améliorée.

Keywords: Dialect; Non-native; Bilingual; Multilingual

1. Introduction

Over the years we have studied speech recognition and speech understanding systems in German, and as more and more multilingual applications are needed, the ISADORA system (Kuhn, 1995; Schukat-Talamazzini, 1995) was also used for multilingual speech recognition (Ackermann et al., 1996b; Nöth et al., 1996).

The need for multilingual speech recognition applications has arisen, for example, due to growing internationalization like within the European Community or in telecommunications. Thus, applications are developed for recognition in a new language, for example dictation systems are ported to a new language or information systems are developed for tourist information at airports and train stations which have to be able to understand a couple of languages.

When developing a recognition system either exclusively for the new language or for the new language in addition to existing languages, the recognition system optimized for the first language has to be adapted to the characteristics of the new language.

During this process, mainly data like the vocabulary, acoustic parameters, language models, and the dialog structure have to be adapted. Most of these adaptations have already been performed before, e.g. when porting a system to a new domain. One topic is still specific to the portation to a new language: the definition and the use of acoustic units. If the recognizer is completely rebuilt for a new language with training material of that language, the definition of new acoustic units arises from the pronunciation of the words in the vocabulary, but when not sufficient training material is available for the new language or when two languages are recognized at the same time, the acoustic units of both languages have to be set in relation.

The variety of approaches makes a definition necessary. We define three approaches to multilingual speech recognition and we will describe our experiences in these three approaches depending on specific fields of applications. The languages used together in multilingual speech recognition must somehow share their acoustic units, and thus a way to relate the acoustic units must be found. This problem and solutions to it will be one central aspect in this contribution together with the description of cross-lingual and simultaneous multilingual recognition systems.

In the following, we will define different approaches to multilingual speech recognition depending on the goal of the application, porting, cross-lingual and simultaneous multilingual speech recognition. These different approaches depend on the goal set by the application, i.e. which and how many languages will be recognized at a time. Furthermore, the chosen approach depends on the available training data with respect to the spoken language, the speaker and the recording conditions. Details will be given for each of the three approaches. Then, we will shortly describe the data we use for our experiments in all three approaches to multilingual speech recognition. In Section 4 we will describe three strategies for phone substitution, namely na(t)ive, phonetic and data-driven. Sections 5–7 show experiments and results in the three approaches to multilingual speech recognition for the seven languages

(Italian, German, English, Slovak, Slovenian, Czech and Japanese).

2. Approaches to multilingual speech recognition

When looking at the approaches in multilingual speech recognition, we propose to cluster them into three groups: porting, cross-lingual recognition and simultaneous multilingual speech recognition. These approaches differ according to the available training data for the development of the recognizer and the application the recognition system is designed for.

2.1. Porting

The first approach to multilingual speech recognition described here, is *porting*. A speech recognition system designed for a language is ported to another language in order to be used in that language. The recognition system is the same for the new language, the training data are only of the new language. This is often done for dictation systems which will be sold also for the recognition of another language. Difficulties to cope with are the characteristics of the new language. For example, if a system is ported to German, it must deal with the compound words, or, for French, with homophones. A system that has been developed for one language is now optimized to be used in another language, and some of the algorithms have to be adapted to work well in that new language. The system for the new language is trained with data of the new language, and there is enough training data in the new language to completely establish a system in the new language.

As shown in Fig. 1, the systems of the old and the new language are separate. The algorithms and principles for the new language are taken from the recognizer of the first language, and only an adaptation of the algorithms takes place in order to achieve an optimized performance also in the new language. Table 1 shows some studies of porting for dictation and dialog systems.

Porting of a recognizer is done for most of the following applications in multilingual speech recognition. The approaches described next deal with additional challenges besides the characteristics of a new language. Difficulties to cope with can be

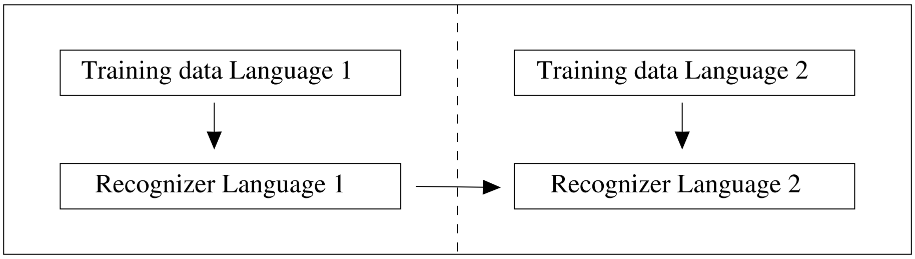

Fig. 1. Sketch of the porting scenario.

Table 1
Approaches to multilingual porting

Approach	Task	Languages
Cerf-Danon et al. (1991)	Dictation	American English, Italian, French, German, Spanish
Glass et al. (1995)	Dialog system	American English, Japanese, Italian
Lamel et al. (1995) and Young et al. (1997)	Dictation	American and British English, French, German
Barnett et al. (1996)	Dictation	American and British English, French, German, Italian, Spanish, Japanese, Mandarin
Fabian (1997)	Dialog system	German, Polish
Ipsic (1996), Klečková (1997) and Krokavec and Ivanecký (1997)	Dialog system	Czech, Slovak, Slovenian

insufficient training data or the recognition of several languages at the same time without knowing which language is actually spoken.

2.2. Cross-lingual recognition

This approach follows the same goal as the porting approach. The difference to the approach above is that insufficient training material is available to train the recognizer in the new language. Thus, in *cross-lingual* recognition, methods must be found to use training material of another language for a modeling of acoustic parameters. Optionally, an adaptation with few data from the goal language takes place.

First, the languages used for the training of the recognizer must be determined. The language(s) leading to the best recognition performance on the new language must be found. A relation between the languages used for training and the language to be recognized has to be selected. Fig. 2 shows the influence of the first language on the recognizer of the new language to be recognized. As in the porting scenario, the algorithms of the recognizer are adapted to the new language. Furthermore, training material of the first language is used for the parameter estimation of the recognition system for the second language.

For this approach, a relation between the training languages and the language to be recognized must be found in order to model the parameters of the language to be recognized using the parameters of the training language. One focus will be put on the acoustic units of the training and recognition languages.

One main problem is to determine identical acoustic units or to model existing acoustic units in a way that a good recognition can be provided. One can think of a couple of different measures to determine the similarity of phones across languages. If some adaptation material of the new language is available, the best adaptation algorithm, that is able to optimally use such data, must be found. Studies of cross-lingual multilingual speech recognition are shown in Table 2.

2.3. Simultaneous multilingual speech recognition

The third cluster of approaches is that of *simultaneous multilingual recognition*. Applications with this approach allow for a recognition of utterances of different languages at the same time. The system does not know in which language an utterance is spoken. A sketch of this approach can be seen in Fig. 3. Training data are available for each language. The result is a single recognizer for all involved languages together.

There are two main strategies for simultaneous multilingual speech recognition, with explicit language identification and with an implicit identification of the spoken language. The first strategy performs a language identification on the speech signal (the language may also be determined by pressing a button, etc.). After the identification of the language, the speech recognition system of the identified language is activated and the utterance is recognized. The advantage of this strategy is a performance that is identical to monolingual recognition as long as the language identification step is performed without errors.

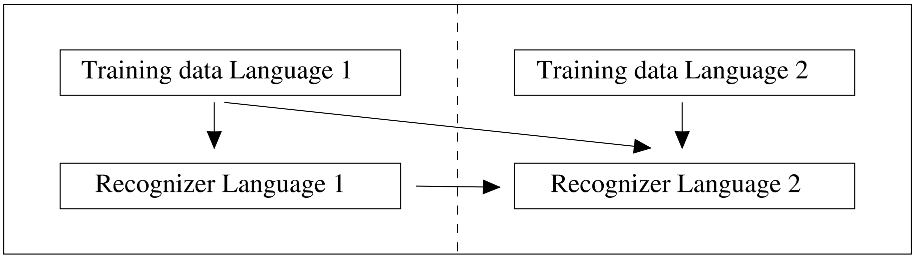

Fig. 2. Sketch of the cross-language scenario.

Table 2
Approaches to cross-language recognition

Approach	Subject of study	Languages
Dalsgaard et al. (1991) ff.	Multilingual acoustic units	British English, Danish, Italian; German, Spanish
Wheatley et al. (1994)	Seed models, phone substitution	English, Japanese
Köhler (1996) and Bub et al. (1997)	Adaptation, multilingual acoustic units, seed models	American English, German, Spanish; French, Italian, Portuguese, Slovenian
Bonaventura et al. (1997)	Multilingual acoustic units	British English, German, Italian, Spanish
Deng (1997)	Phonological feature set	American English, Canadian French, Mandarin/Cantonese Chinese
Schultz and Waibel (1998)	Multilingual acoustic units	Croatian, German, Japanese, Korean, Spanish, Turkish

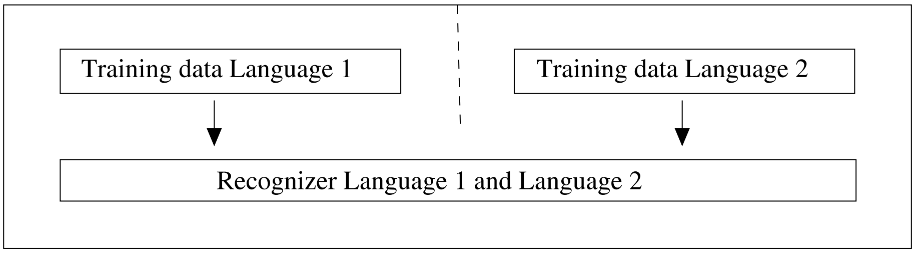

Fig. 3. Sketch of the simultaneous multilingual scenario.

The other strategy performs an implicit language identification. The words of all involved languages can be recognized equally or by language model distribution. A transition between the languages is possible.

The spoken language may be determined according to the recognized words. For this strategy, the same acoustic models may be used for the involved languages. Furthermore, one multilingual language model may be used instead of a cluster of different monolingual language models sharing a common start and end node.

The strategy performing best within this approach may vary depending on the given languages and data. For example, if there are only few data for one language, acoustic units may be shared across languages. If the languages are similar or if there are non-natives among the speakers, multilingual units may lead to a better performance. On the other hand, if languages are similar, it may be more useful to separate the languages as much as possible in order to avoid confusions among the languages.

For the use of language models, it may be better to design monolingual language models in order to avoid confusions. On the other hand, if a wrong decision has been made after language identification or according to better values in the search path, the loss in performance may only be minimized if the language model allows a transition to the other language. Consequently, there is a big variety of strategies and for each application, another one may perform best. Approaches to simultaneous multilingual speech recognition are shown in Table 3.

3. Data bases

The data used in our experiments result from three projects: the European Union (EU) project Spoken Queries in European languages (SQEL), the EU project Speech Recognition for Data-Entry (SPEEDATA), and from the BMBF (Ministry of Education, Science, Research and Technology) project VERBMOBIL.

Table 3
Approaches to simultaneous multilingual recognition

Approach	Subject of study	Languages
Ackermann et al. (1996a)	Multilingual acoustic units, multilingual LMs in a bilingual region	Italian, German
Weng et al. (1997)	Multilingual acoustic units, multilingual LMs, language identification	American English, Swedish
Harbeck et al. (1997)	Multilingual codebook, multilingual LMs, language identification	Czech, German, Slovak, Slovenian
Schultz and Waibel (1998)	Multilingual acoustic units	Croatian, Japanese, Korean, Spanish, Turkish
Ward et al. (1998)	Multilingual acoustic units	American English, American French

The SQEL project covers the languages Slovak, Slovenian and Czech in an information system for train and flight time tables. The SPEEDATA project covers the languages Italian and German, both spoken by dialect and non-natives speakers. The task of the project is the entry of land register data in the bilingual region of South Tyrol in the original language, thus the rate of non-native speech will always be around 50%. The VERBMOBIL project deals with date scheduling among humans in Japanese, English and German including automatic translation between the languages.

An overview of the training data used from these projects is given in Table 4. With these data, we cover seven languages (German (G1, G2), Italian (It), Slovak (Sa), Slovenian (Se), Czech (Cz), Japanese (Jp) and English (En)), while German is covered twice. The German data assigned with G1 result from the SPEEDATA project and contain also dialect and non-native speakers whereas the data set G2 from the VERBMOBIL project covers only native German speech.

The data consist of spontaneous speech for most of the languages, only for G1 and Italian read speech was recorded. Due to the high amount of non-natives and dialect speakers trying to speak the standard language there are still a couple of hesitations and corrections.

The size of the vocabulary differs much among the different tasks and languages. The smallest vocabulary size is observed for the train/flight information domain with around 1000 words per language. For the other domains, land register data-entry and date scheduling the vocabulary are higher and vary among 2000 and 7000 words depending on the language.

4. Phone substitutions

Each language has its own characteristic set of phonetic units. In order to perform cross-language or simultaneous multilingual speech recognition, a relation among the acoustic inventory of the involved languages must be found. This section will show strategies to determine similarities across sounds. The relation between the sounds of different languages is based on these estimated similarities.

In cross-language recognition, there are trained phones of one language, and untrained (or not well trained) phones of the language to be recognized. The sounds of each of the languages to be recognized must be replaced by the most similar trained sound of the other language. Among the languages used in this work, the sound /y/, as in süß, is unique to German. When performing recognition in German having trained only with other languages, the most similar sound to /y/ must be determined from the set of vowels.

Table 4
Acoustic data for each language

Language	G1	It	Sa	Se	Cz	Jp	En	G2
Data/h	8.6	7.6	5.1	6.1	7.2	27.4	9.6	28.5
Distinct vocabulary	5455	6748	1061	955	1323	3207	2157	7444

In simultaneous multilingual speech recognition, monolingual phones are trained for each of the language to be recognized. Some of the sounds of different languages may be similar enough to be represented as the same sound in order to reduce the number of parameters of the recognition system.

In speech recognition, usually phonemes are used for the representation of the pronunciation of the words. Phonemes are defined as the smallest unit in speech leading to a difference in the meaning of a word, whereas phones are characterized according to their acoustic properties. Thus, a phoneme may be realized by different phones, for example the phones /x/ and /c/ may be represented by the same phoneme. The relation between the phones and the phonemes of a language differs across languages. For example, in Japanese, no distinction is made between /r/ and /l/ and they would thus belong to the same phoneme class in that language. In other languages, however, they represent each a phoneme class on their own. There, a semantic difference occurs such that words get a new meaning when e.g. /r/ is replaced by /l/. Some sounds are unique to a language, for example the vowel /y/ appears only in German within the languages involved here.

In order to establish relations between the sounds of different languages, we must determine for each sound of the recognition language the most similar sound of the training language. For the substitution of a sound of the recognition language, we can use one (1:1) or more ($n:1$) sounds of the training language. For the latter type of substitution, the parameters of /y/, for example, are estimated with both /I/ and /u/ of the training language. In this work we will refer to the first strategy of a 1:1 mapping. In the following sections, we will present three possible strategies.

4.1. Na(t)ive approach

This approach of phone substitution follows the principle a non-native follows when speaking a second language. A non-native should use the phonetic inventory of the spoken foreign language. The accent of a non-native speaker is, among others, determined by the phonetic inventory of his mother tongue he uses when speaking the foreign language. A non-native tries to use the phonetic inventory of the new language. Still, he may fall back to his native phone inventory in stress conditions or within difficult words. For example, Japanese speaking English or German often confuse the use of /r/ and /l/ (Kawai and Hirose, 1998).

For some sounds there is a similar one in the native language of the speaker, and the speaker may not learn the small difference for the new sound and will only use the sound of his native language. This may happen for vowels with a small difference in the mean values in their formant frequencies between the languages (Delattre, 1965). Comparing for example English and German, we find that the English vocalic system is less rounded than the German one. Despite the elaborated vowel system in both English and German, there is no big overlap in the vowel inventory of both languages.

For example, German /y/ is not used in any of the other languages of this study. A non-native speaking German would usually replace this sound with /I/ or /i/ of his native language, if he does not pronounce the sound correctly. According to phonetic features, these sounds share most characteristics with /y/. Still, there may be other sounds with the same number of shared phonetic features which could also serve for substitution.

When replacing sounds in this way it is necessary to have some knowledge of the involved languages. Furthermore, it is necessary to have listened to people speaking as non-natives with the same combination of languages as the recognition system. Thus, for recognition of language B having trained the recognizer with language A, it is necessary to have listened to a speaker with native language A speaking language B.

4.2. Phonetic approach

This strategy follows principles in the production of sounds in the human vocal tract. Similarity is based on a similar production of sounds. When describing the production of sounds, usually the *place* and *manner* of the production are determined. The place of production describes where obstacles are put in the air flow and which organs are involved in the production of sounds. The manner of production describes how the obstacles act, e.g. if a complete or partial closure of the air flow is caused.

Thus, consonants can be distinguished with regard to the manner (stop, fricative, approximant, lateral, rhotics and others) and place (labial, dental, alveolar, palatal, velar, alveolar and others). Another criterion is the voicing of consonants which can be either voiced or unvoiced. For vowels, different tongue positions are distinguished such as front, central, back, and for the opening of the mouth among close, close-mid, open-mid, open as well as between rounded and unrounded to describe the shape of the lips. More details about the production of sounds are given, for example in (Ladefoged and Maddieson, 1996).

The difference between consonants is clearer than between vowels, for example a plosive has a complete closure, while others like fricatives do not have a complete closure. Since there is no gradual transition between a 'complete closure' (plosive) and 'no complete closure' (fricative), there is no sound between a plosive and a fricative. For vowels, however, the position of the tongue can gradually change, and the distinction and classification of vowels can become more difficult.

The trained sound that matches the largest number of phonetic features is substituted for the untrained sound. For example, /p/ (plosive, labial, unvoiced) may be replaced by /b/ (plosive, labial, voiced) or by /t/ (plosive, dental, unvoiced).

As in the above example, if /p/ shall be replaced by /b/ or /t/, it must be determined which of the sounds with one difference in the production features shall be used instead of the original sound. A study on the importance of phonetic features to determine the similarity in order to achieve a better performance is presented in (Dalsgaard et al., 1998).

For the choice of the sound that is regarded as the most similar one out of a set of sounds with phonetic features similar to the sound to be replaced, we followed two strategies. The first prefers sounds with an identical manner over those with an identical or similar place. The second strategy follows the principles of non-native speech, and chooses that sound that would be used in non-native speech.

After the most similar sound has been determined, it must be checked if this sound occurs in the language with which the recognizer was trained. Otherwise, the second similar sound must be checked until a sound is found that was trained. For Japanese, this may lead to a problem, since Japanese only shows few sounds compared to the other languages used here. The German vowels /2/, /9/ and /@/ are all modeled by Japanese /e/ since no other more similar vowel exists in Japanese.

4.3. Data-driven approach

This approach determines the similarity among phones with the data given by the trained recognizer. This approach is only possible if there is at least some data available for the new language to be recognized in cross-language recognition. For simultaneous multilingual recognition it can be used to determine the similarity of sounds and to set a threshold, such that any degree of joining sounds into a multilingual acoustic set can be realized.

Measures for the similarity can be estimated from the Gaussian densities or the codebook parameters of a trained recognizer. Therefore a recognizer must be trained with all languages, and for all observations of a language-dependent sound the similarity parameters must be estimated. According to a distance measure the most similar units may be joined. This merging of units can happen in one or more steps and it may also be allowed to split units. The advantage of this approach is that there is no human knowledge or manual work necessary to estimate similarities.

Ideally, the labeling of the speech signal into sounds should be done without errors, since a wrong classification of the time frames may lead to unsatisfying parameters concerning the similarity of sounds. An inconsistent pronunciation of a sound may also lead to an unsatisfying parameterization and, thus, to an unsatisfying substitution or clustering of sounds.

4.4. Comparison of the phone substitution approaches

The phonetic description better separates consonants than vowels into classes while the classification of vowels correlates with formant frequencies and there are no definite boarders between the vowels. Any formant frequency between two (cardinal) vowels can be realized and the classification of a sound in between is difficult. However, this characteristic may make it easier to calculate the parameters of untrained sounds by averaging the parameters of the sounds that average in the same formant frequency.

Another decision is about the type of acoustic units that will be used for the recognizer, especially if the units ought to be mono- or multilingual. Having for example *n* /a/'s of *n* languages trained, it can be decided if the sound /a/ of the target language shall be modeled as the /a/ of a particular language or from a mixture of /a/'s of different languages.

With the data-driven approach it may be determined according to the data if all or only a couple of /a/'s shall have an influence on the modeling of the new /a/, whereas for the first two approaches more complex studies have to be done for this decision.

An example of phone substitution is shown in Table 5. There, Slovak is recognized with Slovenian as training language, thus the Slovak sounds must be modeled by Slovenian sounds. For the data-driven approach, we chose the most similar sound of the training language, regardless if a sound of the spoken language was assigned as more similar. The substitution for this example shows that all data-driven substitutions lead to pairs of sounds that can be explained by phonetic features, although often not the closest sound was chosen.

An interesting substitution is found for the sound /8/, that is often described as pronouncing /t/ and /r/ at the same time. Contrary to the phonetic classification as a plosive, this sound is regarded as more similar to the other sound /r/.

Table 5
Strategies for the three substitution approaches

Slovak ← Slovenian			
Sound	Na(t)ive	Phonetic	Data-driven
7	t	t	d
8	d	d	r
J	nj	N	n
L	j	l	l
x	h	h	h

Comparing the results of these different strategies for phone substitution it can be found that the na(t)ive and phonetic approaches are quite similar, of course depending on the priorities set for substitution to manner or place in the phonetic approach. Differences occur mostly when the orthography proposes the pronunciation of another native sound than the similarity according to acoustic features would propose. For example, in the na(t)ive approach, /u/ may be replaced by /U/ according to the same orthographic spelling [u] rather than to the possibly phonetically closer /o/.

Approach 3 is only possible if a certain amount of data is available for all languages; in general it is used for the design of multilingual acoustic units. Errors in this approach can occur if there is not sufficient data available for each language and thus the parameters have not been well estimated. Another source of error for the third approach may be given when the labeling of the speech material according to acoustic units is not completely correct, for example with automatic segmentation. Sometimes, silence is assigned to a certain sound and changes the statistic properties of that sound.

Another source for errors are different recording conditions. A consequence may be that sounds of the same language without respect to their phonetic features are estimated as more similar than any sound of the other language. In our experiment, this happened for Slovenian sounds which were classified in many cases as more similar within Slovenian than any sound of another language.

One special phenomenon that has arisen in data-driven decision is the similarity of /j/ and /z/ which have quite different phonetic characteristics (approximant–palatal–voiced versus fricative–alveolar–voiced), which has also been shown in several other approaches (Dalsgaard et al., 1998; Jo et al., 1998), thus some other measures may be important besides the phonetic features determined so far.

We performed recognition experiments with the three types of phone substitution with the project data. The highest word accuracy was achieved with the na(t)ive approach, it performed slightly better than the phonetic approach. The data-driven approach led to satisfying results only for the data from the SpeeData project. The results of the following experiments will use the na(t)ive approach for phone substitution.

Table 6
Word accuracy for porting experiments with language models for seven different languages

It	G1	Sa	Se	Cz	En	Jp	G2
94.2	87.9	88.3	90.3	88.6	48.2	64.5	37.1

5. Porting

The first step in multilingual recognition is porting if sufficient training material is available for the new language. This way, we have a baseline system with which we can compare our results in cross-language and simultaneous multilingual recognition. Our recognizer ISADORA (Kuhn, 1995; Schukat-Talamazzini, 1995) designed for German, is ported to Italian, Slovak, Slovenian, Czech, English and Japanese. For German, we have trained two recognizers, one from the VerbMobil project with native speakers, the other from the SpeeData project with natives and non-natives.

The first step is to transliterate the utterances and determine acoustic units. Usually, the utterances will be transliterated at the word level. For Japanese, there is a choice between word-like units or syllables as models for the transliteration of utterances. Here, we used word-like units for Japanese.

Then, the pronunciation of the words or word-like units must be determined. Therefore, the phonetic inventory of each language is created. For many languages a transcription in SAMPA notation is already available. We took the pronunciation lexica from the projects SpeeData, Sqel and VerbMobil which all have transcription similar to SAMPA. A slight difference in the degree of granularity of the acoustic units for the two German recognizers was eliminated. Furthermore, the use of context-dependent acoustic units, polyphones, is studied. For most experiments mentioned here, we show results with the modeling of context-free monophones.

We performed our experiments with semi-continuous Hidden Markov Models and language models. The experiments performed for this contribution are done without optimization, i.e. without using the technique of polyphones for acoustic units, without using a polygram verification for language modeling and without optimizing the training procedure in order to obtain recognizers trained at the same level. Thus, the results given here, do not correspond to the optimally trained recognizers, but are comparable to each other with respect to modeling and training. The results for porting are shown in Table 6.

The performance largely varies across the involved languages. On the one hand, this is due to a different degree of difficulty of the task caused by a different size of the recognition vocabulary, different type of speech, and the speakers themselves with native and non-native speakers. In the SpeeData task the best recognition is achieved, followed by Sqel and finally by VerbMobil.

On the other hand, the languages themselves have a different difficulty for recognition, some languages may be easier to be recognized than others due to the phonetic structure, word length and other reasons. For example, Italian seems to be easier for automatic speech recognition (also see (Barnett et al., 1996)) with a better performance for Italian than for the other languages. Within the Sqel project, Slovenian seems to lead to the best performance.

Further on, we will compare the results in the following sections with this baseline performance.

6. Cross-language recognition

In this section, we will present experiments and results in the field of cross-language recognition. Recognition is performed on a language that was not used for training. Training is performed with one or more languages of the available seven languages. We compare both the performance of different monolingual recognizers as well as the difference to multilingually trained recognizers.

The first step in cross-language recognition is to find a relation between the trained sounds and the sounds of the language to be recognized.

We considered sounds represented by the same SAMPA symbol in the training and recognition languages as identical. The use of such a sound of the training language in the recognition language is not counted as a sound substitution opposite to using sounds with different production criteria. Furthermore, we did not count replacements for the length of phones, i.e. if there existed only a long vowel like /i:/ and the short correspondent /i/ was needed, we did not count this as substitution. The same is done for Italian geminates, thus /nn/ was set equal to /n/ and the substitution was not counted.

Cross-lingual recognizers contain the same number of sounds as those resulting from porting. There may be, however, a difference with respect to the trained parameters. For example, in the cross-lingual recognition of Italian, both /n/ and /nn/ must be modeled. If the training language does not distinguish between /n/ and /nn/, only trained parameters for /n/ are available. In this case, the Italian sound /nn/ is also modeled with the parameters of /n/ of the training language. Since these sounds /n/ and /nn/ have identical production criteria, this replacement is not counted as a substitution, but is regarded as a difference in the degree of granularity in the acoustic modeling.

In Table 7 the number of substitutions across languages is shown. There are no substitutions between G1 and Italian since they share proper names of both languages and thus phones of both languages are modeled for each recognizer. Between G1 and G2 there are two substitutions for originally Italian phones (/J/, /L/) which are used in the G1 recognizer. There is a high number of substitutions between the Germanic languages (English, German) on the one side and the Slavic languages (Slovak, Slovenian, Czech) on the other side, once due to the high number of consonants modeled in the Slavic languages and the high amount of vowels in the Germanic languages. For multilingually trained recognizers, less substitutions have to be made than for the corresponding monolingual recognizers.

Furthermore, we can observe that, using the Japanese recognizer for the recognition of any of the other languages, a high number of substitutions has to be made, since the phone inventory of the Japanese language is small in comparison to those of the other languages. In this case, several sounds are modeled with one Japanese sound. On the other hand, for the recognition of Japanese with any other recognizer, only a small number of substitutions has to be performed. In this case, only a part of the trained parameters is further used for Italian, since the speech with sounds not occurring in Japanese is not further taken into account.

We performed experiments on the different strategies for phone substitution discussed before. For the data-driven technique, we found satisfying

Table 7
Substitution of phones with different languages and recognizers

Rec\Lg	It	G1	Sa	Se	Cz	Jp	En	G2
It	0	0	3	1	4	0	8	0
G1	0	0	3	1	4	0	8	0
Sa	10	10	0	4	6	4	12	11
Se	9	9	5	0	7	2	9	8
Cz	12	12	7	5	0	3	11	11
En	11	11	8	3	7	3	0	9
Jp	12	12	9	6	9	0	13	10
G2	2	2	4	0	5	0	7	0
It–G1	0	0	3	1	4	0	8	0
Se–Sa	7	7	0	0	3	2	8	7
Sa–Se–Cz	7	7	0	0	0	2	8	7
G2–En	2	2	3	0	4	0	0	0
G2–En–Jp	2	2	3	0	4	0	0	0

results only for the Italian and German SpeeData recognizers. For other languages the similarities do not correspond to phonetic properties. For Slovenian, for example, the phones classified as most similar were in most cases also Slovenian phones, probably the recording conditions dominated over the phonetic similarities. Restricting the estimation of similar sounds of other languages than Slovenian, we partly find similarities that can be explained by phonetic properties, but also some similarities that cannot be explained at all with production similarities.

We performed the phonetic approach with different hierarchies on the replacement of the sounds, like proposed by Dalsgaard et al. (1998). With the native substitution approach, which differs only in few sounds from the phonetic approach, we found slightly better word accuracies than for the phonetic approach. The results we present in the following have been achieved with na(t)ive phone substitution (Table 8).

For all languages besides German G2, recognition is best for the monolingual recognizer trained with data of that language and domain. For G2, recognition was shown to be better for the bilingual German–English recognizer under these conditions.

The recognition rate decreases in cross-lingual recognition in a different degree depending on the language that was used for training. In most cases, the cross-lingual performance is better for related languages (e.g. Slovak, Slovenian and Czech), i.e. using one of the languages for training and the another one for recognition.

In order to compare the performance of the cross-lingual recognizers trained with one language we averaged the performance of all recognizers besides the one of the original language and domain. Best cross-lingual recognition averaged over the seven other recognizers was achieved for Italian with 78.7%, worst performance was achieved for G2 with 11.4%.

Both these calculations are difficult for interpretation since the similarity of languages and thus the recognizability cannot be taken into account. For example, we have two German recognizers in the cross-lingual experiments but only one of the other languages. Assuming a higher similarity among the Slavic languages, the cross-lingual performance should be higher when recognizing with Slavic recognizers for the Slavic languages than with other recognizers. Furthermore, the cross-lingual recognition of Japanese could be worse because there are no languages similar to Japanese used for recognition.

In Table 8, we can observe, that starting with a poor recognition rate for monolingual recognition, the performance for cross-lingual experiments suffers more than for languages and domains where the monolingual performance is already higher.

Table 8
Word accuracy for cross-lingual experiments[a]

Rec\Lg	It	G1	Sa	Se	Cz	En	Jp	G2
It	$\boxed{94.2}$	**70.7**	22.2	38.6	59.0	7.4	18.3	18.7
G1	81.0	$\boxed{87.9}$	28.1	31.0	55.3	8.5	17.9	**20.6**
Slovak	77.6	57.1	$\boxed{88.3}$	**71.0**	**68.9**	7.5	20.2	2.6
Slovenian	**86.6**	60.7	**66.6**	$\boxed{90.3}$	52.3	8.9	**30.2**	2.0
Czech	81.5	57.1	35.0	58.5	$\boxed{88.6}$	10.5	22.0	5.9
English	41.4	36.1	35.4	26.8	42.7	$\boxed{48.2}$	20.3	3.1
Japanese	83.3	56.8	40.7	44.1	36.3	5.5	$\boxed{64.5}$	1.1
G2	81.5	67.6	39.6	59.4	53.5	**12.2**	25.2	$\boxed{37.1}$
G1–It	94.1	86.7	28.2	43.2	63.9	8.8	27.3	19.5
Sa–Se	85.8	60.6	84.2	88.0	62.4	7.4	28.0	1.8
Sa–Se–Cz	86.7	65.0	84.1	85.7	83.9	7.2	30.0	1.5
En–G2	84.4	77.1	38.3	60.7	62.3	24.5	29.6	47.0
En–G2–Jp	87.9	73.9	46.4	64.2	64.4	24.1	52.4	46.5

[a] Lines: performance of each recognizer, columns: performance of a language with different recognizers); performance of $\boxed{\text{p}}$orting approach, best **cross-language** performance.

Averaging the performance of cross-lingual recognizers on different spoken languages, we find that, for monolingually trained recognizers, the best cross-lingual performance was achieved by the Slovenian recognizer which lead three times to the best cross-lingual recognition, whereas Czech, English and Japanese never performed best. Thus, the Slovenian recognizer seems to be best for cross-lingual recognition in this task. Only Slovak and Slovenian showed mutually the best performance for cross-lingual recognizers and may therefore be assumed similar for this speech recognition task, although theoretically, Slovak and Czech should be more similar than Slovak and Slovenian due to their close linguistic relation.

For other languages, there is no such symmetry observable, even the two German recognizers do not lead to highest reciprocal results: G1 recognizes best G2, but not vice versa. This may be due to different speaking and recording styles, but more probable to the different speaker characteristics, since the speakers of G1 speak with a dialect and with a non-native accent, while the G2 speakers are German natives and do not speak with a strong dialect.

With multilingual recognizers trained with several languages, performance is worse than with the appropriate monolingual recognizer. Having the target language not included into training, the performance is better than with cross-lingual monolingual recognizers. Unfortunately, for some of the languages with a high cross-lingual performance, no multilingual recognizers were trained, thus often the best monolingual cross-lingual recognizers perform better than the best multilingual recognizers trained in these experiments.

Of the available multilingual recognizers, the G2–English–Japanese recognizer performs best for these data, possibly due to a larger variety in the models provided by Japanese in addition to the Germanic languages.

7. Simultaneous multilingual recognition

There are several techniques for simultaneous multilingual speech recognition (from top to bottom the languages are treated more in common):

1. *Language identification first*. Language identification is performed on the speech signal or in another way. Then, recognition is performed on monolingual recognizers.
2. *Two separate monolingual recognizers that share a common start and end node*. There is no parameter sharing and no transition possible between the recognizers. Language identification is done implicitly by the chosen path and, thus, by the chosen recognizer.
3. *There is one set of acoustic units* (*some or all may be shared across the languages*). Monolingual language models are used. Some of the parameters of the recognizer are shared across the languages, but no transitions between the language models and, thus, between the languages are possible.
4. *Acoustic units and language models are shared across the languages*. All parameters of the recognizer are shared across the languages. Transitions between the languages are possible at any point.

In this study, we concentrate on the last two approaches. Sharing of acoustic units in approach 3 can reduce the number of parameters and guarantee better trained parameters, if we can use similarities between the involved languages. If the languages are very different and no acoustic units are shared, there is no difference to approach 2. Sharing the language models in approach 4 again reduces the number of parameters and the complexity of the recognition process. This approach may be useful, if the languages are very similar or if they share some words like in a bilingual region. With this approach, transitions between languages can be recognized, especially utterances started in one language and continued in another language. The number of parameters is further reduced, for instance when some words like proper names occur in both languages.

We compared the performance of approaches 3 and 4. Approaches 3 and 4 differ mainly in the use of language models. While approach 3 uses one language model for each language (the language models are put in parallel for multilingual recognition (explicit)), approach 4 uses one language model for all involved languages (implicit). The recognizer of approach 3 is larger and more precise, but does not allow transitions between the languages and does not take into account similarities between the languages.

In most cases, approach 3 performed better, even in the SpeeData project of the bilingual region of South Tyrol in Italy where the languages influence each other (Table 9). There, we joined both 2 and 6 language models in both languages for different text types according to the strategy of approaches 3 and 4. The test data are different from the ones in this study, so the recognition rate differs from the one given before. The accuracies given in this table, however, show the difference when combining language models. The performance does not decrease much using approach 3 compared to the baseline system with separate languages, whereas the performance of approach 4 is about 10% worse than the other approaches. In the following, we will describe our experiments with the recognizer of approach 3.

Results in the simultaneous recognition of two and three languages are shown with the examples of the Slavic languages Slovak, Slovenian and Czech in Table 10.

Evaluation is done with monolingually and multilingually trained recognizers. Acoustic units with the same symbol are treated as identical, phone substitution was performed with the na(t)ive approach. Evaluation was done for a bilingual (Slovak and Slovenian) and a trilingual recognizer (Slovak, Slovenian, Czech). Results are given on average and split with respect to the spoken language.

Looking only at the recognized languages, the performance is slightly worse for the trilingual task compared to the bilingual one for most recognizers. Compared also to the monolingual task in cross-lingual recognition in Table 2, there is a loss in performance when adding languages. For example, when recognizing Slovak with the Italian

Table 9
Word accuracy without combining language models (baseline system), joining 2 or 6 language models[a]

Word accuracy	Baseline			2 language models			6 language models		
	It. + G1	It.	G1	It. + G1	It.	G1	It. + G1	It.	G1
Explicit	87.2	91.1	82.3	86.6	91.0	80.9	85.2	89.9	79.1
Implicit	87.2	91.1	82.3	80.2	84.6	74.5	76.1	81.1	69.6

[a] Explicit: two language models in parallel, implicit: one common language model for different languages.

Table 10
Word accuracy for simultaneous multilingual experiments[a]

Rec\Lg	Sa–Se	(Sa	Se)	Sa–Se–Cz	(Sa	Se	Cz)
It	26.7	20.3	33.8	33.6	16.6	32.2	53.0
G1	24.0	20.2	28.3	29.0	11.5	25.4	51.1
Slovak	78.8	88.1	68.4	74.0	88.1	68.1	64.7
Slovenian	77.3	65.4	90.7	66.7	64.0	90.7	46.9
Czech	44.9	34.1	57.0	51.7	16.1	52.8	88.8
Japanese	35.4	34.1	36.9	29.1	33.8	36.7	16.8
English	25.1	31.0	18.5	27.6	26.2	14.8	41.4
G2	46.6	36.9	57.6	44.1	28.5	57.2	48.4
It–G1	29.1	21.6	37.6	35.9	19.0	36.8	53.1
Sa–Se	84.9	82.6	87.6	76.5	82.3	87.3	59.9
Sa–Se–Cz	83.6	81.9	85.4	83.7	82.0	85.4	83.9
En–G2	44.7	31.6	59.5	47.4	29.1	58.8	56.1
En–G2–Jp	44.3	30.7	59.6	46.5	27.6	59.2	54.7

[a] Lines: performance of each recognizer, columns: performance of a language with different recognizers.

recognizer the word accuracy is 22.2% for monolingual, 20.3% for bilingual and 16.6% for the trilingual task. No general rule can be found for the amount in which the recognition decreases when adding languages for the recognition process.

In two cases, the performance increases insignificantly when going from monolingual to trilingual recognition: recognizing Slovenian resp. Czech when training with the corresponding language only.

The best overall recognition was achieved when using those languages for training that are going to be recognized, i.e. Slovak and Slovenian for the bilingual task and Slovak, Slovenian and Czech for the trilingual task. In some cases, the trilingual recognition performs better in the overall evaluation because a higher accuracy is achieved for Czech than for the other languages.

The highest performance for this trilingual task *without* using training material of the languages to be recognized is achieved with the English–Japanese recognizer obtaining an accuracy of 47%.

8. Conclusion

In this contribution, we performed experiments in several dimensions of multilingual speech recognition. We pursued different approaches to speech recognition with the seven languages. We ported our recognition system to new languages, using one language for training and recognizing that language. Furthermore, we used multiple languages for training and recognized one language that was not included into training. Finally, we trained with several languages and were able to recognize different languages at the same time.

In addition, we performed studies on the substitution of phones which is necessary for cross-lingual and simultaneous multilingual recognition. For simultaneous multilingual recognition, we studied different approaches to recognize multiple languages at the same time.

We performed our experiments with non-optimized recognizers (only monophones, no polygram verification in the language models, no optimization in the training).

In porting, we achieved quite different recognition rates, since the degree of difficulty in the projects varied and the languages themselves have a different recognizability for automatic speech recognition.

In cross-lingual recognition, performance is best for monolingual recognizers, besides for the German G2 task. When monolingual recognition is already bad, cross-lingual performance gets even worse. Thus, for Italian, the average decrease in performance is 15%, whereas for G2 only one-third is recognized with respect to the monolingual recognizer. Cross-lingual performance does not show a strong symmetry in the recognition, only Slovak and Slovenian recognize utterances of the other language better than any other language.

When recognizing with multilingual cross-lingual recognizers, performance gets better than with the corresponding monolingual recognizers. Unfortunately, we have not trained all combinations of recognizers, so the combination of the best monolingual cross-lingual recognizers could not always be tested.

If we want to perform a recognition on a language without having acoustic data, we found the best performance when we include several languages with similar recording conditions into training. The sounds of the training languages with identical production criteria are merged to the respective sound of the language to be recognized. In this case, both acoustic units are modeled with more variety and more training material.

In simultaneous multilingual speech recognition, we studied different methods to combine the languages at the level of acoustic units and language models. We found best performance when sharing acoustic units and separating language models. This way, no transition between the languages to be recognized is possible, but a wrong decision at the beginning recognizes the wrong language for the complete utterance. Still, the recognition is best for this approach.

As an example, we show results for the recognition of Slavic languages for bilingual and trilingual recognitions. The recognition of Slovak and Slovenian decreases in most cases slightly when adding further languages for recognition. No rule could be found for the amount of loss in performance with different training languages. The best overall performance was achieved when using the same languages for training and recognition, thus Slovak and Slovenian for the bilingual and Slovak, Slovenian and Czech for the trilingual recognizer. Best cross-language performance was achieved when training with English and German resp. English, German and Japanese for bilingual and trilingual recognition.

Multilingual speech recognition is a very important aspect for the future of speech recognition. There are many applications that need porting, cross-lingual or simultaneous multilingual speech recognition. Among others, two things are important for a satisfying performance: the right choice of the languages for training, if not sufficient training material of the target language is available. Therefore, the phones of the different languages have to be set in relation and a good algorithm must be found for establishing this relation. In this contribution, we have presented some possible and promising algorithms. A second important feature is to design a suitable architecture for simultaneous multilingual speech recognition. Depending on the application, it may be necessary to allow for transitions between languages. The usage of common acoustic units and common language models may be recommended. In our approach, we found it best to combine acoustic units and to separate language models.

A further feature that has to be studied in the future, is to determine the similarity of languages in order to use the most similar language for training and establish a pool of multilingual acoustic units of which the best suiting will be chosen for cross-lingual applications. Further strategies for simultaneous multilingual recognition must be found in order to achieve a performance that is equal (or even better) than the monolingual one in order to provide reliably working systems for the many international applications in the future.

References

Ackermann, U., Angelini, B., Brugnara, F., Federico, M., Giuliani, D., Gretter, R., Niemann, G.L.H., 1996a. SpeeData: multilingual spoken data entry. In: Proc. Int. Conf. on Spoken Language Processing, Philadelphia, PA.

Ackermann, U., Brugnara, F., Federico, M., Niemann, H., 1996b. Application of speech technology in the multilingual SpeeData project. In: 3rd Crim-Forwiss Workshop, Montréal.

Barnett, J., Corrada, A., Gao, G., Gillick, L., Ito, Y., Lowe, S., Manganaro, L., Peskin, B., 1996. Multilingual speech recognition at dragon systems. In: Proc. Int. Conf. on Spoken Language Processing, Philadelphia, PA.

Bonaventura, P., Gallochio, F., Micca, G., 1997. Multilingual speech recognition for flexible vocabularies. In: Proc. European Conf. on Speech Communication and Technology, Rhodes, Vol. 1, pp. 355–358.

Bub, U., Köhler, J., Imperl, B., 1997. In-service adaptation of multilingual hidden-Markov-models. In: Proc. ICASSP'97, Munich, Vol. 2, p. 1451.

Cerf-Danon, H., De Gennaro, S., Feretti, M., Gonzalez, J., Keppel, E., 1991. TANGORA – a large vocabulary speech recognition system for five languages. In: Proc. European Conf. on Speech Communication and Technology, Genova, Vol. 1, pp. 183–186.

Dalsgaard, P., Andersen, O., Barry, W., 1991. Multi–lingual label alignment using acoustic–phonetic features derived by neural network technique. In: Proc. Int. Conf. on Acoustics, Speech and Signal Processing, Toronto, pp. 197–200.

Dalsgaard, P., Andersen, O., Barry, W., 1998. Cross-language merged speech units and their descriptive phonetic correlates. In: Proc. Int. Conf. on Spoken Language Processing, Sydney, Vol. 6, pp. 2627–2630.

Delattre, P., 1965. Comparing the Phonetic Features of English, French, German and Spanish. Julius-Groos-Verlag, Heidelberg.

Deng, L., 1997. Integrated-multilingual speech recognition using universal phonological features in a functional speech production model. In: Proc. Int. Conf. on Acoustics, Speech and Signal Processing, Munich, Vol. 2, pp. 1007–1010.

Fabian, P., 1997. Adaptation of the SQEL speech recognition system for the Polish language. In: Proc. of the 2nd SQEL Workshop on Multi-Lingual Information Retrieval Dialogs, Plzen, pp. 96–103.

Glass, J., Flammia, G., Goodine, D., Phillips, M., Polifroni, J., Sakai, S., Seneff, S., Zue, V., 1995. Multilingual spoken-language understanding in the MIT Voyager system. Speech Communication 17, 1–18.

Harbeck, S., Nöth, E., Niemann, H., 1997. Multilingual speech recognition. In: Proc. 2nd SQEL Workshop on Multi-Lingual Information Retrieval Dialogs, Plzeň, pp. 9–15.

Ipsic, I., 1996. Slovenian word recognition. In: Pavesic, N., Niemann, H. (Eds.), 3rd Slovenian-German and 2nd SDRV Workshop. Faculty of Electrical and Computer Engineering, University of Ljubljana, Ljubljana, pp. 87–96.

Jo, C.H., Kawahara, T., Doshita, S., Dantsuji, M., 1998. Automatic pronunciation error detection and guidance for foreign language learning. In: Proc. Int. Conf. on Spoken Language Processing, Sydney, Vol. 6, pp. 2639–2942.

Kawai, G., Hirose, K., 1998. A method for measuring the intelligibility and nonnativeness of phone quality in foreign language pronunciation training. In: Proc. Int. Conf. on Spoken Language Processing, Sydney, Vol. 5, pp. 1823–1826.

Klečková, J., 1997. Language model for recognizer of spontaneous Czech speech. In: Proc. of the 2nd SQEL Workshop on Multi-Lingual Information Retrieval Dialogs, Plzeň.

Köhler, J., 1996. Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds. In: Proc. Int. Conf. on Spoken Language Processing, Philadelphia, PA.

Krokavec, D., Ivanecký, J., 1997. Semantic processing of sentences in Slovak language. In: Proc. of the 2nd SQEL Workshop on Multi-Lingual Information Retrieva l Dialogs, Plzeň.

Kuhn, T., 1995. Die Erkennungsphase in einem Dialogsystem, Infix, St. Augustin, Vol. 80.

Ladefoged, P., Maddieson, I., 1996. The Sounds of the World's Languages. Blackwell Publishers, Oxford, Great Britain.

Lamel, L., Adda-Decker, M., Gauvain, J., 1995. Issues in large vocabulary, multilingual speech recognition. In: Proc. European Conf. on Speech Communication and Technology, Madrid.

Nöth, E., Harbeck, S., Niemann, H., Warnke, V., Ipšić, I., 1996. Language Identification in the context of automatic speech understanding. In: Pavesic, N., Niemann, H., Kovacic, S., Mihelic, F. (Eds.), Speech and Image Understanding. IEEE Slovenia Section, Ljubljana, pp. 59–68.

Schukat-Talamazzini, E.G., 1995. Automatische Spracherkennung – Grundlagen, statistische Modelle und effiziente Algorithmen. Künstliche Intelligenz, Vieweg, Braunschweig.

Schultz, T., Waibel, A., 1998. Multilingual and crosslingual speech recognition. In: Proc. DARPA Broadcast News Workshop, Washington.

Ward, T., Roukos, S., Neti, C., Gros, J., Epstein, M., Dharanipragada, S., 1998. Towards speech understanding across multiple languages. In: Proc. Int. Conf. on Spoken Language Processing, Sydney, Vol. 5, pp. 2243–2246.

Weng, F., Bratt, H., Neumeyer, L., Stolcke, A., 1997. A study of multilingual speech recognition. In: Proc. European Conf. on Speech Communication and Technology, Rhodes, Vol. 1, pp. 359–362.

Wheatley, B., Kondo, K., Anderson, W., Muthusaamy, Y., 1994. An evaluation of cross-language adaptation for rapid HMM development in a new language. In: Proc. Int. Conf. on Acoustics, Speech and Signal Processing, Adelaide, pp. 237–240.

Young, S., Adda-Decker, M., Aubert, X., Dugast, C., Gauvain, J.L., Kershaw, D., Lamel, L., Leeuwen, D., Pye, D., Robinson, A., Steeneken, H., Woodland, P., 1997. Multilingual large vocabulary speech recognition: The European SQALE project. Comput. Speech Lang. 11, 73–89.