MULTILINGUAL ACOUSTIC MODELS FOR SPEECH RECOGNITION AND SYNTHESIS

S. Kunzmann, V. Fischer, J. Gonzalez, O. Emam, C. Günther, E. Janke

IBM Pervasive Computing
European Voice Technology Development
Gottlieb-Daimler-Str. 12, D-68165 Mannheim, Germany

kunzmann@de.ibm.com

ABSTRACT

In this paper we review the design of a common phone alphabet for up to fifteen languages and describe its application in two important components of a seamless multilingual conversational system, namely speech recognition and synthesis. We report on experiments that demonstrate the advantages of multilingual acoustic models both for the recognition of foreign names and non-native speech, and describe the usefulness of a common phone alphabet for the construction of unit selection based mono- and bilingual speech synthesis systems.

1. INTRODUCTION

Both the tremendous growth of the Internet and the simultaneous convergence of mobile phones and palm-sized computers has promoted speech recognition and synthesis into the rank of a key technology for easy and natural access to information from anywhere for everyone. However, the nature of applications such as voice enabled Internet portals, tourist information systems, or automated directory assistance imposes significant new challenges on these technologies: while speech recognition must cope with, for example, an increased number of non-native speakers with many different accents, speech synthesis must provide natural sounding speech output for foreign words or phrases from a multitude of languages. Clearly, the availability of such systems in many languages is also important from an economic point of view, which may be considered as one reason for the recent interest in the creation of databases that include a rich set of dialects as well as non-native speech, for example, [1].

Multilingual acoustic modeling facilitates the development of speech recognizers for languages for which only little training data is available, and also allows reduced complexity of applications by the creation of acoustic models that can simultaneously recognize speech from several languages [2]. The use and combination of multilingual acoustic models has also proven advantageous for the recognition of accented speech produced by a wide variety of non-native speakers with different commands of the system's operating language [3].

Whereas the design of a common phone alphabet and the sharing of (training) data from several languages is a well established method in multilingual speech *recognition*, it is an only emerging concept in the field of unit selection based speech *synthesis*. Initial work on a common diphone inventory for a seamless multilingual (or *polyglot*) speech synthesizer is described in [4], but today's systems usually achieve speech output in multiple languages by use of two or more language dependent synthesizers (see, for example, [5]), which is frequently accompanied by switching to a different voice.

Aiming at the better utilization of potential synergies between speech recognition and synthesis technologies, cf. [6], we developed a common phonetic alphabet for fifteen languages that can be used in both of these fields. After a discussion of the main design issues in the next section, we briefly review multilingual acoustic modeling techniques and describe some experiments that demonstrate the benefits of the chosen approach in a non-native speech recognition task. Section 4 discusses the use of the common phonology in the construction of a unit selection based speech synthesis system and describes some initial work towards a bilingual concatenative speech synthesizer. Finally, Section 5 gives a conclusion and an outlook on further work towards multilingual conversational systems.

2. COMMON PHONOLOGY

The definition of a common phonetic alphabet for multilingual speech recognition has to consider two conflicting design issues: on one hand the different sounds of each language should be covered separately in order to achieve high recognition accuracy, while on the other as many phones as possible should be shared across languages both for efficient utilization of training data and to achieve reasonably small acoustic models.

Starting from the existing phonetic alphabets for seven languages (Arabic, British English, French, German, Italian, (Brazilian) Portuguese, and Spanish) we have designed two common phonetic alphabets of different detail [7]. For that purpose, the language specific phone sets were first simplified following available SAMPA transcription guidelines [8]; Arabic SAMPA has meanwhile been standardized as part of the Orien*Tel* project [1]. With this approach, languages were affected to different degrees: While the native French phone set remained unchanged, we gave up syllabic consonants for German, and at the same time introduced new diphthongs for British English. In a second step, language specific phones mapped to the same SAMPA symbol were merged into a common unit. This yielded a common phonetic alphabet of 121 phones (65 vowels, 56 consonants) for the seven languages, cf. Table 1, which provides an overall reduction of 60 percent compared to the simplified language specific phonologies.

	total	En	Fr	Gr	It	Es	Pt	Ar
vowels	65	20	17	23	14	10	20	14
cons.	56	24	19	26	32	30	22	29
total	121	44	36	49	46	40	42	43

Table 1. Number of vowels and consonants for seven languages in the detailed common phone set (a). Languages are British English (En), French (Fr), German (Gr), Italian (It), Spanish (Es), Brazilian Portuguese (Pt), Arabic (Ar).

In a less detailed common phonetic alphabet, cf. Table 2, we gave up the distinction between stressed and unstressed vowels for Spanish, Italian, and Portuguese, and represented all long vowels and diphthongs as as sequence of two (identical) short vowels. In doing so, the average number of languages that contribute to the training data for each of the 76 phones (the *sharing factor*) increased from 2.28 to 2.53, or — if Arabic is not considered — from 2.74 to 3.56, while the average word error rate increased by 7 percent on an in-house database if compared to the more detailed common phone alphabet [7].

	total	En	Fr	Gr	It	Es	Pt	Ar
vowels	31	13	15	17	7	5	12	11
cons.	45	24	19	23	28	24	22	28
total	76	37	34	40	35	29	34	39

Table 2. Number of vowels and consonants for seven languages in the reduced common phone set.

Given that this reduced phonology can be applied to new languages with little or no change at all, we believe this degradation to be tolerable and adjustable by improved acoustic modeling, and have integrated eight additional languages with two more vowels and 12 more consonant phones, cf. Table 3.

	Cz	Jp	Fi	El	Nl	Da	No	Sv
vowels	5	5	8	5	14	14	17	17
cons.	27	23	19	25	22	20	23	24
total	32	28	27	30	36	34	40	41

Table 3. Number of vowels and consonants additional languages integrated into the reduced common phonetic alphabet: Czech (Cz), Japanese (Jp), Finnish (Fi), Greek (El), Dutch (Nl), Danish (Da), Norwegian (No), Swedish (Sv).

3. MULTILINGUAL SPEECH RECOGNITION

Acoustic modeling for multilingual speech recognition to a large extend makes use of well established methods for (semi-)continuous Hidden-Markov-Model training. Methods that have been found of particular use in a multilingual setting include, but are not limited to, the use of *multilingual seed HMMs*, the use of *language questions* in phonetic decision tree growing, *polyphone decision tree specialization* for a better coverage of contexts from an unseen target language, and the determination of an appropriate *model complexity* by means of a Bayesian Information Criterion; cf., for example, [2, 9] for an overview and further references.

Having now reached a certain maturity, the benefits of multilingual acoustic models are most evident in applications that require both robustness against foreign speakers and the recognition of foreign words. We have simultaneously explored both of these when creating a Finnish name dialer whose application directory consists of a mix of 6,000 Finnish and foreign names, and which is used by native and non-native speakers.

For that purpose, we created acoustic models with different proportions of speech data from Finnish (SpeechDat-II), US-English, UK-English, German, Italian and Spanish. Table 4 gives details of the amount of training material that was chosen based on some experience gathered in previous experiments. While F70 is a mono-lingual Finnish acoustic model, M01 through M10 incorporate increasing amounts of data from the remaining languages (except Spanish). The largest model (M10b) contains a total of 280,000 utterances, equivalent to approx. 191 hours of speech (silence excluded).

The multi-lingual phonetic alphabet for the 5 languages under consideration is a subset of 61 phones of the reduced alphabet described above (29 phones are used for Finnish), and the average number of languages that share a phone is 3.05 for models M01 to M05 and 3.45 in case of M10 and M10b. Test data for UK-English (En), German (Gr), Italian (It), and Spanish (Es) also included native names

model	Fi	US	En	Gr	It	Es
F70	70.0	0	0	0	0	0
M01	70.0	3.0	0.8	1.5	0.5	0
M02	70.0	6.0	1.6	3.0	1.0	0
M05	70.0	15.0	10.0	7.5	2.5	0
M10	70.0	30.0	20.0	15.0	5.0	0
M10b	140.0	45.0	45.0	25.0	10.0	15.0

Table 4. Number of training utterances (× 1000) used in the various acoustic models.

	word error rate [%]				
	Fi	En	Gr	It	Es
F70	2.63				
M01	2.88	28.20	28.50	20.81	31.40
M02	2.25	21.70	21.50	17.54	24.10
M05	2.44	14.90	10.60	9.41	21.30
M10	2.44	11.10	7.70	7.83	14.70
M10b	2.07	11.50	10.40	4.86	6.40

Table 5. Word error rates vs. amount of non-Finnish training data for a 6000-name grammar task.

and foreign names from any of the other languages. While the name dialer task was of primary interest in our investigations, for Finnish (Fi) we also experimented with other recognition tasks (digit strings, natural numbers, and phonetically rich sentences, cf. also Table 7) in order to get a better insight into the decoding of native speech when incorporating more foreign speech material.

Table 5 shows word error rates (WER) for a 6,000-name grammar task as a function of the amount of non-Finnish speech used in training. As expected, the error rate decreases as the amount of training speech increases. The benefits of multilingual modeling also become evident in the case of Spanish, even though the models were not trained with any Spanish data. This suggests that the models learn from the other languages, providing robustness against native Spanish speakers.

Table 6 provides further insight, demonstrating that in general native speakers achieve better recognition rates for native names than for foreign names. However, in case of Spanish, which includes many foreign names from the other languages, the accuracy on this test set increases.

Finally, Table 7 demonstrates the effects of an increased amount of foreign training data on the decoding of native speech uttered by native Finnish speakers, all recorded under the same conditions. While (almost) no degradation is observable for names and digit strings, we obtained an increased error rate for numbers, which disappeared only when the amount of training data was increased (see M10b). Results in the rightmost column ("rich") refer to the recognition of phonetically rich sentences (recorded under different conditions), and should be understood as work towards even more ambitious applications.

	word error rate [%]			
(a)	En	Gr	It	Es
M01	24.70	24.00	23.76	31.50
M02	20.90	18.20	15.71	25.10
M05	12.30	9.20	10.64	21.40
M10	9.80	6.50	7.16	15.10
M10b	8.50	10.80	4.77	5.70

	word error rate [%]			
(b)	En	Gr	It	Es
M01	31.27	24.69	17.51	17.32
M02	28.26	23.08	16.84	15.18
M05	21.40	15.92	11.27	15.18
M10	15.22	13.15	10.88	9.64
M10b	16.56	13.06	10.21	10.18

Table 6. Word error rates for (a) native and (b) foreign names in four languages.

	word error rate [%]			
	digits	names	numbers	rich
F70	1.13	2.63	12.12	7.38
M01	0.80	2.88	12.50	7.63
M02	1.11	2.25	13.27	7.88
M05	0.96	2.44	14.76	9.01
M10	1.04	2.44	14.94	8.76
M10b	1.10	2.07	11.65	9.13

Table 7. Word error rates for native Finnish speech.

4. MULTILINGUAL SPEECH SYNTHESIS

Considering the proper pronunciation of foreign words as the speech synthesis equivalent of the accurate recognition of non-native speech, a common phonetic alphabet and multilingual acoustic models suggest themselves also for the use in unit selection based speech synthesis. So far, we have applied this idea mainly to system construction, which includes the automatic sub-phonemic labeling of the voice data and the training of binary decision trees for the definition of acoustic contexts and prosodic target values [10, 11].

During the creation of the German version of the synthesizer a better agreement between the speaker's actual pronunciation and the lexicon was achieved by the introduction of nine English phonemes and the use of multilingual seed HMMs in the alignment procedure. Since not yet produced by the German front end during runtime, the positive effect of these changes is mainly due to the avoidance of segmental errors in native German words; however, the new introduced phones are also accessible for the proper pronuncia-

tion of English loan words via exception dictionaries.

More recently, we have started to extend this approach to the construction of a bilingual English/German speech synthesizer. Since, for example, the distinction between long and short vowels clearly is important for speech synthesis, we used the more detailed common phonetic alphabet (cf. Section 2) for that purpose. Based on a small common data base of approx. 2.25 hours of speech (English: 1.25 h, German: 1.0 h) produced by a non-professional, native German male speaker, the system can switch almost arbitrarily (not yet within a word) between two separate linguistic front ends, and is thus able to synthesize mono-lingual as well as mixed-lingual speech.

Speaker dependent bilingual acoustic models turned out to produce more accurate alignments than mono-lingual models, and — consequently — informal listening tests unveiled almost no degradation when comparing the bilingual synthesizer and its two mono-lingual counterparts, when acoustic contexts are determined by use of a decision tree with a relatively small number of leaves.

	lang. dep. leaves [%]		splice rate [%]	
leaves	English	German	English	German
8000	18.8	20.0	19.1	20.0
12000	21.7	23.3	21.2	21.5
24000	27.5	29.2	23.9	25.7

Table 8. Acoustic tree size vs. percentage of language dependent contexts and splice rate (averaged over 50 sentences) for bilingual synthesis.

Table 8 illustrates that a careful control of the acoustic tree size is essential for the production of good quality speech output in both languages: while attempting to provide a better discrimination between both languages by creating a larger number of sub-phonemic contexts, larger acoustic trees also tend to produce an increased number of splices, which are known to be a major source of noticeable distortions in concatenative speech synthesis.

5. CONCLUSION AND FUTURE WORK

In this paper we reported on the use of two different common phonetic alphabets for multilingual speech recognition and synthesis, and gave experimental results that confirm the superiority of multilingual acoustic models for the recognition of non-native speech and words from foreign vocabularies.

Expecting more language and dialect data to be available in the future, we will continue to explore both acoustic modeling techniques (e.g. articulatory features [12]) and the integration of these languages into our common phonetic alphabets. For the latter, the focus of our work will be on the development of alphabets that can be used jointly for recognition and synthesis.

Considering the work described here as initial steps towards a seamless multilingual conversational system, further activities will also deal with the creation of multilingual front ends for speech synthesis (cf., for example, [13]) as well as with the development of multilingual tools for natural language understanding.

6. REFERENCES

[1] R. Siemund *et al*, "OrienTel – Multilingual access to interactive communication services for the Mediterranean and the Middle East," in *Proc. of the 3rd Int. Conf. on Language Resources & Evaluation*, Las Palmas, Spain, 2002.

[2] T. Schultz and A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition," *Speech Communications*, vol. 35, 2001.

[3] V. Fischer, E. Janke, and S. Kunzmann, "Likelihood Combination and Recognition Output Voting for the Decoding of Non-native Speech with Multilingual HMMs," in *Proc. of the 7th Int. Conf. on Spoken Language Processing*, Denver, Colorado, 2002.

[4] C. Traber *et al*, "From Multilingual to Polyglot Speech Synthesis," in *Proc. of the 6th Europ. Conf. on Speech Communication and Technology*, Budapest, 1999.

[5] L. Mayfield Tomokiyo, A. Black, and K. Lenzo, "Arabic in my Hand: Small-footprint Synthesis of Egyptian Arabic," in *Proc. of the 8th Europ. Conf. on Speech Communication and Technology*, Geneva, 2003.

[6] M. Ostendorf and I. Bulyko, "The Impact of Speech Recognition on Speech Synthesis," in *Proc. of the IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, Ca., 2002.

[7] F. Palou Cambra *et al*, "Towards a common alphabet for multilingual speech recognition," in *Proc. of the 6th Int. Conf. on Spoken Language Processing*, Beijing, 2000.

[8] C.J. Wells, "Computer-coded Phonemic Notation of Individual Languages of the European Community," *Journal of the International Phonetic Association*, vol. 19, pp. 32–54, 1989.

[9] V. Fischer *et al*, "Towards Multilingual Acoustic Modeling for Large Vocabulary Continuous Speech Recognition," in *Proc. of the IEEE Workshop on Multilingual Speech Communications*, Kyoto, Japan, 2000.

[10] R. Donovan, "Trainable Speech Synthesis," PhD. Thesis, Cambridge University Engineering Department, 1996.

[11] E. Eide *et al*, "Recent Improvements to the IBM Trainable Speech Synthesis System," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Hong Kong, 2003.

[12] S. Stüker, T. Schultz, F. Metze, and A. Waibel, "Multilingual Articulatory Features," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Hong Kong, 2003.

[13] B. Pfister and H. Romsdorfer, "Mixed-lingual Text Analysis for Polyglot TTS Synthesis," in *Proc. of the 8th Europ. Conf. on Speech Communication and Technology*, Geneva, 2003.