TOWARDS A COMMON PHONE ALPHABET FOR MULTILINGUAL SPEECH RECOGNITION

F. Palou, P. Bravetti, O. Emam, V. Fischer, E. Janke

IBM Voice Systems, European Speech Research,
Avenida Republica Argentina 25, E-41011 Sevilla, Spain
palou@es.ibm.com

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

ABSTRACT

New automatic speech recognition applications, mainly for small and medium vocabulary sizes, demand the capability of recognizing speech in several languages simultaneously. We have started exploring the possibility of building acoustic models that integrate multiple languages (up to seven in the initial stage), using speech transcriptions based on a common phoneme alphabet across all the languages. To reach a common alphabet, we start from the previously existing alphabets for each one of the seven languages. We first proceed to simplify some of them, partially following SAMPA transcription guidelines, and then to merge phones present in several languages that correspond to the same IPA symbol. We study and compare two variants of the common phoneme alphabet. The first of these two alphabets is closer to the starting ones, and includes the use of diphthong phones for English and German, and long-vowel phones for Arabic, English, and German. The second one avoids the long-vowel and diphthong phone models, and also the stressed vowel models. We present and discuss the results of decoding large vocabulary dictation tests, comparing the two alphabet variants, and also the multilingual decoding results with the corresponding monolingual acoustic models.

1. INTRODUCTION

The pronunciation representation in speech is a key necessary ingredient in the preparation of acoustic model builds for automatic speech recognition. In our past experience we have usually approached this problem in the framework of given a language, or even reducing the scope to a dominant dialect of the language, or other specific dialects.

But there are more and more reasons to start dealing with two or more languages in the same acoustic models. Some of the first attempts aimed at couples of languages spoken in the same area, like English and French in Canada [5]. In such situations, the practical advantages are obvious. There are other situations where application domains do not require large vocabularies, and efforts to bring together a larger number of languages have been undertaken [2, 4].

The work in this paper is part of an effort to extend the number of languages, and possible reach a coverage of wider areas, like the European Union. In particular, we have started building acoustic models with seven languages, which are Arabic, British English, French, German, Italian, Brazilian Portuguese, and Spanish.

When the number of languages grows, we need a general framework that allows us to code word pronunciations in a systematic way. We also need, for practical reasons, to work with coding rules that are reasonably intuitive. Our efforts in this direction are the main subject of this paper. The remainder of the paper is organized as follows: Section 2 will discuss our initial common alphabet (designed as CPA-1), and how it relates to the previously existing monolingual alphabets of the seven contributing languages. Section 3 will discuss an alternate version of a common alphabet (designed as CPA-3), with a reduction, compared to the first one, in the number of phonetic units that are needed. This reduction helps increasing the phonetic coverage of additional languages, and helps making the coding of pronunciations more intuitive. Section 4 will present some results on the performance of acoustic models that we have built using both phonetic alphabets, including monolingual acoustic models, and seven-language acoustic builds. Section 5 presents our conclusions and some comment on the direction of future work.

2. INITIAL ALPHABET (CPA-1)

The starting point was the already existing alphabets for each one of the seven languages, with a total of 132 phone models for vowels or diphthongs and 224 phone models for consonants. We proceeded to compare them, for most of these languages, with SAMPA [6] transcription guidelines and alphabets, available in the SAMPA web pages, in order to introduce some language-dependent simplifications, reducing the detail of phonetic transcription. In the case of English, however, new diphthong models were introduced at the same time. Table 1 shows the sizes of these phonetic alphabets, before and after these simplifications.

In the previous process, we also assigned IPA [1] phonetic alphabet symbols, by means of their SAMPA representations, to the majority of the phonetic units coming from the language-dependent alphabets. This will enable us to

(a)	total	En	Fr	Gr	It	Es	Pt	Ar
vowel	132	18	17	23	22	14	24	14
cons.	224	31	19	37	48	35	24	30
total	356	49	36	60	70	49	48	44

(b)	total	En	Fr	Gr	It	Es	Pt	Ar
vowel	118	20	17	23	14	10	20	14
cons.	182	24	19	26	32	30	22	29
total	300	44	36	49	46	40	42	43

Table 1: Number of vowel and consonant models in the phonetic alphabets of seven languages: British English (En), French (Fr), German (Gr), Italian (It), Spanish (Es), Brazilian Portuguese (Pt), and Arabic (Ar): (a) for the previously used phonetic alphabets; (b) for the simplified alphabets integrated into CPA-1.

proceed to the following step, consisting in merging those units that have been mapped to the same IPA symbol, and in this way build a more compact phonetic alphabet (121 phonetic units, 65 of them for vowels or diphthongs, and 56 for consonants), that applies to all seven languages. Table 2 shows that there is, for several reasons, in each language, a part, relatively small, of the phonetic units, which is not shared with the other languages.

	total	En	Fr	Gr	It	Es	Pt	Ar
vowel	30	9	3	6	–	–	6	6
cons.	23	–	1	2	8	4	–	8
total	53	9	4	8	8	4	6	14

Table 2: Number of vowel and consonant models in CPA-1 that belong only to one of the seven languages.

The number of consonant models is compressed by a factor of 3.25, to be compared to 1.82 for the vowels. The picture of the vowel area is much more complex because we distinguish short and long vowels (English, German, Arabic), stressed vowels (Spanish, Italian, Portuguese), nasal vowels (French, Portuguese, German), and include models for some diphthongs (English, German), and a few nasal diphthongs (Portuguese). The details of this picture can be found in Table 3.

(a)	total	En	Fr	Gr	It	Es	Pt	Ar
short	24	7	13	10	5	5	5	7
long	14	5	–	8	–	–	–	7
nasal	7	–	4	2	–	–	5	–
stressed	9	–	–	–	9	5	7	–
(b)	total	En	Fr	Gr	It	Es	Pt	Ar
oral	8	8	–	3	–	–	–	–
nasal	3	–	–	–	–	–	3	–
total	65	20	17	23	14	10	20	14

Table 3: Different nature of the CPA-1 vowel (a) and diphtong (b) phone models, depending of the language.

3. REDUCED ALPHABET (CPA-3)

Here we have introduced some changes in the coding of the consonants, but a more severe change in the coding of the vowels. All the diphthong models have been dropped, after splitting them into two short vowel models. Most of the long vowel models also become sequences of two identical short vowel models. Experimental results for English and German, presented in Section 4, to some extent back up these changes. Also the stressed vowel models, which introduced asymmetry between languages, have been dropped. Table 4 shows some statistics across languages for phonetic alphabet CPA-3, which are to be compared to the contents of Tables 1 and 3 corresponding to CPA-1.

(a)	total	En	Fr	Gr	It	Es	Pt	Ar
vowel	31	13	15	17	7	5	12	11
cons.	45	24	19	23	28	24	22	28
total	76	27	34	40	35	29	34	39

(b)	total	En	Fr	Gr	It	Es	Pt	Ar
vowel	14	3	2	2	–	–	3	4
cons.	16	–	1	1	6	–	–	8
total	30	3	3	3	6	–	3	12

Table 4: (a): Number of different vowel and consonant models used in CPA-3 (total number, and by language). (b): Non-shared models (to be compared with Table 2).

Table 5 shows the new picture of the different types of vowels, to be compared to Table 3. The compression factor for vowels is now 3.81 (relative to the original 118 language-dependent vowels), very close to the 4.04 value for consonant (relative to 182 language-dependent consonants).

(a)	total	En	Fr	Gr	It	Es	Pt	Ar
short	21	12	11	15	7	5	7	9
long	3	1	–	–	–	–	–	2
nasal	7	–	4	2	–	–	5	–
stressed	–	–	–	–	–	–	–	–
(b)	total	En	Fr	Gr	It	Es	Pt	Ar
oral	–	–	–	–	–	–	–	–
nasal	–	–	–	–	–	–	–	–
total	31	13	15	17	7	5	12	11

Table 5: Different nature of the CPA-3 vowel (a) and diphthong (b) phone models, depending of the language.

This reduced alphabet (CPA-3) could be applied, for example, to Dutch, without needing to add new phonetic units for (SAMPA coded) diphthongs "Ei", "9y", and "Au", as would be the case for CPA-1 alphabet. The reduced complexity of this alphabet also makes the coding of pronunciations more intuitive, and less prone to errors.

4. DECODING EXPERIMENTS

The study of many different multilingual acoustic models that have been built using these common phonetic al-

	En	Fr	Gr	It	Es	Pt	Ar
CPA-1	11.2	9.7	9.9	9.2	4.8	9.6	17.1
CPA-3	10.7	13.6	8.9	11.0	5.8	10.3	19.3

Table 6: WER results of decoding test sets for equivalent monolingual acoustic models (LDA) when using phonetic alphabets CPA-1 and CPA-3.

phabets is the subject of separate publication [3]. This reference gives more details on the characteristics of the models that have been used for the decoding experiments described here. Each language provided similar amounts of training data, ranging between 12 and 20 hours of speech. The acoustic models considered here were built either using the whole data set (multilingual acoustic models) or only the part corresponding to one of the languages (monolingual acoustic models), using similar procedures (including a linear discriminant analysis method, LDA), except for the differences in the pronunciation representation when the phoneme alphabet chosen is different. Each language provided a decoding vocabulary (between 22,000 and 100,000 words) and a set of test speakers reading medium-perplexity scripts. The word error rates in Tables 6 and 7 correspond to these decoding tests, using four different acoustic models for each language's set of tests.

Table 6 shows the collected word error rates corresponding to the decoding tests performed with seven CPA-1 and seven CPA-3 monolingual acoustic models. Rather surprisingly, CPA-3 performs better for English and German, in spite of the diphthong and long vowel splits, while CPA-1 performs better for the other languages.

Table 7 shows the collected word error rates corresponding to the decoding tests performed with one CPA-1 and one CPA-3 multilingual acoustic models, the same one, in each case, for all seven languages. In this case CPA-3 performs better for Brazilian Portuguese, and CPA-1 performs better for the remaining languages. The averaged word error rate for CPA-3 reaches 107% of the corresponding value for CPA-1, which represents a rather small difference in this situation.

	En	Fr	Gr	It	Es	Pt	Ar
CPA-1	26.0	17.8	13.7	13.9	9.4	16.2	22.0
CPA-3	26.9	20.6	14.8	15.6	9.5	15.9	24.3

Table 7: WER results of decoding test sets for equivalent multilingual acoustic models (LDA) when using phonetic alphabets CPA-1 and CPA-3.

While the results in Table 6 give useful information for the comparison of both phonetic alphabets, those in Table 7 are more relevant to analyze our progress towards building multilingual acoustic models that can be exploited in real applications, summarized in the following section.

5. CONCLUSIONS

The comparison of multilingual decoding results (Table 6) with the corresponding monolingual results (Table 7) indicates that switching to such multilingual acoustic models, for the complex tasks that we have tested, leads to an important loss of accuracy. We need to understand better what are the reasons that lead to those results, in order to get accuracy improvements in the future. But the absolute scale of the multilingual error rate give us already hopes that important application areas will soon benefit from the use of multilingual acoustic models. The performance comparison between the CPA-1 and CPA-3 alphabets gives a practical draw between them. Other considerations, already exposed, regarding the easier coding of pronunciations, give some preference to the reduced CPA-3 alphabet, and especially taking into account that in the future the multilingual alphabet will have to grow, as we intend to incorporate new languages into the picture, which cannot be fully transcribed without adding some more phonetic units.

Acknowledgment. We would like to thank our colleagues in the IBM European Speech Research Group (located in Cairo, Heidelberg, Hursley, Paris, Rome, and Seville) and in the IBM Human Language Technology Research Group (Thomas J. Watson Research Center, Yorktown Heights) for many valuable suggestions and the continuous exchange of ideas.

6. REFERENCES

1. International Phonetic Association. *Handbook of the International Phonetic Association*. Cambridge University Press, Cambridge, 1999.
2. P. Bonaventura, F. Gallocchio, and G. Micca. Multilingual Speech Recognition for Flexible Vocabularies. In *Proc. of the 5th European Conference on Speech Communication and Technology*, pages 355–358, Rhodes, Greece, 1997.
3. V. Fischer, J. Gonzalez, E. Janke, M. Villani, and C. Waast-Richard. Towards Multilingual Acoustic Modeling for Large Vocabulary Continuous Speech Recognition. In *Proc. of the MSC2000 Workshop on Multilingual Speech Communications*, Kyoto, 2000, *to appear*.
4. J. Koehler. Language Adaptation of Multilingual Phone Models for Vocabulary Independent Speech Recognition Tasks. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Seattle, 1998.
5. T. Ward, S. Roukos, C. Neti, J. Gros, M. Epstein, and S. Dharanipragada. Towards Speech Understanding Across Multiple Languages. In *Proc. of the 5th Int. Conf. on Spoken Language Processing*, Sydney, 1998.
6. J. Wells, W. Barry, M. Grice, A. Fourcin, and D. Gibbon. Standard Computer Compatible Transcription. ESPRIT project 2589 (SAM), Doc. no. SAM-UCL-037, 1992.