A MULTILINGUAL ACOUSTIC-PHONETIC MODEL BASED ON PHONE-TO-PHONE TRANSITION CLASSES

Giorgio Micca[1], Enrico Palme[2], Juan Mari[3]
[1] *CSELT, Turin, ITALY*
[2] *Universita` di Pisa, ITALY*
[3] *Universidad Politecnica de Catalunya, Barcelona, SPAIN*

ABSTRACT

The paper presents an acoustic-phonetic-model based on classes of transitions between phones. The model is to be used in the development of speech recognition systems for multilingual applications. The problem that has been tackled is twofold: how to deal with application vocabularies containing multiple language entries and how to cope with scarce training material in a new target language. The Italian, Spanish and British English languages have been considered in this work. Monolingual models, based on phone transition classes, show recognition performance comparable or even better to alternative models requiring larger inventories of units. Results for a three-lingual model, consisting of a global set of acoustic-phonetic classes where similar transition sounds for the three languages have been unified, are also reported. The approach is being experimented in "blind" mode with the Rumanian language, where the phonological structure of this language is the only information source feeding the recognition model.

1. INTRODUCTION

A consolidated trend in the field of Automatic Speech Recognition (ASR) is the search for higher precision in the acoustic modeling of speech patterns; for instance, the concept of context-dependence of phones has been extended to cover larger and larger ranges (polyphones). The aim is to increase the accuracy of the representation of intra- and inter-word co-articulation phenomena. These efforts have produced noticeable improvements in terms of recognition performance, at the cost of heavier computational load and larger memory requirements. The complexity is proportionally augmented as many languages are to be simultaneously processed by the ASR system, and the application vocabulary consists of words belonging to different languages. This happens, for instance, in voice activated enquiry systems providing information on international travel or finance services. Furthermore, multilinguality also impacts the economy of ASR systems in applications requiring efficient procedures for cross-language transfer of speech technology. The goal here is to avoid or limit the burden of collecting huge amounts of speech data for each of the new languages to be included in a multilingual service or product.

Several different techniques have already been developed for cross-language portability of speech recognition models. In most cases, the starting point is represented by the search for similarities among sounds of different languages. After the pioneering work by Wheatly and al. [1] and the introduction of the concept of poly-phonemes [2], experimented with four European languages (Dutch, British English, German and Italian), several other approaches followed where different combinations of acoustic density clustering and cross-language phonetic lexica mappings were designed and experimented. In [3], we presented an approach similar to [4], and we developed a context-independent multilingual phoneme inventory covering Italian, English, Spanish and German, based on a combination of HMM (Hidden Markov Model) distance measures introduced to compute similarities of acoustic-phonetic units belonging to multiple languages. We also showed how these similarities can be exploited to interpolate acoustic models for a new, undertrained language. In [5] a language-independent approach was attempted, by combining up to eight languages in a global set of polyphones and then by using this model for cross-language transfer purposes. The procedure performs well in the target language, but the large size of the phonetic unit inventory (a few thousands) is an obstacle to a full generalization of the approach towards the direction of language independence. In this paper, we extend the method presented in [3] to transitory units.

2. THE METHOD

2.1 Transitory units

Stationary and Transitory units [7] explicitly represent the central, more stable section of phone realizations and the transition from one phone to the adjacent one. For instance, in Italian, the word "bene" ("well"), */b'ɛ n e/*, is transcribed as

 # #b b b'ɛ 'ɛ 'ɛn n ne e e# #

where odd components represent stationary events - # is the silence - and even components represent transitions. *#b* and *e#* are the positional units at the beginning and at the end of the word. In this case, for instance, *b* is the voice bar and *b'ɛ* represents the transition to the following front vowel. This structure has proven to perform well as far as all the units in the inventory can be properly trained. It may happen that the occurrence frequency of some units is below a minimum threshold in a given language-specific training database; in this case those units would be undertrained. In fact, the minimal statistical coverage requirement can be challenged by the scarceness of data for rare sounds. Furthermore, this drawback is highly emphasized when we look for a global model suitable for multiple languages.

In our method, phones are classified in classes, similar sounds are merged into one class, then these classes are used to build up the set of phone-to-phone transitory classes. Resuming the previous example, the word "bene" can be transcribed as

SL SLVP b VPFV ’ɛ FVNA n NAFV e FVSL SL

where *SL*, *VP*, *VF*, *FV* and *NA* correspond to “silence”, voiced plosive, front vowel and nasal phonetic classes. In [3], the acoustic model was based on context-independent units, therefore all the sounds whose cross-language distance resulted to be below a given threshold were merged into a single class. In the method presented in this paper, classes are introduced only for transitory units. Stationary units are not clustered because they convey the information on the lexical identity of a word. The phonological structure of a given language is preserved, and it is therefore maintained in the multilingual inventory of units. Small classes of transition units preserve a higher degree of acoustic precision than large classes, but reduce the compression factor of multilingual inventories because less units are merged into a single transition class. An optimal trade-off between average size of classes and accuracy of acoustic modeling has to be found in order to guarantee a specified level of statistical robustness - trainability - of units without loosing too many details in the model.

2.2 Classes of transitions

Several models were tested according to the design criteria described in the following. We started developing monolingual inventories, then moved to the multilingual case. In the monolingual experiment, we developed two types of unit sets for each of the three languages.

- **Basic class set,** corresponding to the classical taxonomy of consonant and vowel sounds: voiced an unvoiced plosives, nasals, laterals/vibrants, voiced an unvoiced fricatives, affricates; front, central and back vowels. This method produced the inventories *en-170*, *it-114*, *sp-140* for English, Italian and Spanish respectively.
- **Improved class set,** designed according to similarity measures computed on the HMMs of phones in each language. Measures were based on a metrics introduced in [3], where up to five different algorithms are applied to compute the acoustic similarity of the sounds of a language. The phone hierarchy derived from this computation is represented by a dendogram, an example is given in Fig. 1 for English. The data-driven method is as follows: for a given transition of type *xy*, the corresponding transitory class is identified by combining the information provided by the dendogram of both left (*x*) and right (*y*) constituent phones. In this stage, the absolute values of distance measures are taken into account. Therefore transitional phonetic classes are generated in data-driven mode, and nearly no phonetic, a priori knowledge is considered as an information source. Two specific classes were introduced for the closure section of plosives (silence or voice bar). The generation of the improved class was carried out in two successive steps. In a first step, the procedure was separately performed for each language and the corresponding HMMs for the transitory and stationary units so obtained were trained. In a first stage, for the English language, we designed the transitory unit classes according to a priori phonetic criteria, and generated the set *en-363-mon*. This model was therefore similar to model *en-170-mon*, but resulted in a finer and more detailed phonetically motivated distinction of classes. Since this model did not yield satisfactory improvements in recognition performance, we moved to the data-driven approach, which produced the set *en-358-mon-dd*. The other two

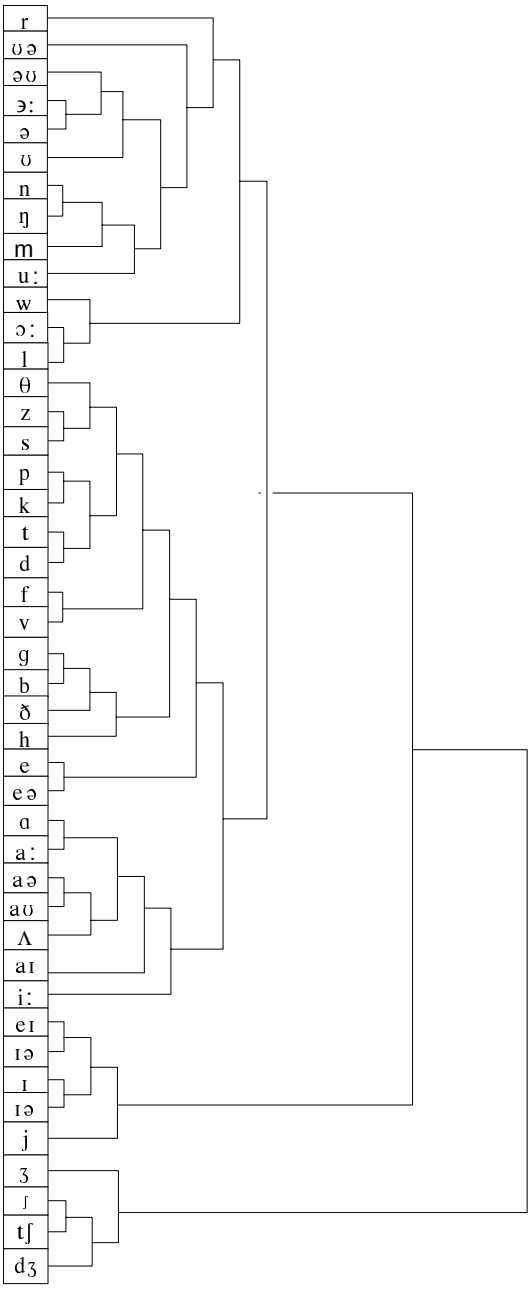

Fig.1. Dendogram for English sounds.

data-driven inventories for Italian and Spanish were it-*220-mon-dd* and *sp-269-mon-dd* respectively. Finally, we obtained the global inventory for the multilingual, multivocabulary model: *mul-670-mul*. It consisted of the combination of the three language-specific sets where classes of different languages, representing cross-language sounds which could be clustered according to the distance measures, were unified. Also closure silences and voice bars were unified across the languages. This cross-language unification operator is represented by the symbol $\oplus$ in the following formula:

mul-670-mul=en-358-mon-dd⊕ it-220-mon-dd⊕ sp-269-mon-dd

3. EXPERIMENTS

3.1 Speech Databases

Training and test databases used in the experiment consisted of a portion of the SpeechDat databases [8] for Italian and Spanish, while the training English component was collected by CCIR–University of Edimburgh. The size of the databases is given Table 1. Two test data sets were used for English: one from SpeechDat and the other one from CCIR.

	ENG		ITA		SPA	
	Train.	Test	Train.	Test	Train.	Test
# utt.	34400	1797	12800	1050	5174	1730

Table 1. Training and Test corpora.

3.2 Initialization of transitory HMMs

Two different bootstrapping methods for transitory units were implemented: coarse (c) and fine (f) initialization. With c-initialization, the left state of a transitory unit is given the density function of the rightmost state of the context-independent HMM, represented by a left-to-right, three state topological structure, corresponding to the left component of the transitory unit. The same process is followed for the right state of the transitory unit (Fig. 2). Stationary units are assigned the density function of the central state of the corresponding context-independent model. With f-initialization, a Viterbi segmentation of training data is performed using the context-independent three-state models. Acoustic sequences, segmented by the rightmost state of the HMM that correspond to the left component of a transitory unit, are assigned to the left state of this unit; a similar relationship holds for the right state and for the stationary unit. Finally, all segments insisting on a given state are processed by a clustering procedure to derive the Gaussian mixture of the state. The process is iterated on all states of the transitory/stationary unit inventory. Since the segmentation is consistent regarding phonotactic constraints (e.g. the leftmost state of transition *xy* is associated to segments of the rightmost state of phone *x* only in contexts where the successive phoneme is *y*), the results is a more precise bootstrap representation of the transitory/stationary units. Anyway, f-initialization requires longer computing time than c-initialization.

3.2 Experimental results

Continuous Density HMMs of acoustic-phonetic units were trained by the K-means algorithm. Each HMM state was represented by a variable mixture density function with up to 32 Gaussians per mixture. The Viterbi decoder generated the N-best scored hypotheses with beam search acceleration. English was the working language for tuning and testing the method; the optimal choices were then extended to Spanish and Italian. Finally the multilingual unit inventory was generated. The multilingual tests were performed on a 535 words vocabulary (475 Italian, 30 English and 30 Spanish). A separate test set for English consisted of a list of 300 railway and underground stations.

3.2.1 Monolingual experiments

The baseline model for English was *en-170-mon*. We tried both c- and f-initialization. Since the latter performed significantly better than the former – Word Recognition rate (WR) of 92.14

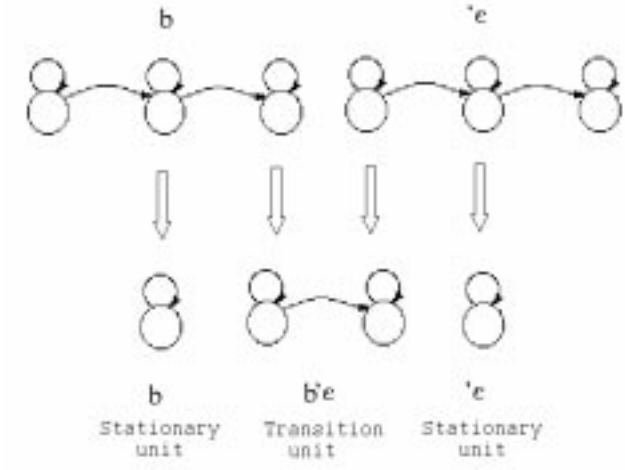

Fig.2. C-initialization of Transition and Stationary units.

compared to 91.59 - we decided to adopt f-initialization in all the successive experiments. The next model, *en-363-mon*, which included quite a larger amount of phonetic knowledge (Section 2.2), brought about only a limited improvement in recognition performance, 7% of Error Reduction rate (ER). Then we moved to the next model, where the new source of information, the distance metrics, was taken into account in the definition of the inventory of phonetic class transitions. Several re-assignations of phones to classes were induced by this procedure. Table 2 shows the different allocation in classes for plosive and some fricative/affricate sounds. The new model significantly outperformed the previous one (WR = 94.21, ER = 20.7), indicating that data-driven criteria can be exploited in the optimization of this type of acoustic-phonetic units.

en-363-mon	*en-358-mon-dd*
d t	**d t**
p b	**p k**
k g	**g**
	b
ʃ ʒ dʒ	**ʃ tʃ dʒ**
tʃ	**ʒ**

Table 2. Different partitions of English sounds in classes.

The method was applied to Spanish and to Italian languages; this time we directly applied the distance criteria. The error reduction observed with respect to the baseline models was 38.9% for Spanish and 38.8% for Italian (Table 3).

ENG	*en-170-mon*	*en-363-mon*	*en-358-mon-dd*
WR	92.14	92.70	94.21
SPA	*sp-140-mon*	-	*sp-269-mon-dd*
WR	95.55	-	97.28
ITA	*it-114-mon*	-	*it-220-mon-dd*
WR	84.54	-	90.54

Table 3. WR performance of different models.

3.2.2 Multilingual experiments

The multilingual phonetic inventory *mul-670-mul* designed according to the method presented in Section 2 was trained by means of the super-corpus obtained by merging the English, Italian and Spanish training corpora. An example of cross-language alignment of sounds is given in Table 4 for some nasal consonants.

	ENG	ITA	SPA
NA Class	**m ŋ n**	**m n**	**m n ɲ**
GNI Class		**ɲ**	

Table 4. Classes for nasal sounds.

Interestingly enough, the Italian sound **ɲ** is not assimilated to the corresponding sound in Spanish, but it is left apart as a single member phonetic class.

These models were tested in two different modes: *monovocabulary* and *multivocabulary*. In the former mode, the test was carried out separately for each language. In the latter mode, all the words of each of the monolingual test vocabularies were merged in a global test vocabulary. The aim of this test was twofold: to probe the preservation of language-specific accuracy of the multilingual models and to evaluate the extent these models might support a multilingual vocabulary recognition task. Results are given in Table 5.

	Test Language		
mul-670-mul	**ENG**	**SPA**	**ITA**
Monovocabulary	**94.26**	**98.44**	**90.17**
Multivocabulary (475 ITA+ 30 SPA+30 ENG)	**90.03**	**89.19**	**89.97**

Table 5. WR performance of multilingual models.

The test in Spanish resulted in a significantly better result than the corresponding monolingual model (98.44 WR compared to 97.28). This effect was explained in terms of a greater robustness of multilingual models than the monolingual models, specifically observed in the test with Spanish utterances. In fact the multilingual models took advantage of the larger size of the multilingual training set, and the Spanish language included the smallest language-specific training corpus of the three languages. The tests in the other two languages did not show relevant deviations from the results observed with the corresponding monolingually trained models. A further test was carried out for English, where a new test set was considered, consisting of 800 utterances with a vocabulary of 300 railway and underground stations. The results was WR = 95.23, which is consistent with the figure of 94.68 which was obtained with language-specific models.

A second series of tests was designed and carried out aiming at evaluating the capability of the multilingual models to strengthen the recognition models of a poorly trained recogniser. To this purpose, we selected about 10% of the training set for Spanish, taking care of including a balanced proportion of male and female speakers. The resulting subset consisted of 517 utterances. WR results are reported in Table 6. This test clearly points out the effect of strengthening of models for the Spanish language due to the contribution of the training material of the other two languages.

	sp-269-mul	Training 557 utterances
SPA	**96.13**	**91.08**

Table 6. Strengthening of poorly trained Spanish models.

The approach is being evaluated in a cross-language recognition model transfer task involving the Rumanian language. The HMMs of the unseen language will be interpolated by mapping its phonological structure into the multilingual set of acoustic-phonetic units described in the paper. In this case, since no acoustic data are available in the target language, also the stationary components will have to be bootstrapped from the stationary constituents of the multilingual model. Eventually the HMMs for the new language will be improved by including a limited portion of Rumanian utterances in the multilingual model.

4. CONCLUSIONS

A method for designing multilingual acoustic-phonetic models for automatic speech recognisers has been presented. The approach extends the concept of phone-to-phone transitions in a given language to multiple languages, where similar sounds are represented by a class of transitions. The procedure increases its efficiency and generality as new languages are added to the model. Experiments with a three-lingual recogniser for English, Spanish and Italian languages outline the capability of the model of combining acoustic precision and generalization towards the direction of language independence. The approach is being experimented in a cross-language transfer of acoustic-phonetic knowledge for the Rumanian language.

ACKNOWLEDGMENTS

The authors are indebted to Luciano Fissore for useful suggestions in the experimental stage of this work.

REFERENCES

[1] Wheatley B., Kondo K., Anderson W., Muthusami Y. 1994. An evaluation of cross-language adaptation for rapid HMM development in a new language. *Proceedings of ICASSP'94*, Adelaide, pp. I-237, I-241.

[2] Andersen O., Dalsgaard P., Barry W. 1993. Data-driven identification of poly- and mono-phones for four European languages. *Proceedings of EuroSpeech'93*, Berlin, pp. 759-762.

[3] Bonaventura P, Gallocchio F., Micca G. 1997. Multilingual Speech Recognition for Flexible Vocabularies, *Proceedings of EuroSpeech'97*, Rhodes, pp.355-358.

[4] Köhler J. 1996. Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds. *Proceedings ICSLP'96*, Philadelphia, pp. 2195-2198.

[5] Schultz T., Waibel A. 1998. Language Independent and Language Adaptive Large Vocabulary Speech Recognition, *Proceedings of ICSLP'98*, pp. 1819-1822.

[6] Schultz T., Waibel A. 1998. Adaptation of Pronunciation Dictionaries for Recognition of Unseen Languages. *International Workshop on Speech and Computer*, St. Petersbourg, october 26-29, pp. 207-210.

[7] Fissore L., Ravera F., Laface P. 1995. Acoustic-phonetic Modeling for Flexible Vocabulary Speech Recognition. *Proceedings of EuroSpeech'95*, Madrid, pp. 799-802.

[8] http://speechdat.phonetik.uni-muenchen.de/, EU Project LE2 4001 SpeechDat.