MULTILINGUAL SPEECH RECOGNITION FOR FLEXIBLE VOCABULARIES

Bonaventura P.[1], *Gallocchio F.*[2] *and Micca G.*[3]

[1] CSELT Consultant, Turin, Italy
[2] Dipartimento di Elettronica e Informatica, Università di Padova, Italy
[3] CSELT, Turin, Italy
patrizia.bonaventura@cselt.it, filippo@luna.cselt.it, giorgio.micca@cselt.it

ABSTRACT

The paper addresses the problem of designing a speech recogniser for multilingual vocabularies. The goal of the research is twofold: future Interactive Voice Recognition (IVR) systems, like a speech activated flight information service, are likely to require multilinguality as a major feature; besides, a general language-independent phonetic inventory might be very useful in bootstrapping phonetic models for a new language for which insufficient training data are available. Metrics were introduced in order to measure cross-language phonetic dissimilarities, and a multilingual phonemic inventory was created. Experiments were run on a speech database including Italian (I), Spanish (S), English (E) and German (G) words. Results clearly show that it is possible to reduce the complexity of a multilingual phonetic recogniser by exploiting phonetic commonalities across different languages, without significant losses in WA for multilingual tasks with respect to single language recognition tasks.

1. INTRODUCTION

One of the most recent trends in the field of Automatic Speech Recognition (ASR) for flexible vocabularies is related to the definition of multilingual phonetic inventories by exploiting similarities among sounds of different languages. A first attempt was carried out in [1] where poly-phonemes were identified across languages (Danish, British English, German, and Italian) on the basis of a data-driven clustering technique; successively, this approach has been extended to a language identification task [2]. A different technique was adopted in [3], where an acoustic-phonetic modeling, based on a method to determine sounds similarities across the E, G and S languages was introduced, by considering language-dependent as well as language-independent properties, using a density-clustering algorithm. Recently the issue of cross-language portability was approached by means of a general integrated language-independent recogniser [4]; also, a classical Bayesian task adaptation procedure was applied to derive a seed model for Slovene language starting from a multilingual HMM recogniser trained on G/S/E speech data [5].

The interest in cross-language transfer of speech technology is growing [6], since the recent state of the art requires the collection of a new speech database for any new language to be added to the system, plus the adoption of language-specific training and test procedures. The commercial diffusion of IVR systems is rapidly impacting the demand of multivocabulary recognition from one side, and of fast and feasible procedures for adaptation to a new language with little or no available data on the other side. A conceivable example for an application of such a system is a flight information service for European cities. In such an application the target vocabulary typically consists of a few hundred city names of different linguistic origin, some of which might have different names in various languages. Besides, the more restricted diffusion of quite a few languages in Europe might limit the extension of IVR applications to these languages. Therefore, there is an increasing need for cross-language portability of existing recognisers, so that the realization of a language-independent device has been attempted by several different techniques.

In our approach, a number of different metrics for measuring similarities among cross-language phonetic models was implemented and compared: weighted Euclidean distances, common hyperspace volume of multivariate Gaussian densities, best match computed as the nearest Gaussian in a mixture, furthest neighbour, multigaussian and monogaussian Bhattacharyya distance [7][9] and the entropy-based information loss [8].

A common phonetic space for I, S, E and G has been identified by a proper combination of the five best metrics. Experiments have been run by using the standard CSELT technology for open vocabulary recognition, based on sub-word unit CDHMMs with variable resolution of the acoustic space.

2. METHOD

2.1. Dissimilarity measures

Dissimilarity measures were subdivided in two classes, according to whether they apply to multi- or to mono-gaussian HMM distributions. In the latter case, the mixture was replaced by a single monogaussian distribution having the same mean and variance of the mixture. Three measures were based on the Bhattacharyya distance [7][9]; the first one applies to multigaussian mixtures, the other two apply to monogaussian distributions. For this latter case the HMM distance is computed either as a sum of each single state distance or by merging the state-dependent gaussians into a single 72 (24x3)-dimensional distribution. The fourth metrics was based on the information loss (il) computed as the entropy variation due to the merge of two models (1):

$$d_{il}(M_1,M_2)= (H_{M_1} + H_{M_2}) - 2H_{M_1 \cup M_2} \quad (1)$$

Finally the fifth distance was obtained as the common N-dimensional acoustic space shared by the two gaussian distributions, averaged over the three states. This corresponds, in a two-dimensional space, to the computations of the common area of two mono-dimensional distributions.

A few other metrics were also implemented: Euclidean distance, Furthest Neighbour, etc.. Then a criterion for choosing the most appropriate distances was determined.

To this purpose, Italian sounds have been classified into the following major phonetic categories [10]: voiced and unvoiced plosive, fricative, affricate, liquid, nasal consonants and front, central and back vowels. For each phoneme, the distance from every other phoneme in the inventory was computed for every metrics. A "phonetic coherence index" was obtained by ordering the sounds according to an increasing phonetic distance and by considering the relative position of all sounds belonging to a given class; finally, all distances belonging to the same class were averaged.

Accordingly, the above-mentioned five distances were selected and then merged into an average one (2).

$$d = d_{Bhus} + d_{Bhss} + d_{il} + d_{Bhm} + \ln(d_{cas}) \quad (2)$$

where **Bh** stands for 'Bhattacharyya' (**us** for 'unified states', **ss** for 'separate states' and **m** for 'multigaussian'); **il** stands for 'information loss' and **cas** stands for 'common acoustic space'.

The adoption of the neperian logarithm for the last distance was required to take into account the larger value range of this metrics.

2.2. Selection of multilingual phonemes

Two phonemes of two different languages could be unified if their distance fell below a given threshold, which was empirically set to 5. Then a second constraint was introduced, that the distance has to be minimal within the entire acoustic space (as defined by the union of the acoustic spaces of the two languages). Since this latter constraint was too tight, a trade-off has been obtained, according to (3) and (4); the merging of two models was performed only if one of these formulae was true:

$$d\left(M_1^{L1}, M_1^{L2}\right) < \left(d_{min}^{L1}(M_1) \text{ and } d_{min}^{L2}(M_2)\right) \quad (3)$$

$$d\left(M_1^{L1}, M_1^{L2}\right) < \left(\left(d_{min}^{L1}(M_1) \text{ or } d_{min}^{L2}(M_2)\right) \text{ and } 5\right) \quad (4)$$

where $d\left(M_1^{L1}, M_1^{L2}\right)$ is the distance between a model belonging to the first language and the most similar model of the second language, $d_{min}^{L1}(M_1)$ is the distance between the first model and the most similar model belonging to the same language, while $d_{min}^{L2}(M_2)$ is equivalent to $d_{min}^{L1}(M_1)$ for the second language.

Formulae (3) and (4) were applied in order to decide whether any phoneme of the N-th language could be merged with the multilingual phoneme inventory, previously built on the N-1 languages.

3. EXPERIMENTAL RESULTS

In this section, experimental results are reported, that have been obtained by the phoneme compression metrics described in (2).

Context-independent models composed of 32-gaussian mixtures have been used. Training has been performed on the following speech databases: SpeechDat (German and Spanish), DbCcir (English), Panda (Italian). The approximate size of these databases was of 4200(G), 2500(S), 11500(E) and 6000(I) utterances respectively. Vocabularies contained 98, 300, 70 e 59 isolated words respectively for I, E, S and G.

The four sets of experiments of multilingual recognition were the following: 1) I-E, 2) I-S, 3) I-E-S, 4) I-E-S-G. Every set was made of a pair of experiments, the first one running on a monolingual vocabulary and the second on a multilingual one; each experiment in the set was carried out both on the basis of distinct phoneme inventories for every language and on the basis of an artificial phonetic inventory; such an inventory was obtained by collapsing selected subsets of phonemes from the different languages on the basis of their acoustic similarity.

In the following sections, WA scores are presented for every set of experiments; the collapsed phonemes are reported in Table 2 (phonemes are notated in IPA symbols [11]). Furthermore, the two following global parameters contribute to a more complete description of the experiments (Table 4):

- global compression rate (GCR) as defined in (5).

$$GCR = \sum_{i=1}^{N} c_i CR_{L_i} \quad (5)$$

where N is the number of languages, CR_{L_i} is the compression rate of the i-th language defined as the ratio between the merged models and the trainable initial models and c_i is the weight attributed to every language, equal to the ratio between the total number of trainable models in the language L_i and the number of the trainable models in N languages.

This parameter provides a global measure of the obtained compression rate: the total number of trainable models (except for the silence model) is 25 for I, 43 for E, 28 for S and 37 for G.

- Error Rate Global Variation (ΔGER): the ΔGER is the average of the error rate variation (ΔER), defined in (6), of the monovocabulary and the multivocabulary cases:

$$\Delta ER = \frac{ER_{mul} - ER_{mon}}{ER_{mon}} \quad (6)$$

where **mul** stands for multilingual and **mon** stands for monolingual. Using this method it is possible to evaluate the effect of every compression in terms of performance variation.

Further experiments were added to the set described above, in the case of I-S recognition: this new set investigates the possibility of exploiting the phonetic similarity between languages in order to increase the recognition performance for a language, whose models were trained by insufficient data.

The following labelling convention have been used: 'I', 'E', 'S' and 'G' (see above) and 'GL' for Global Results (for WA, GL refers to the ratio between the total number of recognized words for the different languages and the total number of words to recognize); monovocabulary monolanguage experiments are indicated by '1', monovocabulary multilanguage experiments by '2', multivocabulary monolanguage experiments by '3', multivocabulary and multilanguage experiments by '4'. For the additional I-S experiment, the meaning of the numbers is different and is explained below.

3.1. Italian-English

Figure 1 presents the WA curves for the I-E experiments; merged phonemes, as resulting from compression, are shown in Table 2.

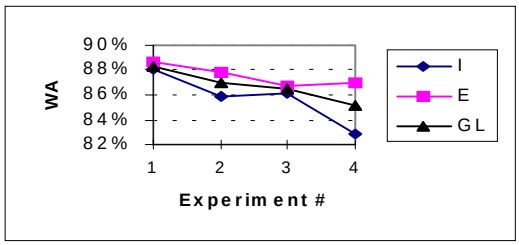

Figure 1: WA curves for experiments I-E

3.2. Italian-Spanish

Figure 2 reports WA curves for the I-S experiments; merged phoneme, as resulting from compression, are shown in Table 2 (in the I-S experiments two more models have been merged: I and S [tʃ]).

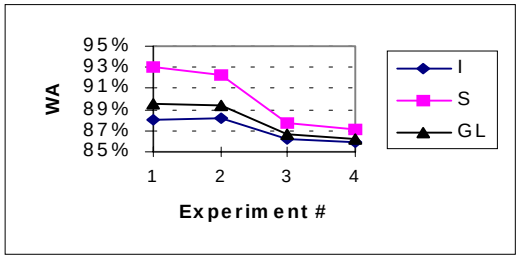

Figure 2: WA curves for experiments I-S

3.3. Italian-English-Spanish

Ex.\La.	I	E	S	GL
1	88%	88.6%	93.1%	89.1%
2	85.6%	86.7%	91.2%	87%
3	84.4%	86.4%	85.1%	85.5%
4	80.1%	85.5%	80.8%	82.8%

Table 1: WA values for experiments I-E-S ('Ex.' = number of the experiment; La.' = Language)

3.4. Italian-English-Spanish-German

I	E	S	G	I	E	S	G
*p**	-	*p**	p	*i**	-	*i**	i:
*r**	-	*r**	r	g	-	-	g
*o**	*ɒ*	*o**	ɔ	*t**	*t*	*t**	t
*n**	*n*	*n**	n	*k**	*k*	*k**	k
*u**	-	*u**	u	*dz**	*θ*	*θ**	s
*l**	-	*l**	l	*m**	*m*	*m**	m
*e**	*ɪ*	*e**	ɛ	*v**	-	*β**	v
f	-	-	f	b	-	-	b
ʃ	*ʃ*	-	-	*ʎ**	-	*ʎ**	-
*a**	*æ*	*a**	a	-	ə	-	ə
z	-	-	z	-	aɪ	-	aɪ
*s**	*s*	*s**	-	-	h	-	h

Table 2: Phonemes merged in the experiments (I-S merges are marked with *, I-E merges are underlined, I-E-S merges are reported in *italic;* finally, merges operated in the I-E-S-G experiments can be inferred from the **G** column, referring to the trilingual models)

Ex.\La.	I	E	S	G	GL
1	88%	88.6%	93.1%	90.7%	89.4%
2	85.1%	87.3%	91%	86.2%	86.9%
3	83.3%	85.5%	83.2%	86.4%	84.7%
4	78.5%	84.7%	80%	85.5%	82.5%

Table 3: WA values for I-E-S-G experiments

3.5. Global parameters

Par.\Exp.	I-S	I-E	I-E-S	I-E-S-G
GCR	64.1%	29.4%	44.8%	53.4%
ΔGER	2%	10.7%	18.9%	19%

Table 4: Global parameters ('Par.' = parameter)

3.6. Reduced training data

Often the speech material available for training is not enough to obtain statistically robust models for a given language. However, the phonetic similarity among languages can be exploited in order to improve performance by the recogniser: in fact, it has been shown [1] that it is possible to train models of a given language using speech material relative to similar models of another language.

The following set of experiments investigates this possibility and evaluates the variation in performance that results from shifting from poorly trained Spanish models (obtained only by Spanish speech material) to multilingual models trained by the same Spanish material and by part of the Italian speech files**.**

Experiments have been run by performing training on different speech material, and have been labeled accordingly: '1' refers to training by 200 files made of Spanish female voices only; '2' refers to training by 200 files of Spanish material equally distributed between male and female voices, '3' and '4' were trained by similar material than '1' and '2', but by a total of 100 files; 300 files (including male and female voices) were always used for training of the Italian component. ER curves for the four experiments and in the two modes (monolingual and multilingual) are shown in Fig. 3.

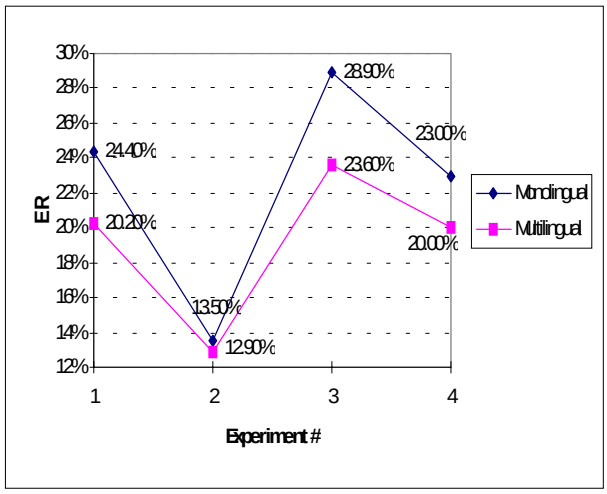

Figure 3: ER curves for exp. with insufficient speech material

Results show a decrease of ER in monolingual vs. multilingual experiments, amounting to **17.2%** in the first condition, to **4.4%** in the second one, to **18.3%** in the third experiment and to **13%** in the fourth one (with a global compression rate between I and S of **64.15%**); these results confirm the efficacy of the multilingual training mode, when the models are trained by insufficient speech material.

4. DISCUSSION

The greater similarity within the I-S acoustic spaces with respect to the I-E ones was evidenced from the first set of experiments. GCR for I-S (64.1%) is significantly greater than GCR for I-E (29.4%); this effect was mainly due to the greater complexity of the English phonetic space [12] with respect to the Italian one. The WA reduction in the multilingual experiment was nearly negligible for the I-S case (2%), while the corresponding value for the I-E case was remarkably higher (10.7%). This result derives from the fact that the common phonetic space between English and Italian was not very cohesive because of the great contextual variations of the phonemes in the two languages, which results in very different coarticulatory effects. These effects provoke a loss in the acoustic resolution of the multilingual models resulting from merging of English and Italian phonemes. Such loss in its turn affected the ER values.

A preliminary experiment was carried out by developing multilingual context-dependent models for I-E-S languages. Recognition results, as reported in table 5, show:

1) an expected improvement in WA scores

2) a reduction of the GCR for the three languages in the context-dependent vs. independent task. This effect was due to the existence of a very small common acoustic-phonetic space between APUCD inventories of E and I.

Ex.\La.	**I**	**E**	**Ex.\La.**	**I**	**E**
1	92.8%	91.2%	**3**	92.4%	89.6%
2	93.3%	90.7%	**4**	95.5%	89.9%

Table 5: WA for experiments with APUCD

Hence it can be argued that the development of an APUCD modelization for the multilingual recognisers would require the extension of the dissimilarity measures described in the paper to APUCD classes.

An important issue that emerged from the previous experiments was the asymptotic trend of recognition performances and of the total number of the multilingual models as a function of the number of languages involved. Also, the gap in performance between the monolingual and multilingual modes asymptotically decreased as new languages were added (as shown in table 4).

5. CONCLUSIONS

An extension of a flexible vocabulary recogniser to a multilingual task has been designed and experimented. We have obtained a reduced set of a multilingual phoneme inventory, covering the I-E-S-G languages, based on a mix of HMM distance measures. The problem of performance degradation due to the loss of acoustic resolution when two models belonging to two different languages are merged was also investigated. Finally, the problem of reduced training data for a given language was addressed, demonstrating that it is possible to exploit cross-language phonetic similarities to statistically strengthen the weak models of the new language. An asymptotic trend by increasing the number of languages, both in terms of absolute recognition performance and in terms of the performance gap between monolingual and multilingual modes, was observed. As a next step, this study will be extended to APUCDs (triphones), where new clustering criteria based on similarity classes, will be required because of the large number of the units involved.

6. ACKNOWLEDGEMENTS

The authors acknowledge the invaluable contribution by Luciano Fissore who provided the training and test environment for the recognition experiments.

REFERENCES

[1] Andersen O., Dalsgaard P. and Barry W. (1993) “Data-driven identification of poly- and mono-phonemes for four European languages”, Eurospeech, Berlin, pp. 759-762.

[2] Andersen O., Dalsgaard P. and Barry W. (1994) “On the use of data-driven clustering technique for identification of poly- and mono-phonemes for four European languages”, ICASSP, Adelaide, pp. I-121 I-124.

[3] Köhler J. (1996) “Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds”, Proceedings ICSLP, Philadelphia, pp. 2195-2198.

[4] Deng L. (1997) “Integrated multi-lingual speech recognition using universal phonological features in a functional speech production model”, ICASSP, Munchen, pp. 1007-1010.

[5] Bub U., Köhler J. and Imperl B. (1997) “In-service adaptation of multilingual Hidden-Markov-Models”, ICASSP, Munchen, pp. 1451-54.

[6] Wheatley B., Kondo K., Anderson W. and Muthusami Y. (1994) “An evaluation of cross-language adaptation for rapid HMM development in a new language”, ICASSP, Adelaide, pp. I-237 I-241.

[7] Mak B. and Barnard E. (1996) “Phone clustering using the Bhattacharrya distance”, ICSLP, Philadelphia, pp. 2005-2008.

[8] Lee K.F., Hayamizu S., Hon H.W., Huang C., Swartz J. and Weide R. (1990) “Allophone clustering for continuous speech recognition”, ICASSP, Albuquerque, pp.749-752.

[9] Mak B. (1996) “A distance measure of speech phones and its application to phonetic context clustering”, Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, Research Proficiency Examination Paper, May, '95.

[10] Leoni F.A. and Maturi P. (1995) “Manuale di fonetica”, La Nuova Italia Scientifica Ed., Firenze.

[11] International Phonetic Association, *The Principles of the International Phonetic Association*, University College, London, 1949 (IPA *Chart, revised to 1993*, in “Journal of the International Phonetic Association”, 23, vol. 1, 1993).

[12] Canepari L. (1980) “Introduzione alla fonetica”, Piccola Biblioteca Einaudi, Torino.