Cross-lingual interpolation of speech recognition models

Giorgio Micca*, Alessandra Frasca[†], Maria Gabriella Di Benedetto[†]

* CSELT, Via G. Reiss Romoli 274, 10148 Torino, Italia giorgio.micca@cselt.it

[†]Universita` di Roma "La Sapienza" Italia

Abstract

A method is proposed for implementing the cross-lingual porting of recognition models for rapid prototyping of speech recognisers in new target languages, specifically when the collection of large speech corpora for training would be economically questionable. The paper describes a way to build up a multilingual model which includes the phonetic structure of all the constituent languages, and which can be exploited to interpolate the recognition units of a different language. The CTSU (Classes of Transitory-Stationary Units) approach is exploited to derive a well balanced set of recognition models, as a reasonable trade-off between precision and trainability. The phonemes of the untrained language are then mapped onto the multilingual inventory of recognition units, and the corresponding CTSUs are then obtained. The procedure was tested with a preliminary set of 10 Rumanian speakers starting from an Italian-English-Spanish CTSU model. The optimal mapping of the vowel phone set of this language onto the multilingual phone set was obtained by inspecting the F1 and F2 formants of the vowel sounds from two male and female Rumanian speakers, and by comparing them with the values of F1 and F2 of the other three languages. Results in terms of recognition word accuracy measured on a preliminary test set of 10 speakers are reported.

1. Introduction

Multilinguality is becoming increasingly important in Automatic Speech Recognition applications, due to two major driving factors: 1) the possible requirement of including words belonging to multiple languages within a single utterance, and 2) the demand for the rapid and economic adaptation of the most relevant Interactive Voice Recognition systems, already available in a few major languages, to a larger community of new languages. An example of the former task might be represented by a worldwide flight information enquiry system, where names belonging to different languages could jointly occur in the same utterance, for instance: "Vorrei gli orari dei voli da Brighton a Clermond-Ferrand domani mattina". The latter requirement should be fulfilled by substantially reducing the amount of speech data to train and validate the recognition models in a new language.

We introduced the concept of sharing multilingual phones exploiting the phonetic similarities of sounds across languages in [5], following previous works on this topic [6]. Then we extended the method to context-dependent acoustic-phonetic units [1], and introduced the concept of Classes of Transitory-Stationary Units. We showed how the statistical richness of a multilingual model based on this type of units can be transferred to a poorly trained recognition model. In the present paper we extend the methodology to the cross-lingual interpolation of the recognition models of a new language under the condition of lacking any training or adaptation acoustic data.

2. The Method

2.1. Building the Multilingual model

The first step in the process of building a multilingual inventory consists in developing language-specific acoustic-phonetic models based on the Transitory-Stationary paradigm. An example of transcription of the words for Italian, Spanish and English languages is given

Word	TS unit transcription		
Cuore	k ku u uo o or r re e		
Acudir	aak kkuuuðððiiirr		
Language	ˈl læ ææŋŋŋgggwwwɪɪɪd d dʒʒ		

Tab.1 Transcription of multilingual words in terms of Transitory-Stationary units.

in Tab. 1. Stationary units represent the central, more stable section of phone realizations, while transitions represent the trajectories across two adjacent phones. This model yields good recognition accuracy scores provided it can be properly trained. This condition depends on the coverage of the training database, which is likely to show some weakness corresponding to the less frequent pairs of adjacent sounds. Therefore, we introduced a metrics for acoustic similarity of sounds, so that similar sounds can be merged into classes. The metrics which we described in [5] is based on a combination of five different acoustic distances; three of them are variants of the Batthacharyya distance, one is based on the information loss (il) computed as the entropy variation induced by the merge of two models:

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

Finally, the fifth metrics is computed as the portion of the N-dimensional acoustic space shared by the probability density functions of the states of the two HMMs. These distances were identified as the most appropriate and were selected from a set of a dozen of different metrics, according to a "phonetic coherence index" described in [5]. The contribution of each metrics is then summed up as follows:

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

where the first, second and fourth components refer to three Bhattacharyya distances, the third component is the information loss and the fifth is the "common acoustic space" (cas) distance. The logarithm here accounts for the higher dynamic range of this component (this distance can be nearly zero). Two variants of the method of merging

CLASS	ENG	ITA	SPA
SL (sil.)	•	•	•
FV1	ii,i	i	i
FV2	e	e	e
CV	a , aa , uh	а	а
BV1	0,00	0	0
BV2	uu , u	u	u
SC	@,@@		
SP	Occlusion	Occlusion	Occlusion
BS	Voice bar	Voice bar	Voice bar
LB	p , b	p , b	p , b
DN	d,t	d , t	t , d , D
PL	k,g	k,g	k , g , G
NA	m, ng, n	m , n	m , n
GN		Ν	Ν
TH	th , s , z	<dz>, s, f</dz>	s , z , T
VF	sh , zh , jh	<dz></dz>	
СН	ch	<t&> , &</t&>	tS
ZE		Z	
YP	у		j,jj
EL	1	1	1
GL		L	L
DH	dh		
VD	W		w
VU	f,v	v	f, B
ER	r	R	r,rr
WH	h		X

Table 2. Classes for the three-lingual model.

transitions into classes were experimented. In variant a), phones of all languages are clustered in classes of similar sounds. These classes are given a mnemonic name, then they are used to create the transition classes where each one of the two constituent sounds is represented by the corresponding class. A table of the classes obtained for our three-lingual model is given in Table 2.

With this variant, the metrics is applied to compute the distance among basic Acoustic Phonetic Units Context Independent (APUCI). In variant b), classes are generated by directly applying the clustering algorithm to properly trained multilingual transition units. In this case, the concept of class of transition component does not apply any longer to the left and right constituents of a transition class, since the merge is data-driven only. Classes are therefore named by a progressive number, and the information on the phonetic identity of left and right components is lost. In our "blind" cross-lingual interpolation task we chose variant a), because we still needed to keep the information on the phonetic class of the components of a transition, in order to be able to map the phones of the target language onto these classes.

2.2. Bootstrapping the Rumanian models

2.2.1. Statement of the problem

It is well known that speech production mechanisms are involved in different ways for different languages. The effects of the differences can be clearly noticed when listening to speech produced by a non-native speaker in a target language, or when testing a speech recognizer with speakers of a mother tongue different from that of the recognizer. The effects can be attributed to the following causes:

- 1. Vowels and consonants may be produced with a wrong place of articulation or manner of production. In particular, the non-native speaker tends to substitute the phonemes of the non-native language with the nearest phonemes belonging to the vowel and consonant systems of his native language.
- 2. Speech may sound unnatural or affected by unusual intonation patterns. This effect is related to differences in prosody of different languages. Intonation patterns may be inappropriate for the target language, as well as phoneme duration patterns, or amplitude variations across segments.

In our case, the problem was in some sense reverse: we wanted to approximate the production of sounds of native speakers by means of the models trained with non-native speakers of several different languages. In other words, we investigated the problem of representing speech segments of a language in a recognition system which was built for other languages. In particular, we focussed the analysis on vowel sounds.

2.2.2. How to find similarities for Rumanian vowels

We based our analysis on the well established theory of vowel articulation indicating that the first two resonances of the vocal tract, F1 (height) and F2 (backness). We examined the similarities of vowel patterns of the Rumanian language with those of the three languages for which the multilingual recogniser had been built. The common latin root of the Rumanian language with the Italian and Spanish components of the multilingual model leaded us to predict a significant overlap of vowel patterns. These similarities can also be quantitatively analyzed by mapping vowel charts [7].

Frequency analysis

The vocoids of the Rumanian and Italian languages are given in Figs 1 and 2 respectively.



Fig. 1 Vocoids of the Rumanian language



Fig. 2. Vocoids of the Italian language







Fig.4. Frequency analysis of the Italian /u/, female speaker

Similarities are apparent for the phonemes /a/, /e/, /i/, /o/ and /u/, while the phonemes / \hat{i} / and /ə/ belong to the internal section of the vocoidal space. We studied these correspondences by means of a frequency analysis of vowels pronounced in isolated mode by two native male and female Rumanian speakers. A sample of the frequency spectrum of the vowel /u/ pronounced by a female speaker is given in Figs. 3 and 4 for Rumanian and Italian languages respectively. The energy peaks for the seven Rumanian vowels corresponding to the same speakers are given in Tables 3 and 4 respectively. The complete set of Rumanian phones and the phones which were used to initialize their models are given in Table 5. The phoneme \hat{i} , which is a central, closed non-round sound was approximated by the closed mid English /I/, and the Rumanian /a/ was approximated by the English schwa /ə/.

Bootstrapping of CTS units

The final step was the generation of CTSUs for the Rumanian language. We mapped the interpolated Rumanian sounds onto the multilingual classes of Table 2, then we obtained the CTSUs for the new language. In principle, CTS units should have provided improved recognition accuracy with respect to context-independent units. On the other hand, we did not know the degree of performance degradation we should expect by interpolating context-dependent models through the "blind" procedure previously described. The increase in precision of the model could not necessarily lead to an improvement in performance of the interpolated model.

vowel	F ₁ (Hz)	F ₂ (Hz)
а	1060	1271
e	528	2666
i	215	2590
0	694	929
u	236	500
ə	620	1250
ĩ	526	884

Table 3. F1 and F2 for Rumanian vowels

vowel	F ₁ (Hz)	F ₂ (Hz)
а	1025	1241
e	428	2791
ε	450	2200
i	249	2740
0	452	898
Э	727	993
u	300	884

Table 4. F1 and F2 for Italian vowels

Rumanian	Example	Approx.	Example
[a]	c[a]p	[a] ITA	p[a]ne
[ə]	m[a]r	[ə] ENG	[a]round
[î]	c[î]nt	[I] ENG	b[e]tter
[e]	m[e]re	[e] ITA	r[e]mo
[i]	[i]nima	[i] ITA	car[i]o
[0]	c[o]rp	[0] ITA	[o]ro
[u]	b[u]n	[u] ITA	s[u]a
[e]	b[e]a	[e] ITA	r[e]mo
[į]	[i]ata	[j] SPA	actuac[i]'on
[o̯]	sc[o]ala	[w] SPA	adec[u]ado
[u̯]	do[u]a	[w] SPA	adec[u]ado
[- ¹]	frat[i]	[i] ITA	car[i]
[b]	[b]un	[b] ITA	[b]ene
[k]	[c]orp	[k] ITA	[c]ui
[Ě]	[cȟ]em	[k] ITA	[c]ui
[t∫]	[c]er	[tʃ] ITA	[c]iao
[g]	[g]ura	[g] ITA	[g]atto
[ģ]	[gh]em	[g] ITA	[g]atto
[dʒ]	[g]eam	[dʒ] ITA	[g]ia`
[h]	[h]aina	[h] ENG	[h]assocks
[3]	[j]oy	[3] ENG	bei[g]e
[1]	[l]up	[1] ITA	a[l]i
[m]	lu[m]e	[m] ITA	[m]ia
[n]	[n]u	[n] ITA	[n]uova
[p]	[p]ot	[p] ITA	[p]uoi
[r]	[r]au	[r] ITA	ma[r]e
[8]	ca[s]a	[s] ITA	[s]esso
[∫]	[s]I	[∫] ITA	[sc]iame
[t]	[t]u	[t] ITA	[t]uo
[tz]	cu[t]it	[ts] ITA	al[z]are
[v]	[v]oi	[v] ITA	[v]oi
[z]	[z]ile	[z] ITA	ro[s]a

Table 5. Initialization of Rumanian models

2.3. Recognition results

The 10 test speakers 5 male and 5 female, were collected on the Italian PSTN. Each speaker uttered 44 isolated words corresponding to the application words defined within the SpeechDat project. We used the CSELT CDHMM recognizer based on variable mixture density functions, with up to 32 Gaussians per mixture. The recognizer was trained with the phonetically balanced component of the SpeechDat databases for each language. We compared three different acoustic-phonetic models: a) APUCI (Acoustic-Phonetic Units, Context Independent), b) TS (Transitory-Stationary) Units and c) CTS Units. Results are presented in Fig. 5. The context-dependent units nearly halved the error recognition rate. The TS units were generated by directly substituting the Rumanian phone-to-phone transitions for the corresponding language-specific transitions.

The test confirmed our twofold hypothesis:

1) similarity metrics are useful in optimizing the process of interpolating the recognition models for an "unknown" language starting from a multilingual inventory of recognition units;

2) even in cases where no data are available to adapt the models to a new target language, the adoption of a context-dependent paradigm for building the recognition units improves recognition accuracy.



Fig. 5. Word recognition performance with Rumanian test speakers, three different models

3. Conclusions

A method for interpolating recognition models for an "unknown" language has been proposed. The procedure exploits the phonetic similarity of sounds across different languages. Classes of context-dependent units can be defined preserving language-specific acoustic identities where necessary and sharing acoustic models where possible. These units can be successfully deployed to derive the models of a new target language for which no training or adaptation data are available. Preliminary tests with mother tongue speakers in Rumanian language have confirmed the viability of the method. Further experiments will be devoted to understand to which extent a multilingual model, based on this paradigm, can be generalized to be useful for interpolating a recognizer of any new language, and to develop a general framework for cross-lingual exploitation of recognition models.

ACKNOWLEDGMENTS

The authors thank Loreta Moisa and Cosmin Popovici for their contribution on the Rumanian language.

4. References

- Micca G., Palme E., Mari J. 1999. A Multilingual Acoustic-Phonetic model based on phone-to-phone transition classes. *Proc. XIV Int. Congress of Phonetic Sciences*, S. Francisco, CA, Vol. 3, pp. 1677-1680.
- [2] Köhler J, 1999. Comparing three methods to create multilingual phone models for vocabulary independent speech recognition tasks. *Proc. MIST ESCA-NATO Workshop*, Leusden, Belgium, pp. 79-84.
- [3] Schultz T., Waibel A. 1998. Language Independent and Language Adaptive Large Vocabulary Speech Recognition. *Proc. ICSLP*, pp. 1819-1822.
- [4] Schultz T., Waibel A. 1998. Adaptation of pronunciation dictionaries for recognition of unseen languages. SPECOM'98. St. Petersburg, pp. 207-210.
- [5] Bonaventura P., Gallocchio F., Micca G., 1997. Multilingual Speech Recognition for Flexible Vocabularies. *Proc. EuroSpeech*, Rhodes, pp. 355-358.
- [6] Köhler J., 1996. Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds. *Proc. ICSLP*, Philadelphia, pp. 2195-2198.
- [7] Di Benedetto M.G., Lienard J.S., 1992. Extrinsic vowel normalization of vowel formant values based on cardinal vowel mapping. *Proc. ICSLP*, Banff, Canada, pp. 579-582.