Phonetic knowledge, phonotactics and perceptual validation for automatic language identification

Martine Adda-Decker$^{\diamondsuit}$, Fabien Antoine$^{\diamondsuit X}$, Philippe Boula de Mareuil$^{\diamondsuit}$, Ioana Vasilescu$^{\spadesuit}$,
Lori Lamel$^{\diamondsuit}$, Jacqueline Vaissiere$^{\clubsuit}$, Edouard Geoffrois^{X}, Jean-Sylvain Liénard$^{\diamondsuit}$
$^{\diamondsuit}$*LIMSI-CNRS, Orsay* $^{\spadesuit}$*ENST, Paris, France*
^{X}*CTA-DGA, Arcueil,* $^{\clubsuit}$*ILPGA, Paris,France*

ABSTRACT

This study explores a multilingual phonotactic approach to automatic language identification using Broadcast News data. The definition of a multilingual phoneset is discussed and an upper limit on the performance of the phonotactic approach is estimated by eliminating any degradation due to recognition errors. This upper bound is compared to automatic language identification based on a phonotactic approach. The eight languages of interest are: Arabic, Mandarin , English, French, German, Italian, Portuguese and Spanish. A perceptual test has been carried out to compare human and machine performance in similar configurations.

Different phoneset classes have been experimented with, ranging from a binary C/V distinction to a shared phone set of 70 phones. Experiments show that phonotactic constraints are in theory able to identify a language (among 8) with close to 100% on very short sequences of 1-2 seconds. Automatic and human performances on very short sequences both remain below the theoretical performances.

1 INTRODUCTION

The work reported here is a contribution towards automatic language identification (LId). LId has been an active research domain for about 30 years, with the pioneering work of Leonard & Doddington (1974), and House & Neuburg (1977). Different sources of information are known to contribute to human language identification: among the most important are acoustics, phonetics, phonotactics, prosody, morphological and lexical knowledge. All of these are of course not equivalently easy to model for automatic LId. Acoustic-phonetic and phonotactic modelling benefit from many decades of research first by linguists to describe languages using compact phoneme systems, and more recently by computer speech scientists elaborating models for automatic recognition. For all these reasons, acoustic-phonetic and phonotactic modelling have become the most popular approach for LId [1, 2]. Other sources of information, such as prosody or morphology, can then be used in addition to a phone-based kernel system, rather than in a stand-alone approach. Another aspect addressed is the comparison of machine and human performances. Assessing performance and understanding changes seems to be important to gain insight into the achieved modelling accuracy and to guide future research.

Acoustic-phonetic and phonotactic modelling raise the question of the phoneset to be used in a multilingual context. For automatic LId using phonotactic constraints, either multiple language-dependent phonesets are used by recognisers in parallel, or a single global phoneset, or even a combination of both [1, 3, 4]. Defining a single global phoneset is an interdisciplinary research issue of its own, which ranges from phonetic and phonological domains [8] to multilingual speech recognition [5].

One aim of this study is to estimate an upper limit of the phonotactic approach, by discarding linguistic noise due to recognition errors. To this end, a priori phone transcriptions obtained via pronunciation dictionaries from orthographic transcripts are used. The link between the elaboration of a global phoneset and the estimation of an upper identification limit of phonotactics for LId is highlighted. The upper limit performances are compared to those obtained from automatically generated phone transcription (by means of phone recognition systems).

2 GENERAL APPROACH AND CORPUS

Broadcast news corpora provide a larger linguistic variety than multilingual telephone conversations which are often used for LId. They are also of higher acoustic quality, which is appropriate for analysing phonotactic modelling.

A multilingual broadcast news corpus has been gathered for the following eight languages: Standard Arabic, Mandarin Chinese, American English, French, German, Italian, European Portuguese and Latin American Spanish. French and Arabic are French DGA resources. English, Spanish and Mandarin are excerpts from LDC Hub4 corpora. German, Portuguese and Italian BN data are resources acquired within various European FP5 LE projects (OLIVE, ALERT) or purchased from ELDA. As resources are most limited for Portuguese (only a few hours), we limited the corpus to about 3 hours per language for most experiments.

Figure 1 gives a simple overview of language identification using a phonotactic approach. Both the estimation of language-dependent phonotactic models (training phase), and the test sequence (HYP) to be identified depend heavily on the acoustic-phonetic decoding accuracy (part A). To as-

ISBN 1-876346-48-5 © 2003 UAB

sess the capacity of the phonotactic decoder without recognition errors, part (A) is removed and canonical phone sequences (REF) are generated via pronunciation dictionaries from orthographic transcripts. These reference phone sequences, used below for the upper limit experiments, need to be expressed in a common alphabet. The first step of defining a shared phone symbol set for the different languages is described in the next section.

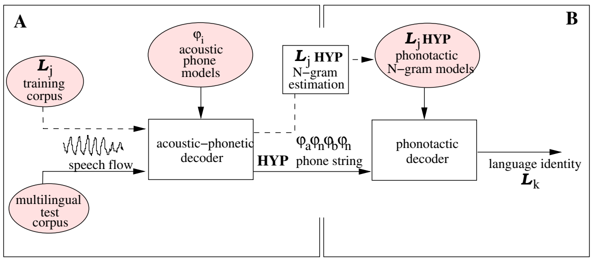

Figure 1: Automatic language identification with a phonotactic approach. The left part (A) outputs a phone sequence which is the input to language identification proper (B). Prior training of language specific phonotactic models is shown in the upper part (dotted lines).

3 COMMON PHONESET

The search for a global typology, a phoneme classification, is one of the fundamental problems of speech sciences [7, 9]: phonemes and their allophones are traditionally defined by the analysis of minimal pairs, linked to the distinctive function within a given language (e.g. in Italian *cui* /kui/ vs. *qui* /kwi/). Each year, standards are published by specialists, to decide to include symbols and diacritics in the International Phonetic Alphabet (IPA). Presiding over the IPA, the principle of typographical economy, although completely sensible in a monolingual context, contributes to inconsistencies in a multilingual framework. This principle recommends that the use of characters other than those of the Roman alphabet be restricted “as far as is practicable” [8]: this means that if a language possesses an /ɛ/ but no /e/, the latter character will all the same be used. It is also suggested to avoid diacritic signs “whenever possible”. These ones should indeed be reserved to a narrow transcription. Nevertheless, for the four emphatic consonants of Arabic, no other symbol than the pharyngealisation diacritic is proposed. Symbols also oppose /b/ to /p/, whereas what is relevant in Mandarin is not the voicing feature but the aspiration feature. In general, the symbols were defined as having the value of the sounds they note in major European languages. A similarity principle has also ruled the IPA and its different reforms since 1888: when we find an identical sound in several languages, it is advised to use the same sign. This also applies to nuances of sounds which are “close” to one another. Nonetheless, what it means for two sounds to be judged “similar” is not specified.

Using the same symbol for close sounds across languages is a central question when attempting to define a common multilingual phoneset. Previous studies carried out at LIMSI proposed to gather “close” phonemes across six languages, on the basis of objective acoustic criteria [6]: they especially displayed similarities between the unvoiced plosives of English, French, German, Italian, Portuguese and Spanish. Grouping what is traditionally noted /ɔ̃/ in French and /õ/ in Portuguese, they confirm what the IPA says, wiz. that this second symbol would be more appropriate for standard French – even though the notation /ɔ̃/ accounts for morpho-phonological changes such as *mention~mentionner*.

3.1 LANGUAGE-DEPENDENT PHONESETS

For each of the eight languages under consideration, phonesets exist for speech recognition purposes. Table 1 specifies the phoneset sizes for the 8 languages. Glides are considered as consonants. Silence and noise symbols are not included. Spanish has the smallest set with only 24 units. In our configuration Mandarin has the largest set with 58 units: the large number of vowel units can be explained by 3 units per vowel, corresponding to globally rising, falling and flat tones each. All 8 languages have large consonant sets (generally more than 20 consonants), with English and Arabic being the most elaborate. The Italian language makes uses of distinct units for geminates which explains the 42 consonants here. German and English have both large C and V sets, Italian, Spanish and Arabic have small V sets. French, Portuguese and Mandarin (when ignoring the tones) have comparable medium size V sets and comparable-size C sets.

Lang:	*En*	*Ge*	*Ma*	*Fr*	*Sp*	*It*	*Po*	*Ar*
Cons.	27	23	22	20	19	21*2	21	31
Vowel	18	23	12*3	14	5	6	14	6
Total	45	46	58	34	24	48	35	37

Table 1: Number of phones in language-dependent phonesets.

For phone modelling, frequency of occurrence is the major criterion for unit selection. Minor criteria are the existence of acoustically “close” units and phonotactic considerations. Xenophones, usually occurring few times, are generally ignored. Similar acoustic units may have distinct distributional properties across languages, which can explain part of cross-language inconsistencies. Problems also arise due to linguistic notation conventions which differ accross languages for acoustically similar events, as exemplified by diphthongs, geminates and affricates. More technically different options have been taken for phone modelling due to frequency of occurrence, or depending on the pronunciation dictionaries used. Taking as an example the German language, /ts/ is not considered as an affricate in the recognition phoneset. Glottal stop is generally not used. A syllabic /n̩/ is used but the less frequent syllabic /m̩/ has not been selected. The xenophone /ʒ/ has been adopted in the German phoneset, whereas the /w/ glide is absent, the short /ʊ/ is considered acoustically close enough. For Arabic, xenophones like /p/ and /v/ have been selected. Italian has special symbols for geminates whereas for the Arabic language the phone symbol is doubled.

ISBN 1-876346-48-5 © 2003 UAB

3.2 LANGUAGE-INDEPENDENT PHONESETS

As phonemes are language-dependent entities, it looks impossible to define a multilingual phoneme inventory. But a sharable inventory of acoustic phone units can be defined more or less accurately. Gathering the existing language-dependent phonesets results in a complex set of over 300 units. Some of these are xenophones or mere allophones in certain languages, and have a phonological status of phoneme in other languages. For the sake of homogeneity and to facilitate automatic processing in a cross-language framework, it is of interest to reduce the number of distinct phones in the global set.

For simplicity reasons we decided to start with broad phone classes, where a consensus across languages is not too difficult to achieve. Among these classes we have used 2 classes (C/V), a set of 10 classes (Vowel, Nasal, Glide, Liquid, Plosive-(v/uv), Affricates-(v/uv), Fricative-(v/uv) with voicing distinction) and a set of 19 "megaphone" classes. The different classes are obtained by appropriate mapping rules applied to the language-dependent phonesets. Compromising linguistic and practical considerations an additional set of 70 multilingual phones has been elaborated. The more complex units, like diphthongs, geminates and affricates are here replaced by a sequence of 2 simple units. A worst case example is provided by the geminate affricates in Italian which are replaced by a sequence of 3 elementary units (e.g. /ddʒ/). Other choices might be more appropriate for temporal modelling, but the retained choice had the advantage of simplicity. The "megaphone" class set has been adapted to the 70 multilingual phones. A question of interest is how the different phones distribute among the different languages. A summary of phone and phone class statistics is provided in Table 3.

Class	*Phones*	*Class*	*Phones*
i	i ì í iː ɪ ĩ y yː	p	p
e	e ə ɛ ɛ̃ œ œː ø	t	t tˤ þ
a	a á à ɑː ɒ æ ã ɑ̃	k	c k g ʔ
o	o ɔ õ	b	b
u	u ú ù uː ũ	d	d dˤ ð ðˤ
j	j	g	g
w	w ʊ	f	f v
l	l əl ʎ	s	s sˤ
r	r ʀ ʁ x ɹ	ʃ	ʃ ç
m	m əm	z	z
n	n ən ŋ ɲ	ʒ	ʒ
h	h ʕ ħ		

Table 2: Cross-language sharable 19 "megaphone" classes.

The /a/ and /e/ vowel classes are significantly more frequent in all languages than the back vowel classes. The /u/ phone class is the least frequent for all languages except for Portuguese. Voiced consonants have about 30% more occurrences than unvoiced ones, but for the fricative and plosive subsets, unvoiced consonants are often twice as frequent as their voiced counterparts. The /n/ class is particularly low for French and Portuguese, suggesting that part of these have "disappeared" to form nasal vowels.

Lang:	*En*	*Ge*	*Ma*	*Fr*	*Sp*	*It*	*Po*	*Ar*
i	10	7	7	6	6	9	6	14
e	12	10	4	17	14	11	14	0
a	7	12	14	11	12	11	15	24
o	3	3	5	6	9	9	6	0
u	2	4	2	4	2	2	**8**	5
p	2	1	1	4	3	3	3	0
t	8	9	7	5	5	9	5	5
k	3	2	2	4	4	4	4	7
b	2	2	1	1	2	1	1	2
d	7	5	5	5	5	5	5	4
g	1	2	2	1	1	1	1	0
m	3	3	1	3	3	2	3	5
n	9	11	11	**3**	8	8	**3**	5
f	4	4	1	3	1	2	2	2
s	5	6	1	6	8	4	4	3
ʃ	1	3	4	1	0	2	**6**	1
h	1	1	4	0	1	0	0	3
z	4	2	1	2	2	1	1	1
ʒ	1	0	2	1	0	1	1	1
l	4	3	2	6	5	6	3	8
r	6	5	2	8	8	7	7	4
w	1	1	9	1	1	1	2	3
j	3	3	11	2	3	2	1	4
%C	63	61	68	56	58	57	51	59
%V	35	36	32	44	42	43	49	41
%Syll	2	3	0	0	0	0	0	0
%Cvoic.	40	36	43	36	36	35	28	35
%Cunv.	23	25	25	21	22	22	23	23
%Plos	19	21	17	20	19	21	20	18
%Pvoic.	6	9	6	7	8	6	7	6
%Punv.	13	13	12	13	11	15	13	12
%Fric	18	16	16	13	11	10	14	13
%Fvoic.	8	4	2	5	0	3	3	2
%Funv.	10	12	14	8	11	7	11	11
%Liq	9	8	2	14	13	13	9	11
%Glid	6	4	21	3	3	2	2	7
%Nas	11	11	13	6	11	10	6	9
%Vnas	-	-	-	7	-	-	6	-

Table 3: "Megaphone" and broad class statistics measured on the BN training data.

4 EXPERIMENTAL SETUP & RESULTS

The multilingual corpus comprises 3 hours of audio data per language. 90% were used for training the phonotactic models and 10% were held out for testing. The test set thus corresponds to about 20 minutes of speech per language, resulting in a multilingual test set of more than 2 hours.

4.1 UPPER LIMIT OF PHONOTACTIC APPROACH

The following experiments aim at measuring the best possible performances for LId using phonotactic knowledge. Reference phone strings are created using different multilingual symbol sets: C/V, the 10-class and the 19-class "megaphone" sets. Synthetic results are shown in Table 4 using 5-gram phonotactic models. All lower order N-grams have been tested and gains have consistently been measured for increasing N. As soon as REF test sequences comprise 20 phonemes and more the upper limit phonotactic approach yields identification results close to 100%.

class set:	*CV*	*10-class*	*19-class*
%LId	35.5	78.6	96.0

Table 4: Upper limit language identification results on REF test sequences comprising only 10 phones.

4.2 AUTOMATIC PHONOTACTIC APPROACH

Test phone sequences are now obtained by automatic acoustic-phonetic decoding. Using a 19-class set, identification rates of 51.9, 83.7 and 93.7 are achieved using phone sequences of length 10, 40 and 80 respectively. These lengths correspond to durations of 0.7, 3 and 6 seconds. With a 70 shared phoneset the results are improved to 63.2, 90.3 and 96.8 % respectively. These results are particulary high, explainable by the use of all data for acoustic model training. Only for phonotactic models, the test data have been held out. However the achieved results highlight the importance of accurate acoustic-phonetic decoding.

5 HUMAN LANGUAGE IDENTIFICATION

A preliminary experimental setup of human perception has been designed to compare human and machine performances in similar test conditions. These similarities are the test stimuli extracted from the BN corpus, length of stimuli and acoustic condition (wide-band high-quality). Whereas the machine models are trained using 3 hours per language, humans disposed of only 20 seconds of speech per language from one speaker. But the test corpus comprises mainly acquainted languages. Short speech excerpts of 1.5 to 2 seconds containing on average 25 phonemes are used. For these short stimuli care has been taken to cut speech on word boundaries. The test corpus consists of 3 speakers per language (2 male/1 female or vice-versa) which results in 24 distinct stimuli. The complete stimuli set was played twice in a random order. 14 French academic natives listened to the 2×24 stimuli producing 14×48 decisions (84 per language if equiprobable). A global correct identification rate of 87.6% is achieved. Details are shown in Table 5.

Human perception results								
stimuli: **p'ceived**	*En*	*Ge*	*Ma*	*Fr*	*Sp*	*It*	*Po*	*Ar*
En	94	-	-	-	-	-	-	-
Ge	1	100	-		-		-	-
Ma	5	-	99	-	-	-	2	1
Fr	-	-	-	100	-	-	-	-
Sp	-	-	1	-	63	10	12	1
It	-	-	-	-	28	78	5	-
Po	-	-	-	-	8	12	78	14
Ar	-	-	-	-	1	-	3	84

Table 5: Human perception confusion matrix on stimuli comprising on average 25 phones (1.5-2 sec). A column shows identification results for a given language.

6 DISCUSSION

We showed that the phonotactic approach could achieve close to 100% identification rates if accurate phone (class) strings were available, even if these strings are expressed using 19 shared multilingual phone classes. The results highlight the importance of accurate acoustic-phonetic decoding for language identification by the phonotactic approach. Comparing the size of the different common phone class sets, significantly better performances are achieved for larger sets. Whereas a direct comparison with machine results is not possible, the perceptual test results show that short segments are difficult to identify for humans and that machines may outperform them in this condition. If considering only the Romance languages subset (minus French, which is the listeners native language) humans achieve only 74%. Future work will further investigate shared multilingual phonesets for acoustic-phonetic modelling. Perceptual tests will be extended, in particular by adding variable acoustic conditions.

Acknowledgements

This work is partly carried out within a CNRS STIC-SHS interdisciplinary programme (MIDL project).

REFERENCES

[1] M.A. Zissman, “Comparison of Four Approaches to Automatic Language Identification of Telephone Speech”, *IEEE Trans. on SAP* 4(1), 1996.

[2] J.-L. Gauvain and L. F. Lamel, “Identification of Non-Linguistic Speech Features”, *Proc. of ARPA Workshop on Human Language Technology*, 1993.

[3] C. Corredor-Ardoy *et al.*, “Language identification with language-independent acoustic models”, *Eurospeech*, vol. 1, (pp.5-8), Rhodes 1997.

[4] D. Matrouf *et al.*, “Comparing Different Model Configurations for Language Identification using a Phonotactic Approach”, *Eurospeech*, Budapest, 1999.

[5] T. Schultz, A. Waibel, "Experiments on Cross-Language Acoustic Modeling", *Eurospeech*, Alborg, 2001.

[6] P. Boula de Mareüil, C. Corredor-Ardoy, M. Adda-Decker, “Multi-lingual automatic phoneme clustering”, *ICPhS*, San Francisco (pp.1209-1213).

[7] P. Delattre, *Comparing the phonetic features of English, German, Spanish and French*, Julius Gross Verlag, Heidelberg.

[8] Handbook of the International Phonetic Association, *A Guide to the Use of the International Phonetic Alphabet*, Cambridge University Press, Cambridge, 1999.

[9] N. Vallée, L.-J. Boë, C. Abry, J.-L. Schwartz, “La matérialité des structures sonores du langage”, *JEP*, Avignon 1996.