6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

THE USE OF RARE SEGMENTS FOR LANGUAGE IDENTIFICATION

Jean-Marie HOMBERT and Ian MADDIESON***
*Dynamique du Langage (UMR5596), CNRS/Université Lyon-2 (France)
**University of California, Los Angeles (USA)

ABSTRACT

Knowledge of the distribution of rare segments across the languages of the world might be used in identifying languages within an open set. Segments which are both discriminatory (i.e. rare) and robust (i.e. easy to identify) are the best targets for efficient language identification. Considering several properties at the same time allows to use more common segments and/or features in a still very discriminatory way.

1. INTRODUCTION

In recent years automatic language identification (ALI) has been a rapidly growing field of research. However, most of the results are based on a limited number of languages, at most 20, and often much less. In this paper we propose to show how the knowledge accumulated in traditional phonetic and phonological studies across the whole range of spoken languages might be applied to the task of improving our abilities to identify languages correctly. It is important to appreciate that what is being envisaged is the task of identifying languages within an open-ended set, rather than within a closed set, as is usual in current approaches to ALI [1] [2] and [3].

Typological data on the world's languages has been accumulating at an accelerating rate in recent decades, so that we now have a good basic knowledge of the phonological patterns of almost all the extant languages. These descriptions are often based on limited familiarity with the language, but nevertheless allow a reasonable approximation of the segmental inventory to be obtained. This knowledge permits the construction of large typological surveys of phonological systems. These surveys permit a well-founded appreciation of which segments are common and which are rare, and how these rare segments are distributed within geographical areas and language families. The original purpose of these surveys was to highlight universal sound patterns [5] but they also necessarily show which segment types are most restricted in their distribution. It is this aspect that can be exploited to extend the possibilities of identifying languages, especially in the context of an open-ended set.

The two databases used in these preliminary studies are those tabulated by Ruhlen [4] and an extended version of the UPSID database originally described in Maddieson [5], which cover respectively 693 and 451 languages. The criteria used to select languages and interpret their phonologies in these databases differ, but their results are highly convergent with respect to the questions considered in this paper. It should be emphasized that these databases are based on phonological analyses only, abstracted from the phonetic richness of the acoustic signal. Consequently a real implementation of the method we are proposing will require a prior stage of automatic or semi-automatic transformation from the phonetic facts to the phonological level. Some approaches to this stage have been addressed [6]. However, even when segments are described by very general phonological categories, these terms do imply some typical acoustic properties that can be expected to be present in the signal. Such properties must be categorized along two dimensions — their discriminatory power and their robustness of identification.

2. DISCRIMINATION AND ROBUSTNESS

In a previous paper [7] we emphasized the importance of separating these two dimensions. Discriminatory power refers to how useful a segment is in narrowing the choice of possible languages; the rarest segments will have the greatest discriminatory power. Robustness of identification refers to how easily a segment's presence can be detected; the more salient and less variable a segment's acoustic pattern is, the easier it is to detect. For example, the presence of /s/-like sibilant fricatives is easy to detect automatically, but since the vast majority of languages have such a sound, this property has little value in discriminating between languages. By contrast, very few languages have voiceless non-sibilant dental fricatives (T), so this is a strongly discriminatory property, but the acoustic signal for this segment is weak and easily confounded with other weak fricatives, such as /f/. Hence, this property also has little value, as it is weakly detectable. The best targets are thus those properties which are both strongly discriminatory and strongly detectable, as indicated in the grid below. The most obvious cases would include clicks, labial-velar stops, and other rare and salient segments. Further suggestions are given in [7]. Other segments will fall into an intermediate category along both dimensions.

weakly detectable	strongly detectable

10.21437/Eurospeech.1999-98

weakly discriminatory	e.g. /f/	e.g. /s/
moderately discriminatory	e.g. /h/	e.g. /S/
strongly discriminatory	e.g. /T/	**e.g. clicks**

By definition, the rarest segments occur in few languages. In order to be able to identify more than these few languages, it is necessary to make use of less powerfully discriminatory features. Less discriminatory features become more powerful when used jointly. For example, consider 3 distinct features which are relatively common — let's say they are each found in 20% of the world's languages. If these features are randomly distributed, then less than 1% would possess all three. In the following sections we will present a detailed illustration of the discriminatory power of features taken in combination in this way.

3. DISTRIBUTION OF FEATURES

The distribution of certain selected features of the consonant systems in 693 languages as given by Ruhlen [4] is shown in Table 1. Similar data for selected properties of vowel systems is given in Table 2. The family grouping are those provided by Ruhlen. Two of the ten consonantal traits illustrated (a voicing contrast in stops and the presence of glottals) are very widespread, one is truly rare (clicks), but seven are found in between 6% and 19% of the languages. Of the vocalic characteristics in Table 2, one — the presence of a vowel length contrast — is found in almost half the languages but the others are found in less than a quarter. Any value between about 25% and 5% might be considered to indicate a moderately discriminatory property. The particular features shown in Tables 1 and 2 are features which we think are likely to be usable in practical recognition tasks, but it is not critical that this be so. We are more concerned with demonstrating the concept of how such distributions can be of value in principle. A further practical issue concerns the length of sample that would be required to detect the presence of given feature in the phonological system. Given that it is desirable that samples used in automatic language identification be as short as possible, it is obvious that only positive detection of features can be taken into consideration. The absence of a given feature in a sample does not necessarily imply its absence from the phonology of the language.

Table 1. Proportion of languages in different families with the consonantal traits identified in the column headings.

	Voice Contrast in Stops	Aspiration Contrast in Stops	Ejectives	Implosives	Prenasalized Stops	Length Contrast in Cons.	Clicks	Labial-Velars	Retroflex Cons.	Glottals
Afro-Asiatic	29/29	0/29	6/29	12/29	4/29	16/29	0/29	0/29	3/29	24/29
Niger-Kordofanian	50/51	4/51	1/51	15/51	13/51	4/51	1/51	36/51	4/51	28/51
Nilo-Saharan	24/25	0/25	2/25	11/25	11/25	9/25	0/25	4/25	8/25	17/25
Khoisan	3/4	2/4	2/4	0/4	0/4	0/4	4/4	0/4	1/4	4/4
Indo-European	71/73	13/73	1/73	1/73	1/73	8/73	0/73	1/73	19/73	41/73
Caucasian	37/37	3/37	37/37	1/37	0/37	17/37	0/37	0/37	0/37	37/37
Uralic	15/23	0/23	0/23	0/23	0/23	12/23	0/23	0/23	2/23	12/23
Altaic	36/39	3/39	1/39	0/39	0/39	5/39	0/39	0/39	0/39	21/39
Paleo-siberian	3/8	1/8	1/8	0/8	0/8	5/8	0/8	0/8	0/8	7/8
Dravidian	10/10	1/10	0/10	0/10	0/10	3/10	0/10	0/10	9/10	5/10
Sino-Tibetan	12/18	15/18	0/18	0/18	1/18	0/18	0/18	0/18	3/18	16/18
Austro-Asiatic	16/17	10/17	0/17	4/17	1/17	0/17	0/17	0/17	8/17	16/17
Indo-Pacific	37/50	2/50	2/50	1/50	18/50	0/50	0/50	1/50	3/50	26/50
Australian	2/24	1/24	0/24	0/24	1/24	1/24	0/24	0/24	19/24	2/24
Austro-Tai	50/67	8/67	1/67	4/67	12/67	6/67	0/67	0/67	10/67	57/67
Eskimo-Aleut	1/5	0/5	1/5	0/5	0/5	2/5	0/5	0/5	0/5	4/5
Na-Dene	7/12	7/12	12/12	0/12	2/12	0/12	0/12	0/12	1/12	12/12
Macro-Algonguian	3/13	1/13	0/13	0/13	0/13	3/13	0/13	0/13	2/13	13/13
Salish	3/10	0/10	10/10	0/10	0/10	0/10	0/10	0/10	0/10	10/10
Wakashan	0/2	1/2	2/2	0/2	0/2	0/2	0/2	0/2	0/2	2/2
Macro-Siouan	4/12	1/12	3/12	0/12	0/12	0/12	0/12	0/12	0/12	12/12
Penutian	26/43	2/43	31/43	14/43	1/43	3/43	0/43	0/43	9/43	41/43
Hokan	5/19	3/19	7/19	0/19	0/19	0/19	0/19	0/19	4/19	19/19
Aztec-Tanoan	7/15	2/15	2/15	0/15	0/15	2/15	0/15	0/15	4/15	15/15
Oto-Manguean	11/14	1/14	1/14	1/14	4/14	0/14	0/14	0/14	3/14	14/14
Macro-Chibchan	7/10	3/10	2/10	1/10	0/10	0/10	0/10	0/10	2/10	10/10
Ge-Pano-Carib	13/24	1/24	1/24	1/24	0/24	0/24	0/24	0/24	8/24	23/24
Andean-Equatorial	19/39	8/39	3/39	2/39	1/39	1/39	0/39	0/39	9/39	35/39
TOTAL/693	501	93	129	68	70	97	5	42	131	523
Percent	72.3	13.4	18.6	9.8	10.1	14.0	0.7	6.1	18.9	75.5

Table 2. Proportion of languages in different families with the vocalic traits identified in the column headings.

	Nasalized V's	Front Rounded V's	Back Unrounded V's	Long V's
Afro-Asiatic	0/29	0/29	0/29	17/29
Niger-Kordofanian	27/51	3/51	1/51	24/51
Nilo-Saharan	1/25	0/25	1/25	12/25
Khoisan	4/4	0/4	0/4	2/4
Indo-European	19/73	15/73	0/73	33/73
Caucasian	14/37	12/37	1/37	15/37
Uralic	0/23	13/23	5/23	13/23
Altaic	1/39	25/39	21/39	23/39
Paleo-siberian	0/8	0/8	0/8	4/8
Dravidian	2/10	0/10	0/10	10/10
Sino-Tibetan	3/18	7/18	5/18	5/18
Austro-Asiatic	6/17	0/17	2/17	7/17
Indo-Pacific	2/50	1/50	0/50	14/50
Australian	0/24	0/24	0/24	9/24
Austro-Tai	2/67	3/67	5/67	38/67
Eskimo-Aleut	0/5	0/5	0/5	3/5
Na-Dene	8/12	0/12	0/12	9/12
Macro-Algonguian	2/13	0/13	1/13	11/13
Salish	0/10	0/10	0/10	3/10
Wakashan	0/2	0/2	0/2	1/2
Macro-Siouan	8/12	0/12	0/12	8/12
Penutian	0/43	0/43	3/43	26/43
Hokan	1/19	0/19	0/19	15/19
Aztec-Tanoan	3/15	1/15	3/15	11/15
Oto-Manguean	13/14	0/14	3/14	3/14
Macro-Chibchan	6/10	0/10	0/10	0/10
Ge-Pano-Carib	9/24	0/24	10/24	6/24
Andean-Equatorial	23/39	0/39	5/39	13/39
TOTAL/693	154	80	66	335
Percent	22.2	11.5	9.5	48.3

In addition to the overall frequency of these features, the tables show how some of them have quite marked variations in their frequency distribution in the different language families listed. These variations also correlate strongly with a geographical pattern of distribution (for the geographical location of language families see, for example, www.sil.org/ethnologue). The labial-velar stops /k͡p, g͡b/ are found almost exclusively in two African families of languages, Niger-Congo and Nilo-Saharan. Implosives are present especially in all three large language families of Africa — Niger-Congo, Afro-Asiatic and Nilo-Saharan — and in the Penutian family of North America. Front rounded vowels occur almost exclusively in five language families mainly present on the Eurasian landmass, Indo-European, Caucasian, Uralic, Altaic, and Sino-Tibetan. Such patterns of distribution mean that, given a knowledge of its segmental features it is often possible to focus in on a likely area in which a language is spoken, or to say which family it belongs to even if it is not possible to identify the specific language.

4. JOINT DISCRIMINATION

Three of the features in Tables 1 and 2 were utilized in a test of their joint discriminatory power. These three features are: the occurrence of nasalized vowels, of labial-velar stops, and of retroflex consonants. In the expanded UPSID database of 451 languages, 22.6% (102) have nasalized vowels in their inventories. There are marked regional disparities; none (0%) of the 72 languages in the major families of Oceania — Australian and Papuan — have nasalized vowels, but 20 of 49 (or 40.8%) of the North American languages (not including Eskimo-Eyak) have this feature.

Labial-velar stops are found in a substantially smaller number of languages than nasalized vowels, 41 or 9.1% of the languages in the database. If we search for those languages with both these traits, there are only only 19 (4.2%). All of these are African. Just these two features eliminate languages from other areas, such as North America or Europe.

When a third feature, the occurence of retroflex consonants is added, only three languages are retained from the original set of 451, which is less than 0.7% of the original search area. To select between the remaining three languages, a large number of traits, some of which are in themselves not at all rare, can be

used. For example, Lelemi has a simple velar nasal /N/ which is not found in Dan or Sango.

5. CONCLUSION

Existing databases of phonological systems can be used to provide a geographical distribution of segments found in the world languages. Segments which are both rare and easy to identify are extremely valuable in an automatic language identification task. But it is also important to point out that even less restricted (found in 5 to 25% of the sample can be very discriminatory when used jointly (e.g. nasalized consonants, labial velar stops and retroflex consonants are found in less than 0.7% of the samples).

6. REFERENCES

[1] Muthusamy, Y. K., E. Barnard & R. A. Cole. (1994) Automatic language recognition: A review/tutorial. *IEEE Signal Processing Magazine* 10/94: 33-41.

[2] Pellegrino, F. & R. André-Obrecht. (1996) Stratégies en identification automatique des langues. Vers une classification automatique des systèmes vocaliques". *XXIes Journées d Étude sur la Parole*, Avignon: 409-412.

[3] Pellegrino, F (1999) Ed. Actes de la Première Journée d'Etude sur l'Identification Automatique des Langues : de la caractérisation à l'identification des Langues".

[4] Ruhlen, M. (1975) *Guide to the World s Languages.* Stanford University.

[5] Maddieson, I. (1984) *Patterns of Sounds.* Cambridge: CUP.

[6] Ohala, J. J. & Marsico, E. (1999) Differentiating phonetic from phonological events in speech. In [].

[7] Hombert, J.-M. & Maddieson, I. (1998) Automatic language identification : a linguistic point of view. *UCLA Working Papers in Phonetics* 97: 119-124.