MULTI-LINGUAL AUTOMATIC PHONEME CLUSTERING

Philippe Boula de Mareuil‡*, Cristóbal Corredor-Ardoy ‡ † and Martine Adda-Decker‡
‡ Spoken Language Processing Group, LIMSI-CNRS, Orsay, FRANCE,
* now with Élan Informatique, Toulouse, † now with Bouygues Télécom, Paris

ABSTRACT

In this article, we describe an approach for automatic multi-lingual phoneme classification. The classes were obtained by agglomerative hierarchical clustering. We used a similarity measure based on the likelihood between the acoustic frames and the Hidden Markov Models. The method was applied to French, English, German, Spanish (IDEAL corpus), as well as to Italian and Portuguese (SPEECHDAT corpus). The analysis of the clusters demonstrated that, despite the acoustic mismatch between these corpora, this approach remains robust. For 90 clusters, the obtained classes correspond, to a large extent, to well defined linguistic groups. A qualitative analysis of the results is given.

1. INTRODUCTION

The search for a typology of phonemes is one of the stakes of speech studies [10, 18], for teaching [6] or automatic language processing [2, 3]. In this article, we describe a clustering algorithm for grouping the phonemes of six European languages (French, English, German, Spanish, Italian and Portuguese) and we give a linguistic interpretation of the results.

The automatic phoneme clustering, within a multi-lingual framework, was initiated by the development of corpora for automatic language identification (LID): OGI-TS [14], CALLHOME, CALLFRIEND, SPEECHDAT, IDEAL [4, 5]... It is possible to replace several phonetic recognition systems by a single, common system for all the languages [4].

In our approach, the acoustic analysis is based on cepstral coefficients, rather than on formant extraction. Indeed, the latter requires a voiced/unvoiced segmentation, the location of vowel nuclei and the extraction of formant values, decisions that are subject to error.

We propose a hierarchical clustering algorithm. Berkling [2] also applied this method to the LID of 6 languages of the OGI corpus, but she uses a similarity measure based on the distances between acoustic vectors. Köhler [12] uses a measure based on acoustic likelihoods; nevertheless, this measure was only applied to 3 out of the 11 languages of the OGI corpus. The normalization introduced in our similarity measure, likewise based on acoustic likelihoods, enabled us to apply the automatic clustering within a more extensive framework: that of the IDEAL [4, 5] and SPEECHDAT corpora. The method used happened to be robust with respect to the different recording conditions and to the linguistic contents of the corpora.

Within a mono-lingual framework, the hierarchical clustering algorithm may assist in analyzing phonetic classes, in defining phone sets for speech recognition, and in grouping a large number of Markov models in context.

Section 2 presents the automatic clustering algorithm. Section 3 is devoted to the experimental framework: corpora and alphabets. Section 4 analyzes the results obtained, having the number of desired classes decrease from the 235 phonemes of the initial set, and details the clustering in 90 classes, before concluding.

2. THE AUTOMATIC CLUSTERING ALGORITHM

2.1. Preview

Since the emergence of the first speech recognition systems based on a Markov phonetic modeling, being able to measure the distance between the models has happened to be important. This measure is used to reduce the number of models in context, to determine a smaller phone set within a language or to establish a phone set common to several languages.

We opted for a hierarchical clustering, even though the latter does not allow backtracking (even badly computed, a class cannot be considered back).

The phonemes of six languages were grouped by a hierarchical automatic clustering [7]. In the initialization phase of the algorithm, each phoneme is assigned to a class. After each iteration, two classes of maximum similarity measure are grouped. The procedure is repeated until the desired number of classes is obtained. Note that several phonemes of a same language may be grouped, since no constraint on the linguistic origin has been taken into account.

2.2. The similarity measure

The definition of a similarity (or dissimilarity) measure between phonemes or allophones is an issue that has been widely addressed by the scientific community. Young [19] defined the dissimilarity between allophones by expressing the divergence of two gaussians as a function of their mean values and their variances. However, this approach only applies to one-state Markov models and to a single gaussian. By replacing the divergence concept by Bhattacharyya’s distance, Mak [13] proposed an alternative expression of dissimilarity as a function of the models’ parameters. This approach applies to Markov models with gaussian mixtures, but with a single state.

The comparison of two Markov models with several states and gaussian mixtures requires a similarity (or dissimilarity) mea-
5.1. Reduction of the number of classes

At each state, the law of emission of observations is defined as the weighted sum of 32 gaussians. For each language, all the sentences were phonetically aligned. This segmentation was obtained automatically by using the orthographic transcription of the sentences, the word list with their related graphemic and phonemic forms, and the acoustic phone models. We recall that the segmentation $\varphi_i$ is used for the similarity measure computation.

We computed the acoustic likelihood matrix, from the cepstral vectors corresponding to the 235 phonemes and from their related Markov models. Before executing this operation, we had to limit the number of acoustic vectors per phoneme to less than 20,000 samples (7,000 cepstral vectors on average), to reduce the computation time. Once the likelihood matrix was computed, we applied the hierarchical clustering algorithm, by having the number of desired classes vary.

5. ANALYSIS OF THE RESULTS

5.1. Reduction of the number of classes

One can group the phonemes of a language or of several languages according to different criteria. It is classical, in phonetics, to adopt a representation in the form of a more or less hierarchical tree, according to the degree of analysis. This is possible with our algorithm, where we choose the number of classes in advance, by having this number decrease. With 2 classes, we roughly find a
vowel-consonant bipartition; with some ten classes, we roughly have the following groups, which are mutually exclusive: open vowels, closed vowels, fricatives, plosives, nasals, liquids, glides. Even though some phonemes are refractory to grouping, and even though other groupings are not the ones we wish, it is interesting to observe the linguistic relevance of the classes produced by a computation on purely acoustic features (the cepstral coefficients).

However, our goal is placed at a lesser level of abstraction: we wish to lose as few phonemic information as possible. We do not pretend to formally define this concept: it is difficult to evaluate it as a way of phonemes and their allophones are traditionally defined for a given language, by the analysis of minimal pairs of commutation, linked to the distinctive function. Recommendations and standard conventions are published every year by specialists, to decide to officially include symbols and dialectics within the International Phonetic Alphabet (IPA). Nevertheless, it is not by chance, intuitively, if neighbor sounds, in several languages, are represented with the same IPA symbol [16]. The algorithm described above allows us to objectively explore this proximity.

Starting from the 235 phonemes of our initial set (for French, English, German, Spanish, Italian and Portuguese), we very quickly have Italian geminates which are grouped with the corresponding simple consonants; but a class /t/ common to the six languages only emerges around 90 classes. With 80 classes, the phonemes /zl/ happen to be grouped, but the /zl/ is attached to /lu/. Since we have to find a compromise (may be somewhat arbitrary), what we are going to describe in more detail is the clustering in 90 classes. Moreover, 90 is roughly speaking the number of different IPA symbols that are used to describe our six languages.

5.2. Interpretation of the 90 classes of phonemes

We have 35 singletons (corresponding almost to as many vowels as consonants): some examples are the French /f/, the English /u/ and /a/, the German /k/, which are characteristic of their languages. Among the remaining 55 classes, reported in Table 2, we can notice that, with a few exceptions like the velar /l/ of Portuguese and the /sl/ (jota) of Spanish, there is no overlap between vowels and consonants, nor between the following groups: fricatives, affricates, plosives, nasals, liquids, glides. Some classes only gather 6 phonemes of identical IPA symbol across the 6 languages (plus a 7th for the geminates of Italian): /ul/, /ul/, /ul/, /ul/, /ul/, /ul/. In these figures, we have to reckon with the noise initially introduced by heterogeneous phone sets (especially as far as duration features, diphthongs, glides, affricates, geminates are concerned), more or less rich in (vocalic) allophones. Affricates and diphthongs often happen to be isolated - even though the German and English /ul/ are grouped. Our model does not allow us to determine whether they would match with two segments (for instance /ul/ and /ul/ for /ul/). The automatically derived classes generally correspond to identical or similar units. Nonetheless, we have to mention that 12 of them are pairs of phonemes of a same language: for example, the English /zl/ and /zl/. This case may be explained by the fact that the essential of the energy of the /zl/ is concentrated below 4 kHz. Therefore, the filtering achieved by the telephone band deprives us of a large part of the information. Inversely, the /ul/, /ul/, /ul/, /ul/, /ul/, as well as the voiced fricatives (especially /f/) are not all gathered.

6. CONCLUSION

We applied an automatic clustering approach to the phonemes of six languages (French, English, German, Spanish, Italian and Portuguese). This method is based on a hierarchical clustering algorithm, which enables the follow-up of the evolution of an agglomerative grouping. Starting with an initialization with 235 phonemes, we had the algorithm iterate while analyzing the linguistic relevance of the clusters obtained at each step. We kept the number of 90 clusters, in order not to lose much phonemic information. In majority, these classes correspond to a linguistic reality.

We also analyzed the clustering results in the lowest levels of the hierarchical agglomeration - where we find the vowel-consonant dichotomy. This is another evidence for the efficiency of this method.

We should also notice the ability of this approach to process data of different origins. Analyzing the results obtained at each iteration, we can observe that we did not get a clustering by corpus, nor a good clustering for the phonemes of certain languages and a bad clustering for the phonemes of other languages. This seems to demonstrate the validity of the similarity measure, facing the acoustic variations of the recording conditions, and the linguistic content of the different learning corpora. Indeed, the probability density function corresponding to the acoustic vectors of the phoneme φ_l, within the expression of the similarity measure, achieves a normalization on all the acoustic data.

This approach was applied within the framework of the automatic identification of 4 languages (French, English, German and Spanish). The results obtained for the computation of phoneme clusters of six languages enable the extension of the decoder, from four to six languages. We can envisage to use the acoustic models corresponding to the phoneme clusters of the known languages, for the phonetic recognition of a new language, for which only speech corpora without phonetic transcriptions are available.

Quantifying the proximity of phonemes and allophones of a same language may contribute to the definition of an economical set of phonemes. Grouping the phonemes of several languages eventually applies to multi-lingual speech synthesis, and to didactic, descriptive or corrective phonetics.

NOTES

1. Our approach therefore enters the maximal clustering type (inter-language + intra-language) proposed by Berkling [3].
2. The similarity measure between two classes may also be computed as the maximum or the minimum of similarity between phonemes [7].
3. Indeed, some phonemes such as geminates have few representatives. For these phonemes, the validity of the statistical models may be questioned.

REFERENCES


Table 2: clusters of at least two phonemes for 90 classes, across the 6 languages studied (French, English, German, Italian, Portuguese) - one cluster per line.