CROSS-LANGUAGE MERGED SPEECH UNITS AND THEIR DESCRIPTIVE PHONETIC CORRELATES

Paul Dalsgaard[1]*, Ove Andersen*[1] *and William Barry*[2]

[1]Center for PersonKommunikation, Aalborg University, Denmark
[2]Institut für Phonetik, Universität des Saarlandes, Germany
{pd, oa}@cpk.auc.dk, wbarry@coli.uni-sb.ge

ABSTRACT

The focus of this paper is to formulate an approach to merging phonemes across languages and to evaluate the resulting cross-language merged speech units on the basis of the traditional acoustic-phonetic descriptions of the phonemes. The methodology is based on the belief that some phonemes across a set of languages may be similar enough to be equated, contrasting traditional phonology which treats phonemes from one language independent from phonemes from another language. The identification of cross-language speech units is performed by an iterative data-driven procedure, which merges acoustically similar phonemes from within one language as well as across languages. The paper interprets a number of merged speech units on the basis of articulatory descriptions.

1. INTRODUCTION

In Crystal [1] the languages of the world are shown in a cluster tree containing families of languages which share aspects of their development. Families may be set up due to their linguistic similarities at any descriptive level. It is the speech sound system similarities that are interesting in the context of this research.

Figure 1 illustrates the two major sub-families of the Indo-European language family without giving any details of possible further sub divisions. The main point here is that the languages applied in this research come from two separate sub-families.

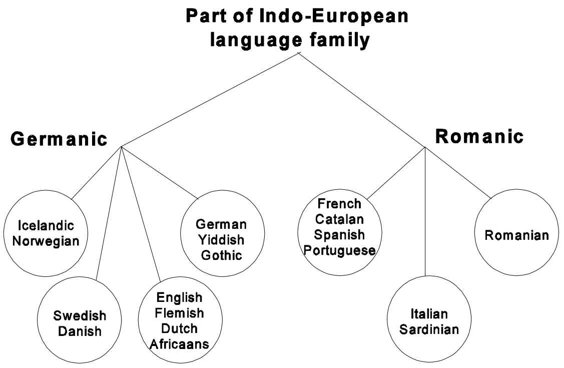
Part of Indo-European language family
Germanic
Romanic
Icelandic Norwegian
German Yiddish Gothic
French Catalan Spanish Portuguese
Romanian
Swedish Danish
English Flemish Dutch Africaans
Italian Sardinian

Figure 1 Symbolic representation of language families

This paper focuses on multi-linguality and takes as its outset the observation that, across languages, the realisation of some phonemes are similar enough to be equated from an acoustic-phonetic point of view. Such "merged speech units" - may be applied in a number of language-specific applications although they are established on the basis of speech corpora covering several languages, see e.g. [2,3,4]. This approach to multi-linguality contrasts most other approaches in speech technology, where a technique or system is referred to as being multi-lingual if it has the ability to process more than one language. There, each language is normally treated independently, not taking advantage of the commonalities that exist between the languages in question.

This paper asks the general question whether it is possible to limit the total number of speech sounds across a number of languages way, i.e. grouping acoustically equateable sounds while still being able to relate the speech units to each monolingual phonological framework.

All speech sounds or phonemes $\mathbf{\Phi}_L$ describing Q languages of the world can theoretically be defined by: $\Phi_L = \Phi_1 \cup \Phi_2 \cup \ldots \cup \Phi_Q$ where the k'th language is described by its own set of language-specific phonemes $\Phi_k = \{\varphi_{k,1}\ \varphi_{k,2} \ldots \varphi_{k,N(k)}\}^T$ and the total number of phonemes is N(k). The total number of phonemes across all Q languages is

$$N = \Sigma^{Q}_{j=1} N(j)$$

Each phoneme is defined by the International Phonetic Alphabet or its equivalent Worldbet (WB) in terms of a description and a symbol. Comparison of the canonical articulatory definitions and the corresponding acoustic realisations for a number of phonemes across the world languages reveals, however, that a number of phonemes have the same or similar articulatory definitions and are acoustically similar. In this paper these facts are submitted to discussion. To simplify the description used in the following we have adopted the following notation:

$$\mathbf{\Phi}_L = \{\mathbf{\Phi}_1, \mathbf{\Phi}_2, \ldots, \mathbf{\Phi}_Q\}^T = \{\varphi_1\ \varphi_2\ \varphi_3 \ldots \varphi_N\}^T$$

2. PHONEME SIMILARITIES

This section presents the data-driven methodology by which some of the phonemes from the total set $\mathbf{\Phi}_L$ of N language-specific phonemes φ_k belonging to Q languages can be merged into a set $\mathbf{\Psi}_L$ of generalised speech sound units of size $K \leq N$.

2.1. General concepts and definition

Applying the data-driven merging methodology to the set Φ_L results in the merging of some of the language-specific phonemes into a common speech unit now representing for example the phonemes φ_p and φ_q. The merging is performed on the basis of acoustic similarities of the speech segments representing the two phonemes represented in the common training speech corpus encompassing all Q chosen languages. One merging thus reduces the total number of phonemes/-speech units by one. Each speech unit is created either as the result of merging two phonemes or a phoneme with an already merged speech unit ψ_j.

Speech units $\mathbf{\Psi}_{cl}$, which is the result of merging phonemes across two or more languages, are termed *cross-language (cl) merged speech units*. The remaining speech units $\mathbf{\Psi}_{li}$, termed *language-internal (li) speech units,* consist of merged speech units and phonemes and of non-merged language-specific phonemes from a specific language. After the merging is finished (see later) there exist a group of language-internal speech units for each of the Q languages. Each element of one of language-internal group $\mathbf{\Psi}_{li,k}$ may as such either correspond to one *language-specific* phoneme, φ_p, or to a speech unit ψ_j resulting from the merging of two or more phonemes from one language only.

The total set $\mathbf{\Psi}_L$ of speech units resulting from the data-driven process can thus be described by:

$$\mathbf{\Psi}_L = \mathbf{\Psi}_{cl} \cup \mathbf{\Psi}_{li,1} \cup \mathbf{\Psi}_{li,2} \cup \, . \, . \cup \mathbf{\Psi}_{li,k} \quad . \, . \cup \mathbf{\Psi}_{li,Q} = \{\psi_1 \; \psi_2 \, . \, . \, . \; \psi_K\}^T$$

where the index $k \in \{1 \, . \, . \, Q\}$ references each individual language and the total number K of speech units is the outcome of the merging process.

2.2. Data-driven definition of cross-language merged speech units

The basis of the data-driven methodology is a number of iteratively conducted phoneme decoding experiments. These apply speech segment models λ_k which are all trained on appropriately annotated speech material encompassing all the languages being included in the experiments. The merging procedure is initialised with N_i=N (*i* is the iteration index) language-specific phoneme models λ_k, each modelled by a Hidden Markov model. The data-driven merging procedure finally defines the set $\mathbf{\Psi}_L$. The size K of the set $\mathbf{\Psi}_L$ can be varied by choosing the number of iterations *i* to be performed.

During each iteration, the strategy is to select and merge those speech sound segments which correspond to the two most similar speech units and/or language-internal speech units. The methodology is briefly outlined below. The merging is based on the results of experimentally established acoustic similarities. The similarity between any pair of HMM models λ_p and λ_q is calculated on the basis of a recognition experiment in which the Viterbi-based log-likelihood score $c(\lambda_p,\lambda_q)$ is calculated across all pairs of speech unit models, λ_p and λ_q, given within the combined training corpus. The phoneme decoder calculates the average per frame values of the log-likelihood scores. Details of the iterative data-driven methodology are given in [5].

The results of applying the iterative data-driven methodology to the combined spoken language corpus is the identification of the set $\mathbf{\Psi}_L$ of speech segment units. Their identification is based solely on the assumption that acoustically similar speech segment can be equated across and within languages. Each element of $\mathbf{\Psi}_{cl}$ corresponds to speech segments contained within the combined corpus of Q spoken languages. In contrast, each element of the set of language-internal speech units $\mathbf{\Psi}_{li,k}$, encompasses speech segments from language *k* of the combined spoken language corpus only. It may thus represent either one of the language-specific phonemes or the merging of two or more language-specific phonemes from that language.

The data-driven, iterative process makes it possible to check the validity of the approach in a flexible way by setting the stopping parameter Γ equal to the resulting number of merged speech units**.**

3. TRAINING AND TEST DATABASE

Having established the underlying general framework for merging cross-language speech units, we now consider a focussed evaluation to demonstrate its validity. Three languages from the Oregon Graduate Institute Multi-Lingual Telephone Speech Corpus (OGI_TS) [6] are used in this research, namely American-English (UK), German (GE) and Spanish (ES), which are all labelled at the phonetic level.

The parts of the OGI_TS corpora that are used are the spontaneous speech utterances. In this research only the files containing the 'before-tone' acoustic material are used. The symbols used for annotating the 'before-tone' files are from the WB set of symbols, which were designed by Hieronymus [7] to provide computer compatible symbols which are consistent with the International Phonetic Alphabet. In the OGI_TS corpus there are not enough examples of all the different speech sounds to establish separate and robust models for all variations of the phonetic realisations. Therefore, the diacritic markers of the phonetic symbols have been removed. This of course increases the variance of the phonetic categories, possibly merging elements which might be viable speech units if sufficient examples were present in the corpus. The basic phonetic/phonemic sound symbols, and the symbols used to annotate 'non-speech' and the glottalisation segments are given in [7]. The 'non-speech' segments and the glottalisation-segments are all treated as one group in the present research. Thus, taken together the three speech corpora contains N=113 phonemic speech sounds; 40 for American-English, 41 for German and 32 for Spanish after the removal of the diacritics.

4. SPEECH MODELLING AND TRAINING

Each of the speech units is modelled by a context independent, continuous density HMM model described by a three state left-to-right structure with one skip. The probability density function in each state is modelled by two Gaussian mixtures. All phonemes/speech unit models are trained on acoustic features based on the use of the RASTA [8] technique.

Each of the ten 'non-speech' segments - including the glottalisation segments - is modelled by an ergodic CDHMM model described by four states each with two Gaussian mixtures to be modelled. HTK2.0 is used for training and testing.

5. SPEECH UNITS AND RESULTS

Figure 2 shows the results of all merging involved when the set Ψ_{cl} of cross-language speech units has been identified by the data-driven methodology. It shows that the set consists of Γ= 21 cross-language speech units after the conduction of i = 40 iterations. The figure shows a cluster tree which may be looked at as an intermediate ‘snapshot’ of the results of the data-driven merging during 40 iterations. Additional merging - allowing for more iterations to take place - will further extend the cluster tree by drawing new candidates from the pool. In the following, examples of the results given in the cluster tree are interpreted on the basis of the acoustic-phonetic characteristics of the individual phoneme as given in WB.

Detailed analysis of the results presented in Figure 2 shows that 9 cross-language speech units are merged solely on the basis of phonemes belonging to the class of vowels (including the diphthongs), and 12 cross-language speech units belonging solely to the class of consonants. It is also observed that speech units include phonemes across the two classes.

The 9 vowel/diphthong based speech units are ψ_21, ψ_35, ψ_15, ψ_17, ψ_37, ψ_24, ψ_29, ψ_36 and ψ_40. The 12 consonant based speech units are: ψ_2, ψ_28, ψ_33, ψ_13, ψ_8, ψ_31, ψ_39, ψ_11,ψ_16, ψ_19, ψ_30 and ψ_32.

The following comments are made to selected speech units from this list of vowel-based speech units:

ψ_15: both phonemes are high and front,
ψ_21: the phonemes are characterised by comparable articulatory characteristics, namely mid-high to high and back rounded. Furthermore they are all long vowels,
ψ_24: this merging is most problematic from traditional phonetic and phonological feature description; the phonemes differ in length (long vs. short), in their diphthongal vs. monophthongal nature, and in their lip-rounding, though standard German />Y/ tends to be only weakly lip-rounded,
ψ_29: both phonemes have a front-central tongue position with mid-high to high tongue height. The difference in lip-rounding is, in fact often most neutralised by rounding of /l/ in labial and in palato-alveolar fricative contexts.
Seen from an overall acoustic-phonetic point of view, the clustering seems to take place on the basis of expected variations within the realisations of the different sounds (e.g. due to reductions) such that a vowel based merging to a large extent can be explained by articulatory features like tongue height and position, and lip-shape, and all seem to be of importance.

Following observations are made to a selection of the consonant based speech units:

ψ_2: these two phonemes have identical manner and place specifications (and compare the early merging of identically specified nasals in ψ_1 and ψ_3),
ψ_13: the four phonemes have the voiceless aspirated stop release in common,
ψ_28: the merging of the nasals and the laterals captures a natural class of front sonorants, merging the alveolars (ψ_23) before including the labial,
ψ_31: the three obstruents share the voiceless post-alveolar fricative property,
ψ_32: the two phonemes are back fricatives, and though nominally differing in voicing characteristics, they both tend to be voiced in sonorant contexts and voiceless after voiceless obstruents.

It can be observed that, in general, comparable place and manner consonants tend to merge after very few iterations (see ψ_1-3 and ψ_6-9).

Merging that occurs after a greater number of iterations groups sounds into larger (mainly) natural classes according to manner or (less frequently) place (see ψ_28; ψ_33; ψ_13; ψ_31; ψ_11; ψ_16; ψ_19 and ψ_30). The most disparate grouping is ψ_39, but even within that case, the previous merger steps can be seen to be closely related groupings (ψ_38, apical obstruents; ψ_27 apical obstruents with frication).

It is also worth mentioning the phenomena that an expected merging did not show-up in Figure 2. One would for instance expect that the American-English phoneme /m/_US would merge with their two counterparts /m/_GE and /m/_ES which were the first two phonemes to cluster into ψ_1 and later being merged with a substantial number of phonemes, but not /m/_US. Further iterations (the results not given here) shows that /m/_US is merged with the cross-language speech unit ψ_19 consisting of /V/_ES and /G/_ES at iteration number i=55.

The results of the merging is illustrated symbolically in Figure 3 in a language-family cluster tree encompassing the three languages used in the experiments and showing the distribution of the speech units across these languages.

6. CONCLUSIONS

This paper addresses the question of multi-lingual techniques in speech processing from an alternative point of view as compared to most work found in the literature on this topic. The significant difference arises in the way the languages in question are dealt with. Most often each language is treated separately. The present paper advocates for an approach that combines the speech sounds within and across languages.
A data-driven approach taking spectral and temporal characteristics into account is proposed. An evaluation of the approach is carried out on a telephone corpus comprising three European languages. It is demonstrated that most of suggested mergers can be explained from an articulatory-phonetic interpretation.

The analysis of similarities across the selected languages is, of course, interesting from a phonetic point of view. But the idea has also some

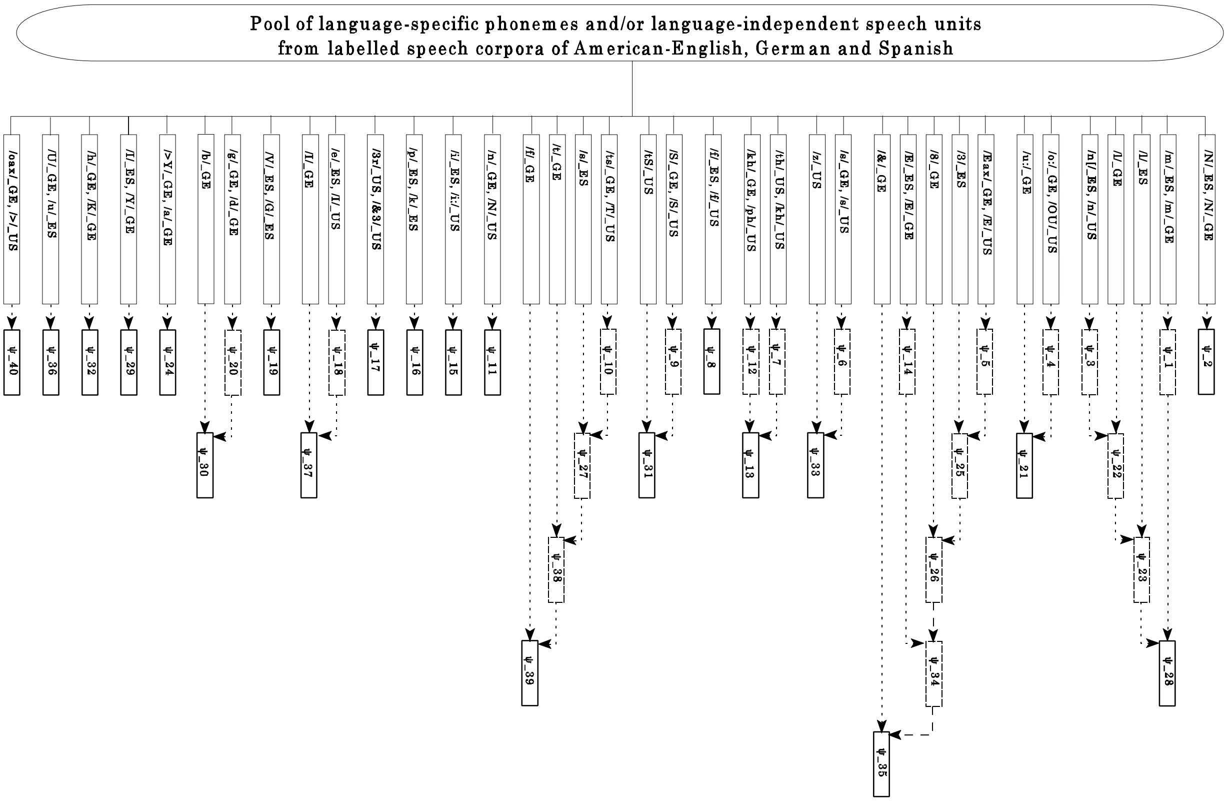

Figure 2 Phoneme/speech unit cluster tree after i = 40 iterations

practical implications as well. In e.g. a language identification system work has been undertaken to analyse whether performance can be improved by putting more emphasis on the language specific units rather than assuming that all phonemes contribute to the task with the same amount of information.

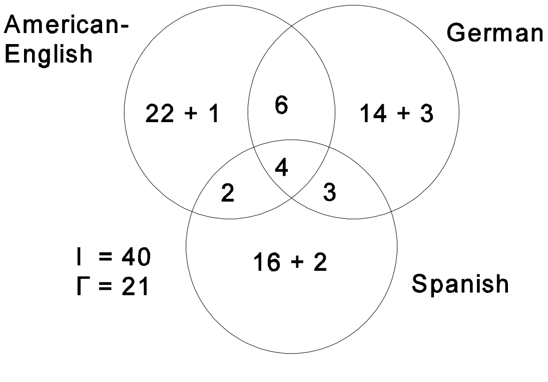

Figure 3 Symbolic representation of speech units across the three test languages

7. REFERENCES

1. D Crystal (1987). ‘The Cambridge Encyclopedia of Langauges’. Cambridge University Press, Chapter IX, The Languages of the World.
2. William Barry and Paul Dalsgaard (1993), "Speech Database Annotation. The Importance of a Multi-Lingual Approach", Keynote speech, In Proceedings of EUROSPEECH93.
3. Paul Dalsgaard and Ove Andersen (1994), "Application of Inter-Language Similarities for Language Identification", In Proceedings of ICSLP94, pp. 1903-1906.
4. Joakim Köhler (1996), "Multi-Lingual Phoneme Recognition Exploiting Acoustic-Phonetic Similarities of Sounds", In Proceedings of ICSLP96, Philadelphia
5. Paul Dalsgaard & Ove Andersen (1996), “On the Identification and use of Language-Independent Speech Units in Language Identification”, The 3’rd CRIM-FORWISS Workshop, October 7-8, Montreal, Canada.
6. Y K Muthusamy, R A Cole, B T Oshika (1992). 'The OGI multi-language telephone speech corpus', International Conference on Spoken Language Processing, Banff, Canada, pp 895 - 898.
7. J L Hieronymus (1993). 'ASCII Phonetic Symbols for the World's languages: Worldbet'. Journal of the International Phonetic Association.
8. H Hermansky et al. (1992).‘RASTA-PLP speech analysis technique’. In ICASSP’92 Proceedings, Pg. 121-124.