# A MULTILINGUAL PHONETIC REPRESENTATION AND ANALYSIS SYSTEM FOR DIFFERENT SPEECH DATABASES

*Toomas Altosaar[1], Matti Karjalainen[1], Martti Vainio[2]*

[1]Acoustics Laboratory, Helsinki University of Technology, Finland
[2]Department of Phonetics, University of Helsinki, Finland

## ABSTRACT

A multilingual phonetic representation and analysis system for different speech databases is presented. The need for such a system is first justified and then one is proposed based on the Worldbet phonetic alphabet. A phonetic class hierarchy is developed and a description of the hierarchical structural representation follows. Database access is based on the latter and is accomplished by defining predicate search functions and applying them to a database. Immediate signal analysis of the results is possible since the multilingual phonetic representation system is seamlessly integrated into a digital signal processing environment.

## 1. INTRODUCTION

Many speech databases already exist for different languages of the world [1] and more are currently being produced. Speech databases provide a wealth of knowledge regarding spoken language. It has been realized that they represent the foundation onto which robust speech analysis, synthesis, and recognition systems can be built. However, due to the existence of different standards, computer readable phonetic alphabets, and specifications used in these speech databases, their use in a common analysis environment has not always been feasible.

We present a new multilingual phonetic representational system that uses Worldbet's [2] phonetic symbols mapped onto a phonetic class hierarchy. Instances of these phonetic classes (e.g., phonemes) are then linked together with diphones, syllables, words, and sentences, to form a structural lattice representing the intrinsic composition of the utterance. Advanced searches can then be performed on the lattice permitting complex analyses to take place.

The system is integrated into our object-oriented QuickSig signal processing platform [3] which is built on top of the Lisp/CLOS programming language. It provides a dynamic and research-motivating environment for the speech scientist. A wide variety of analyses and tasks can be performed in QuickSig such as spectral analyses, database operations, neural network studies, etc.

Each different database, such as TIMIT, EUROM, Kiel, etc., requires a specific parser that converts the phonetic transcriptions to Worldbet phonetic class instances. These instances can display themselves using different phonetic symbols, such as IPA, Worldbet, SAMPA, etc., if a corresponding mapping exists.

Since different signal storage formats are also used in the databases, converters are required for the actual audio data as well. If desired, an existing database can be saved as a new QuickSig database with no loss of information since all types of database-specific data are readily supported.

To perform a speech analysis the user first defines a context search predicate. A search is then initiated over all or part of the database's phonetic lattices, and all matching environments are collected. Any phonetic or signal processing analysis that exists in QuickSig can then be applied to the results of the search. These analysis methods include, e.g., segmental durations, average spectral characteristics, and prosodic studies. Other analyses can be added since QuickSig is an open system and is readily extendible. Systems for database access that use a different formalism also exist [4].

The knowledge gained from having flexible access to several different speech databases in a uniform environment is of valuable use in areas such as speech synthesis, recognition, and phonetic studies, e.g., in contrastive phonetics. This paper describes the system in detail and presents some analyses that have been performed on different speech databases.

## 2. MULTILINGUAL PHONETIC REPRESENTATION SYSTEM

### 2.1. Worldbet

Due to the lack of any universal standard in speech databases (regarding the symbols used in labeling as well as the audio data) multilingual speech database access systems have been difficult to develop. Worldbet [2] addresses the labeling deficiency by providing a common formalism for representing the sounds of any of the world's languages.

Worldbet, an ASCII version of the International Phonetic Alphabet (IPA) has in addition broad phonetic symbols not presently in the IPA [2]. Worldbet can be seen as a symbolic representation for all different speech sounds in the world, with each sound having a separate symbol. In addition, allophonic variation can be described with the use of diacritics. Therefore existing phonetic alphabets used for labeling speech corpora such as TIMITBET, SAMPA, etc., can be mapped onto Worldbet.

Through the common description of a sound provided by Worldbet any phone labeled in one phonetic alphabet can be mapped to the corresponding symbol in another alphabet. E.g., a phone transcribed in SAMPA can be viewed in the TIMIT representation given that the appropriate mapping exists. This may be of benefit for a researcher studying another speech database that has been transcribed in an unfamiliar alphabet.

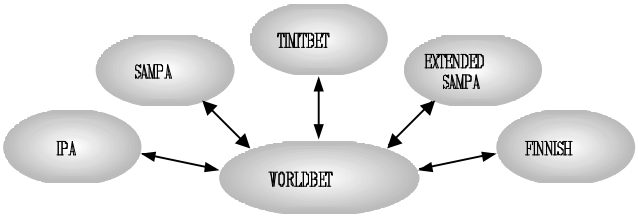


**Figure 1.** Existing phonetic alphabets can be based on Worldbet.

## 2.2. Phonetic Class Hierarchy

A specific Worldbet symbol can be seen as a class which, when further broken down, reveals its constituent phonetic features. A class is thus defined by combining certain features, e.g., /ɛ/ (/E/ in Worldbet) combines the features open-mid, front, unrounded, monophthong-vowel as well as voiced.

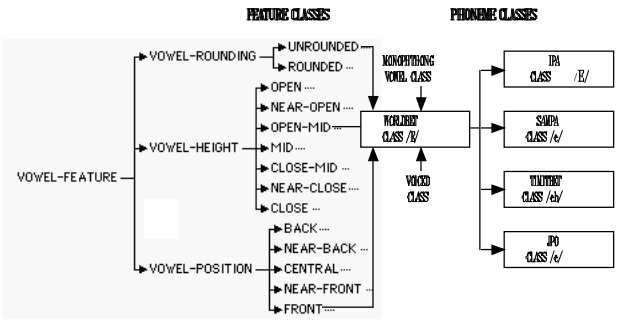


**Figure 2**. Part of the phonetic class hierarchy used to define the Worldbet class /E/ (/ɛ/ in IPA) as well as other subclasses.

A specific class, e.g., /eh/ in TIMITBET, inherits the features found in the base Worldbet class /E/. Worldbet inherited classes, although not necessary, are used in our system to define phonetic alphabet specific subsets in a compact manner. E.g., alphabets found in the EUROM corpuses are derived from the SAMPA classes.

## 2.3. Parsers and Class Instances

Once the alphabet specific class subsets have been defined, e.g., SAMPA for Dutch, instances of these classes can be created to represent actual data found in the corpus. The mapping from a phonetic symbol, e.g., [eh] in TIMITBET, to its corresponding class is performed by a specific parser implemented for each phonetic alphabet.

Diacritics attached to a phonetic symbol are used to depict allophonic variation. An instance of a phonetic class inherits its class' default characteristics as well as any overriding or additional features. E.g., the phone [ɛ̃] can be transcribed as being unvoiced and nasalized using diacritics and the subsequent class instance retains this knowledge.

## 2.4. Hierarchical Structure Representation

In addition to the above-mentioned phonetic classes, other structural classes such as *sentence*, *word*, *syllable*, and *segment* are used to define a continuum of discrete levels over the different temporal scales of speech. These units can all have internal structure, i.e., they may include phonetic subunits.

Also available are a set of *di-units* that are used to relate a pair of primary units. E.g., a *diphone* relates two consecutive phonemes while a *disyllable* relates two syllables. Other units such as spacers and pauses are used to signal pauses in the signal.

With the above mentioned classes a hierarchically structured phonetic representation can be constructed. For Finnish the hierarchical order of the units has been: *sentence* → *words* → *syllables* → *phonemes* → *segments*. Di-units used have been limited to diphones and disyllables. In general, the different hierarchical levels that are built are determined by the available information supplied in the transcriptions of a specific utterance in a corpus.

Figure 3 shows part of the phonetic structure that results for the TIMIT utterance "She washed your suit in greasy wash water all year". From the sentence, word, and phoneme levels the hierarchical network is formed by the parser. Some links are computational while others are "hardwired" to increase database access efficiency.

## 2.5 Speech Database Access

It is important to be able to define search predicates precisely and efficiently which can then be applied to speech database material. Since this system supports a multilingual phonetic model through an implementation of Worldbet, search predicate forms can be formulated in several different phonetic alphabets: in the native alphabet (i.e., using the same symbols as found in the transcriptions), in Worldbet, or in another alphabet – as long as a 1-to-1 mapping exists between the foreign and native alphabet (i.e., both classes are derived from the same Worldbet class). Also, mixed alphabet syntax is also supported, e.g., a search form can be defined in both TIMITBET and SAMPA. The actual search takes place using Worldbet classes.

A library of primitive functions, relations, and types is used to construct the predicates, e.g., *typep*, *prev-phoneme*, and *vowel*, or broader classes such as *tremulant*. The system has several simple predicates already defined which users can use. For more complex searches new predicates can be designed using the same formalism.

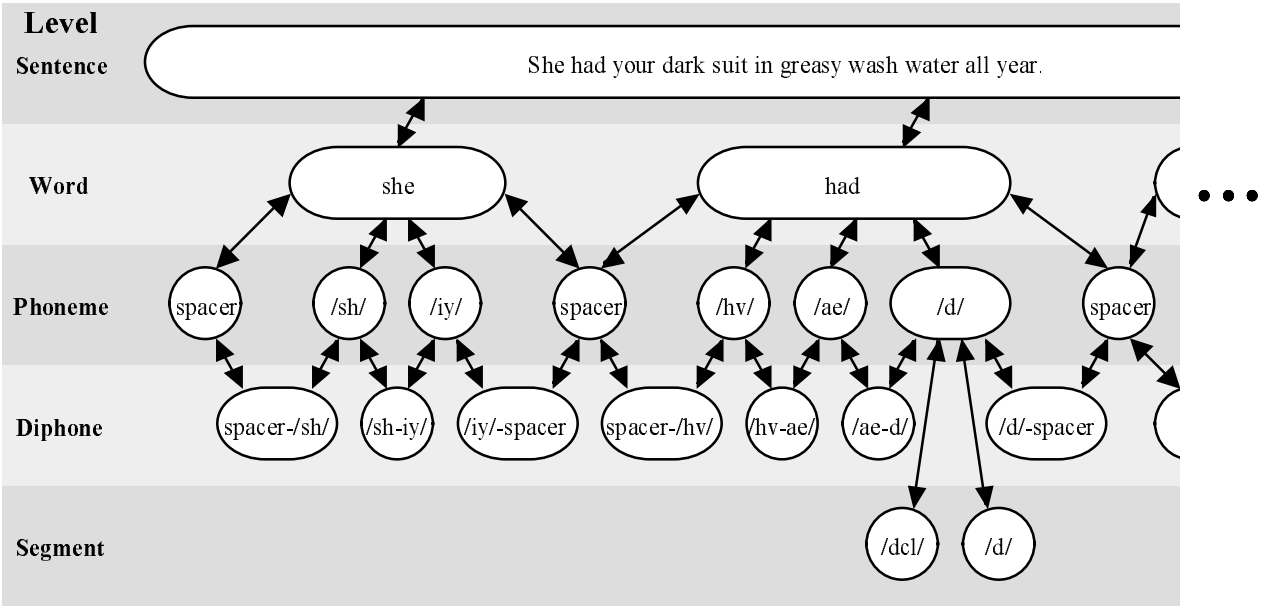


**Figure 3**. Part of the hierarchical structure representation for the TIMIT sentence: "She had your dark suit in greasy wash water all year". Ovals and circles represent instances of phonetic classes while double ended arrows indicate bi-directional links. Some computational links, e.g., *next-phoneme*, are not shown but are rather computed during an application of a predicate.

Searches can be applied to an entire database, or, selectively to only certain parts. A common graphical interface exists for all databases and allows the user to control the search space effectively [5]. The search engine operates with the phonetic structures and their classes (that have been created by the parser), and not the transcriptions that exist in string format in a corpus. This makes performing searches both flexible and efficient.

An example of a predicate form that can be used to search for V/m/V occurrences for male speakers is shown below. The variable x represents a phoneme object onto which the predicate is applied.

```
'(lambda (x)
   (and (typep x (Worldbet "m"))
        (typep (prev-phoneme x) 'vowel)
        (typep (next-phoneme x) 'vowel)
        (eq (gender (speaker x)) 'male)))
```

In this example a test is first made to see whether the phoneme being tested (the variable x) is a subclass of the Worldbet class /m/. If so the test continues to check whether the previous and next phonemes adjacent to /m/ are both vowels. Finally, a test is made to ensure that the gender of the speaker is male. If all tests passed then the predicate function returns a non-nil value indicating success.

A search returns a list of objects that matched the predicate form. These objects can be any instances of any phonetic classes, e.g., phones, words, diphones, etc., and can be operated upon with the signal processing environment described in the next section.

## 2.6. Signal Processing in QuickSig

To be able to perform analyses on the actual speech signals a digital signal processing (DSP) environment must be readily available, preferably in the same environment. In our DSP environment called QuickSig [3], many required signal processing operations are available. QuickSig is an object-oriented signal processing system that supports object classes for signals, filters, spectrograms, etc. all of which can be used by the speech researcher.

QuickSig is implemented in Lisp/CLOS and provides a dynamic and motivating and environment in which speech scientists can perform analyses. QuickSig is readily extendible and new signal processing methods may be added incrementally to the rest of the system on-line and in a matter of seconds.

Our multilingual phonetic representation system is seamlessly integrated with the rest of QuickSig enabling signal analyses to be applied immediately to the results of a search.

**Signal Analysis**

Once a transcription has been parsed into its corresponding phonetic structure the speech signal is accessible through links from any object at any level of representation. This allows for signal processing operations available in QuickSig to be applied to the actual speech signal, or a feature calculated

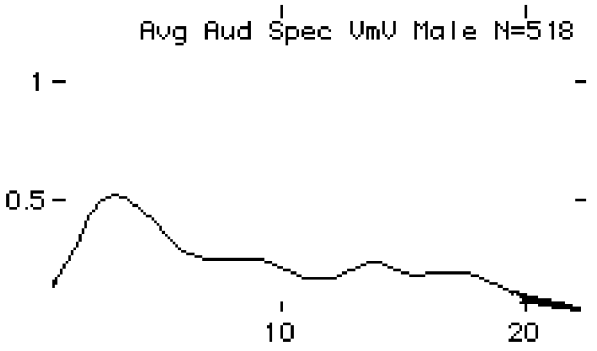


**Figure 4**. Average auditory spectra and distribution for /m/ in a V/m/V context. Male TIMIT speakers only. Loudness (relative sone) vs. pitch (Bark).

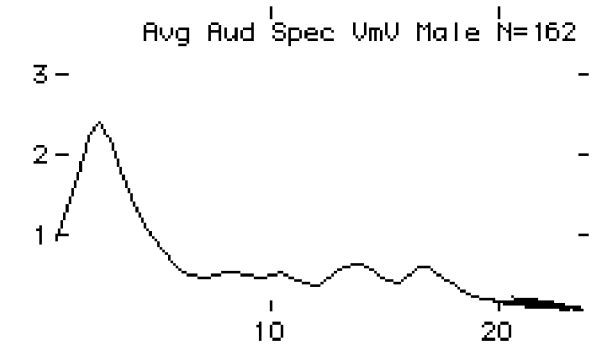


**Figure 5**. Average auditory spectra and distribution for /m/ in a V/m/V context. Male Finnish speakers only. Loudness (relative sone) vs. pitch (Bark).

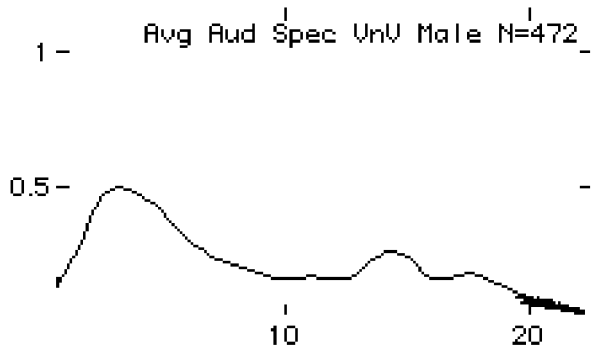


**Figure 6**. Average auditory spectra and distribution for /n/ in a V/n/V context. Male TIMIT speakers only. Loudness (relative sone) vs. pitch (Bark).

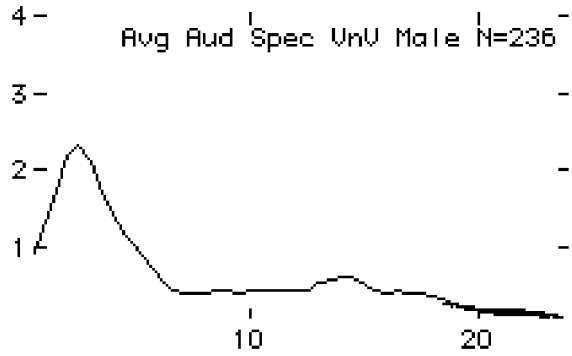


**Figure 7**. Average auditory spectra and distribution for /n/ in a V/n/V context. Male Finnish speakers only. Loudness (relative sone) vs. pitch (Bark).

from it, e.g., F0, loudness, auditory spectrum, etc. In the following section examples are presented that illustrate the flexibility of the phonetic system as well as the signal processing environment.

## 3. EXAMPLES

In this section spectral analyses are applied to the results of searches performed on different speech databases. Using the predicate that was defined in section 2.5 for male V/m/V searches, and applying it to the TIMIT database yields 518 context matches. Figure 4 shows the result of calculating an auditory spectrum for each of the 518 phoneme and displaying the average as well as the distribution of the spectra. By applying the same predicate to our Finnish speech database and performing the same analysis, figure 5 is obtained. If the predicate is modified slightly to search for only /n/ phonemes, i.e., `{typep x {Worldbet "n"}}`, then by performing the same analyses on the different databases figures 6 and 7 are obtained.

## 4. DISCUSSION

Currently we are expanding the number of parsers in the system so that other significant speech databases, e.g., the EUROM series, Kiel, etc. can be analyzed. By using a universal phonetic alphabet such as Worldbet identical analyses can be readily performed over different databases.

## REFERENCES

1. Documentation for different speech databases: TIMIT, EUROM, Kiel.

2. Hieronymus, James. L., "ASCII Phonetic Symbols for the World's Languages: Worldbet", Bell Labs Technical Memorandum 1993.

3. Karjalainen, M., "DSP Software Integration by Object-Oriented Programming: A Case Study of QuickSig." IEEE ASSP Magazine, April, 1990.

4. Hendriks, Jan. P.M., "A Formalism for Speech Database Access," Speech Communication 9 (1990) 381-388. Elsevier Science Publishers B.V. (North-Holland).

5. Karjalainen, M. and Altosaar, T. "An Object-Oriented Database for Speech Processing," Eurospeech-93, Berlin, 1993.