FORMING GENERIC MODELS OF SPEECH FOR UNIFORM DATABASE ACCESS

Toomas Altosaar[1] and Martti Vainio[2]

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

ABSTRACT

This paper presents a formalism that models speech from different databases generically. For each utterance in a speech database a communication framework is first constructed which is composed of a set of communication planes, such as acoustic, orthographic, linguistic, and phonetic. Each plane in turn is made up of a set of levels to represent the plane's structural hierarchy, e.g., for the linguistic plane, levels such as sentence, word, syllable, and phoneme may exist. Information from speech databases is parsed and compiled into such objects and exhibit both individual and class inherited behaviour. Once placed into the framework these objects can have their relationships to other objects explicitly defined by links on the same level, across different levels, and across different planes. Speech from databases covering different languages and annotation styles can therefore be modelled generically allowing for uniform database access. Searches can be performed on the framework and the results used for further analyses.

1. INTRODUCTION

Speech databases may include time-aligned transcriptions for some or all of their audio files. These representations of the speech process, symbolic in nature and consisting of textual characters, provide a way to locate segments in the speech waveform. Database queries are typically accomplished by reading in label strings and searching for certain characters [1]. By relying on a purely symbolic textual representation of the speech process, complex search functions may become difficult to formulate and prone to errors especially above the word level [2]. Furthermore, subsequent analyses of database query results that are not linked explicitly into the speech process may be non-flexible to use thereby limiting the speech scientist's power of expression.

This paper presents a system for modelling the relationships between speech units explicitly rather than leaving the representation on a purely symbolic and textual level alone. Based on object-oriented programming, speech units are modeled as objects which contain inherent knowledge about their "genetic" makeup and behavior. For example, phones and phonemes know of their constituent distinctive features and therefore database searches are free to exploit their class inheritances and dependencies. By defining all possible human sounds using a set of phonetic features, search predicates can be expressed both compactly and efficiently in any existing phonetic alphabet. Additionally, sentences, words, syllables, etc., are also modelled as objects which are aware of their immediate contextual environment through links. For example, a phoneme is connected to its neighbouring phonemes as well as to the syllable or word it belongs to. By parsing and compiling symbolic and textual representations of speech into hierarchical object structures, a more flexible and useful model is available on which to conduct analyses.

The formed hierarchical network models are generic in the sense that they are independent of language, character set, and phonetic alphabet. Furthermore, different transcription styles such as linear, non-linear, with or without overlap, and ones that support componential features, can be represented in the same framework. Database access becomes uniform regardless of the corpus and thereby facilitates comparative studies of spoken language.

The richness of the symbolic description of speech, i.e., what levels of annotation detail are supplied for an utterance within a corpus, determines which hierarchical structures are created. For example, individual structures for phonetic, linguistic, orthographic, and prosodic representations, are constructed if these information sources exist in the corpus. Objects from these different hierarchical structure planes are linked to other objects when possible, e.g., a link from a phoneme to a phone from the linguistic and phonetic planes, respectively, can be used to indicate its actual spoken realisation. Inter-object links form a detailed multi-dimensional network model of speech where relationships that define the essence of spoken language are represented explicitly.

Database searches over these networks are performed by pattern matching using objects. Functions are used to define search templates which traverse the structures looking for desired contexts. Objects within the structures can be tested against their local and class inherited properties. Queries return the actual objects within the structures allowing for their immediate analyses since all object links are retained and available for use. For example, a search returning a set of back-vowels in some specific context can immediately have their pitch contours calculated since each phone is linked to the speech waveform and is aware of its temporal extent. These contours can then be used, e.g., for training neural networks to generate microprosody [3].

The methods used to form generic models of speech are described in the following sections. First, the concept of a communication framework is developed where information from speech databases can be stored. Since databases exist in a diverse variety of formats section 3 describes a formalism that models the storage of speech material existing in files. Once read into memory, the symbolic and textual information from different databases can be viewed in a labelling frame. Section 4 describes the methods used to parse and interpret symbolic information from speech databases that are represented using different annotation standards, e.g., phonetic alphabets. Creation of the hierarchical network structures is then described. Examples of a search query formulation is finally given. Figure 1 shows the above process graphically.

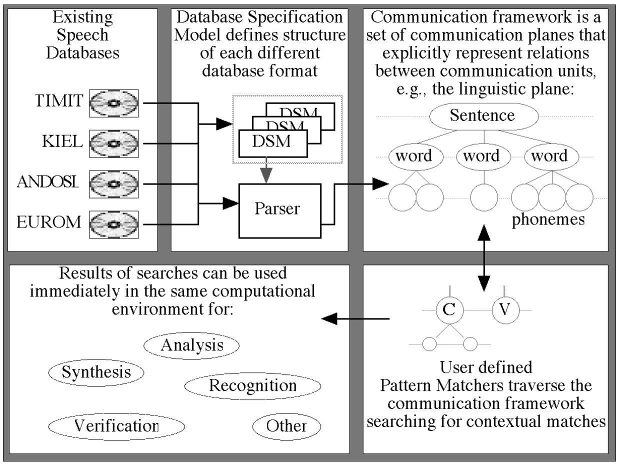
Existing Speech Databases
TIMIT
KIEL
ANDOSL
EUROM
Database Specification Model defines structure of each different database format
DSM
DSM
DSM
Parser
Communication framework is a set of communication planes that explicitly represent relations between communication units, e.g., the linguistic plane:
Sentence
word
word
word
phonemes
Results of searches can be used immediately in the same computational environment for:
Analysis
Synthesis
Recognition
Verification
Other
C
V
User defined Pattern Matchers traverse the communication framework searching for contextual matches

Figure 1: Overview of transforming speech material from databases into generic representations of spoken language. Results of searches performed on communication frameworks can be used in different applications.

2. DEFINITIONS

Communication in humans is carried out by messages coded as detectable and creatable changes in the senses. Mainly the high bandwidth senses, hearing and sight, are used for communicating via spoken and written natural language. For a message to be understood information in a sense channel is transformed from a low level representation, e.g., an acoustic waveform, to a higher one, e.g., semantic. *Communication planes* defined by theories of natural languages exist, e.g., phonetic, linguistic, prosodic, orthographic, etc., and are used to define different representation spaces. Each communication plane may have one of more distinct *communication levels* that define a plane's structural hierarchy across different scales. For example, the linguistic plane may contain the following levels among others: sentences, words, syllables, and phonemes. Finally, each communication level contains a set of similar communication units, e.g., phones.

Planes can be envisioned to be aligned along the temporal axis but the information stored in them need not be chronologically ordered since unit locations can be represented freely by forward and backward links, e.g., in semantics where interpretation is possible only when an entire utterance has been processed.

2.1. Purpose

Speech databases contain information that can be transformed into knowledge to further improve the understanding of spoken and written natural language. The efficiency of this transformation is dependent upon the quality of the representation which is formed of the data. An inadequate or structurally incorrect representation will severely limit the transformation and allow only a fraction of a speech database's potential inferences to be formed.

The purpose of a communication framework is twofold. It primarily serves as a structure that binds and links together the different communication planes of an utterance much like a multi-layered blackboard, see figure 2. Information from different communication planes can then be analysed in a context-rich environment, i.e., in the presence of other related information. Inferences generated from analysing the links between different communication planes, levels, and units can then be used to generate other levels. These new levels, possibly higher or lower, on the same or different plane, can be utilised in actual language applications, e.g., the transformation of linguistic words to phones via a phoneme level in speech synthesis, or, the suitability of different word hypotheses in a semantic context in speech recognition as a function of time, etc.

Many competing theories exist for both spoken and written natural language. Differences exist in the constraints they place on the representational resources required to create models of actual speech data. The second purpose of the communication framework is to support and allow for the instantiation of these different theories in the form of models so that theories can be developed and evolved in a common environment.

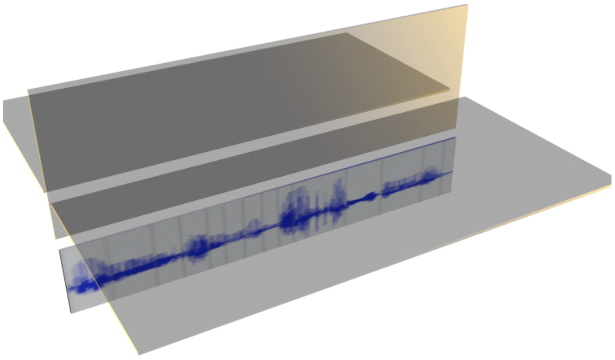

Figure 2: A newly created communication framework containing only acoustic waveform data. Parsed objects from an utterance in a database can now be stored in the other planes, e.g., linguistic, phonetic, and orthographic.

3. FROM STORAGE TO GENERIC LABELLING FORMAT

A diverse variety of speech database formats presents the first hurdle in developing a system that can read in different speech databases independent of file format, character set, language, phonetic alphabet, etc. This problem is solved by defining a *database specification model* (DSM) for each different format of database. The model encompasses information such as:

- file directory, name, and extension structure
- waveform information description format, e.g., AIFF, WAV, NIST headers, etc.
- symbolic information description format for each type of information as well as plane/level/unit storage destination in a communication framework

For example, in the Kiel corpus symbolic annotations for an utterance reside in the same file while in TIMIT separate files are used—this knowledge is stored within a DSM. A parser using specification information supplied by a DSM segments the annotation data into communication units which are objects that can be presented to the user in a labelling frame for viewing and editing if necessary. These units are also possibly given information regarding their language, phonetic alphabet, transcription style, if applicable, and temporal interval, if determinable. The DSM also indicates to the parser where data is to be stored in the communication framework. For example, after a Kiel waveform and its associated annotation file for an utterance have been read in, the system outputs a new communication framework object consisting of an acoustic and four symbolic planes: linguistic, phonetic, orthographic, and prosodic, since this information is supplied in the database. The symbolic planes contain levels and units where units are composed of textual strings and interval objects specifying their temporal extent. A signal model supporting deferred loading is used to transfer waveforms into memory only when samples are actually required via specialised methods that handle binary data efficiently.

4. PARSING COMMUNICATION UNITS

After speech data has been transformed into the generic labelling format described above, communication units are formed. These units are instances of classes that are given knowledge about their genetic makeup. Most notably, phonemes and phones inherit a rich set of phonetic features that can be effectively used during the database search phase. Figure 3 shows some of the primitive feature classes that an open-mid, front, unrounded monophthong inherits (ɛ in IPA).

4.1. Languages and Phonetic Alphabets

Since Worldbet [4] is used to define all possible speech sounds it is a straightforward task to define speech sound inventories for each language, e.g., Finnish, English, German, etc. Similarly, all known phonetic alphabets can be mapped onto the Worldbet phonetic alphabet since it represents the set of all speech sounds. Therefore, mapping from one phonetic alphabet to another is possible when the same sound exists in both alphabets, e.g., /e/ (SAMPA) and /eh/ (TIMITBET).

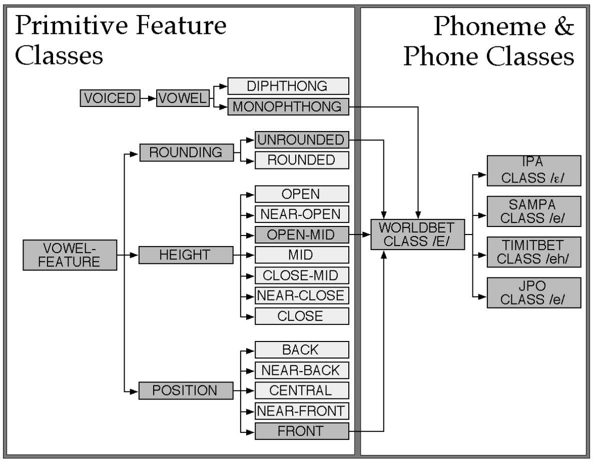

Figure 3: Partial class inheritance for the /ɛ/ (IPA) class.

4.2. Parsing Phonetic Labels into Units

With class-based resources defined such as languages and phonetic alphabets, parsers can be formed to interpret phonetic strings and diacritics that code allophonic variation. A parser decodes the sequence of characters that symbolically describe an annotation and generates one or more communication units that are placed into a level of a communication plane. Figure 4 shows part of the state machine used to parse symbolic data from the Kiel database. In Kiel, phonetic, linguistic, and prosodic information is coded into the same annotation string.

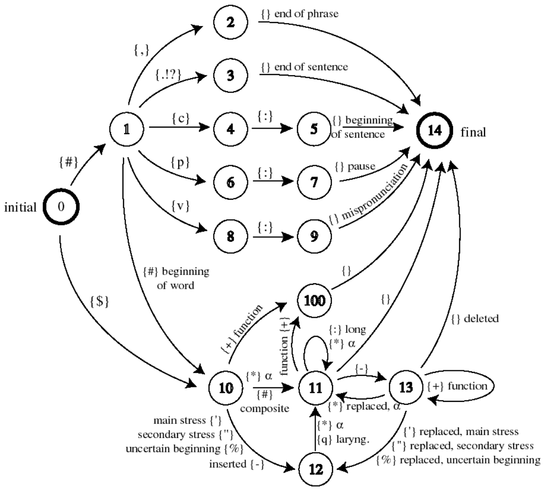

Figure 4: Part of the state machine that parses segment label data from the Kiel database.

4.3. Parsing Units in other Planes

Communication units of other planes can be parsed as well according to required criteria. For example, textual orthographic strings can be easily broken down into paragraphs, sentences, phrases, words, syllables, and graphemes by simply searching for whitespace. Since each communication plane has a predefined list of structurally ordered communication levels, this information can be used to indicate to parsers how far down in the hierarchy to proceed. Communication units are formed and placed into planes and levels as in the case of phonetic data but temporal information is not necessarily available at this point since no inferences in the form of cross-planar links have been made yet.

5. LINKING UNITS

When all possible communication units have been extracted from the symbolic descriptions of a speech utterance and placed into a communication framework, three different phases of linking are performed to generate a communication framework on which searches can be performed:

1. *Vertical linking*. Units that can be structurally organised according to a predefined hierarchical scale have their relationships identified using bidirectional links.

2. *Horizontal linking*. Temporal order is asserted by bidirectionally linking neighbouring objects on the same communication level.

3. *Cross-Planar linking*. If possible, communication units on different planes are linked to indicate their realisation in another representation space.

The top-right pane of figure 1 shows graphically the concept of vertical and horizontal links. Vertical and horizontal linking is performed during the parsing stage for some communication planes and levels. Vertical linking in its simplest form is accomplished by comparing temporal intervals and building a tree structure. Horizontal links are represented computationally for additional flexibility. Cross-planer linking is the most complex of these linking procedures since knowledge from at least both communication planes must be made available to the linker. For example, canonical phonemic level and phonetic phone level units can be related to each other to indicate whether the phone was elicited as expected, or whether an insertion, deletion, or replacement occurred. In some cases simple 1-to-1 links do not suffice to represent complex phenomena, e.g., a phoneme realised as two distinct phones. In such cases other classes of link objects are used. Componential residues—an indication that some remnants of a deleted phone still exist—can be modelled by functional closures or by specifying the remnants explicitly as local unit properties.

6. DATABASE ACCESS

Search predicates can be applied to linked communication frameworks. Since units are linked to one another, and the units know of their local and class inherited properties, functions can be defined to search out desired contexts. For example, to find all consonant-back-vowel-fricative contexts within a database the following Lisp function is applied to all phone level units in each communication framework:

```
(defun consonant-backvowel-fricative-p (x)
  (and (typep (prev x) 'consonant)
       (typep x 'vowel)
       (typep x 'back)
       (typep (next x) 'fricative)))
```

In the above case a search returns matching phone units. Since phones are temporally defined they have access to acoustic speech waveform and signal processing methods can be applied directly to them. The same search function can be applied to databases containing speech from different languages and transcribed in other phonetic alphabets since communication frameworks model actual speech and not database annotation dependent information. Defined search functions form a library that can be called from more complex functions. In the above example only horizontal links were utilised but vertical and cross-planar links are available for use as well. See [5] for examples of analyses.

7. FURTHER WORK

Currently the system is able to generate communication frameworks for some or all communication levels supplied by the following speech databases: TIMIT, Kiel, ANDOSL, and our own Finnish database. Work is underway to write parsers for the missing levels and include other databases as well.

8. REFERENCES

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

2. Kohler, K., Personal communication. University of Kiel. July, 1998.

3. Vainio, M., Altosaar, T., Modeling the Microprosody of Pitch and Loudness for Speech Synthesis with Neural Networks. In these proceedings.

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

5. Altosaar T., Karjalainen M., Vainio M. A Multilingual Phonetic Representation and Analysis System for Different Speech Databases. In Proceedings of ICSLP 96, Philadelphia, 1996.