A KNOWLEDGE-BASED APPROACH TO UNLIMITED VOCABULARY SPEECH RECOGNITION FOR THE FINNISH LANGUAGE

Toomas Altosaar and Matti Karjalainen
Helsinki University of Technology
Acoustics and Speech Processing Laboratory
Otakaari 5A, 02150 Espoo, Finland

ABSTRACT

This paper describes a strategy for the development of a speech recognition system for Finnish that is based on several different knowledge sources which include auditory, phonetic, and linguistic details. Auditory knowledge is derived from the application of computational models which simulate the human peripheral hearing system. Phonetic knowledge is represented by rule-based analysis, parsing and classification of phonetically relevant units and structures from the output of the auditory model. Finally, linguistic knowledge is used to filter the word forms generated by the phonetic level by accepting or rejecting hypotheses through the use of morphological analyses. The main components of the system including the structure of the rules are presented.

INTRODUCTION

Speech recognition methods based on signal processing and numerical algorithms have been more successful in practice than methods where AI, knowledge-based and symbolic computations are utilized. However, to solve the most fundamental problems in advanced speech recognition the knowledge-based approach seems to hold the most potential.

Automatic recognition of unconstrained speech is a very difficult pattern recognition problem [1]. Research efforts in this domain normally must use some kinds of constraints for the input speech. Typically for example, the allowed speaker population is small, the words need to be pronounced with distinct pauses between them, and the recording quality is required to be high and constant over talking sessions. The human recognition system on the other hand can adapt to a wide range of speakers and languages, and can perform the transformation from acoustic wave to implied message with relative ease. It is important to try to understand and simulate the human process, if possible, since its performance far exceeds any artificial recognition system. Auditory modeling - the simulation of the human peripheral auditory system on a computer allows us to investigate this area by deriving models and comparing them to the human system.

Several different ways exist in which auditory modeling can be realized [2]. A natural way is to simulate the physiological function of the basilar membrane and the cochlear processes up to the neural levels. Another method is to use the results from psychoacoustics, the study of auditory perception at the psychological level, which relates acoustic signals to what the human listener perceives. A third approach to modeling, which is emphasized in this paper, is to hypothesize higher level functional principles that possibly could be found in the hearing system.

The human auditory system is able to accept and correctly interpret signals over several resolutions in the time and frequency domain. Therefore in a computational model the analysis should be carried out without any preassumed scale, frame-length, duration or frequency resolution. The input itself should create and define a hierarchical and structural description which can be analyzed by a set of rules that contain speech specific knowledge. By creating a natural representation of the speech higher levels of performance can be achieved in later processing stages.

The application of Multiple-Resolution Analysis (MRA), the study of signals over many different scales and resolutions, in the domain of speech processing is especially appropriate since experimental evidence exists which shows that the human auditory system may function in a similar fashion. The human system can apply context-dependent resolutions in both the time and frequency domain, e.g. forward and backward masking, and Chistovich's 3.5 Bark spectral integration of formant clusters [3]. Several studies, such as Witkin's scale-space filtering for signal interpretation [4], the dendrogram by Glass and Zue [5] for acoustical segmentation of speech, the wavelet transform to analyze and synthesize signals by Grossmann et al. [6], and the curvature primal sketch by Asada and Brady [7] for image analysis, emphasize the importance of allowing the data itself to form its own and natural representation. Another feature of MRA is that it avoids thresholding away data as much as possible allowing access to information that may be required in later processing. We believe that in human perception unconstrained and natural representations are formed at many different levels - from the acoustic level up to the thought level.

In this paper we describe an experimental knowledge-based unlimited vocabulary speech recognizer for the Finnish language that is based on knowledge-based and MRA concepts.

RECOGNITION STRATEGY

The recognition system is composed of several different units and its structure is shown in Figure 1. At the bottom level there exists an auditory model which transforms the time waveform into a series of spectra. These spectra form the input for a dendrogram construction algorithm which creates a multilevel-segmentation of the speech. The nodes of the dendrogram are then coarse classified and a path found through the structure that represents a probable phoneme sequence. Fine classification then takes place in a neighborhood of the path and may form several possible phoneme sequences. These sequences are then fed to a morphological filter that dramatically improves the performance of the recognizer. The system is implemented on a Symbolics Lisp machine using the QuickSig object-oriented signal processing package [8]. Objects as well as a powerful set of relation-mixin flavors are used to represent the structures.

Auditory Model

The auditory model acts as the front-end processor converting the acoustic wave into a representation that emphasizes aspects of the signal that are important to human perception. In our model we use 48 channels spaced equally across the Bark (critical band) scale at 0.5 Bark intervals [9]. The model includes the most important aspects of the peripheral system: the ear’s sensitivity curve, critical band shape according to Schröder et al. [10], and loudness scaling. Forward and backward masking details have been left out in this recognizer’s front-end - their effect is represented to some degree in the rules found in the phonetic stage. An auditory spectrogram of the Finnish word /yksi/ can be seen in the middle window of figure 2. A spectrum is calculated every 5 ms and yields sufficient temporal resolution. A recent study has shown that a speech recognition system can attain better recognition rates when using an auditory front end as compared to a more conventional Fourier transform representation [11]. This provides good motivation to continue investigating the best auditory front end known as of yet.
The rules themselves are represented as method functions of a node object class that know how to evaluate themselves within a certain context. A typical evaluating rule to determine a node's fricative membership value may be written as:

\[
\text{fricative-membership (node)} \equiv \begin{cases} 
  (\text{fricative-index} \cdot (\text{duration-index} \cdot \text{rms-normalized inner product})) & \text{if} (\text{duration-index} \geq 0.08 \text{ and } \text{rms-normalized inner product} \geq 0.3) \\
  0 & \text{otherwise}
\end{cases}
\]

This rule places some loose constraints on what properties an unvoiced fricative should have, e.g., its duration should be roughly between 80 and 300 ms, its voicing level should be minimal, and that its loudness level should be more than 20% of the local neighborhood's level. It combines these facts into a product and assigns the node a fricative membership value.

After the initial coarse classification has been performed fine classification of the nodes may be performed. However, since there are a large number of nodes many hypothetical phoneme paths exist from left to right through the tree. To reduce the computation time required to perform the fine classification an active path is first found through the dendrogram which indicates a likely phoneme sequence. The path finding algorithm performs three passes: the first pass determines for each node a sub-structure stability index which indicates the similarity of a node's offspring according to loudness levels, spectral characteristics, and temporal durations. The second pass forms a new estimate for the stability of a node by forming a product composed of its own first pass stability and a weighted combination of its children's first pass stabilities. The final pass forms a derivative along the increasing scale dimension to indicate where the largest change in stability has occurred. The final value is also weighted by the node's duration and units between 40 and 350 ms are accepted without penalty.

A path is then found through the tree by searching for the nodes with the highest third pass values. These nodes are included in the path while ensuring that a complete covering in terms of time is achieved, i.e. a continuous path exists through the tree from left to right. The nodes which are outlined in the dendrogram of figure 2 indicate the path that was found.

This algorithm has been found to be quite reliable and generates only a few deletion and insertion errors. In one sense finding a single path through the dendrogram and using it only is contradictory to the multiple-resolution analysis philosophy since a kind of data thresholding has been applied. However, the number of all possible paths is so large that some type of limiting mechanism is required to guarantee a practical computation time. The active path or maximum stability covering, should be interpreted only as a good first estimate to a possible phoneme sequence to which necessary perturbations may be made. When path errors exist it is the responsibility of higher level processing to make corrections.

Since a relatively good estimate of phoneme-like units has been formed in the dendrogram structure a more careful evaluation may be made in the active path's neighborhood regarding the specific auditory characteristics of nearby nodes. Gross spectral characteristics have been used as the basis for determining phoneme type by filtering spectra with a second derivative bandpass Gaussian kernel and calculating a rms-normalized inner product. Diphones (the transition from one phoneme to another) have been used as the basis for templates. Each template consists of three filtered auditory spectral samples taken from real speech signals and includes natural coarticulation effects. A set of 11 diphone class groups as shown in table 2 have been used as a preliminary rule set and cover only a part of the transitions which exist in the Finnish language thereby restricting the vocabulary of the recognizer.

Each of the 11 diphone groups has a matching rule associated with it that compares a node from the dendrogram structure with a three spectra template taken from real speech. The evaluation is performed from right to left, i.e. the stop-fricative class evaluates a fricative node preceded by a stop. The membership values from the coarse classification stage defined in a broad sense what type of segment is in

![Figure 1 Recognition strategy.](image-url)
question and are now used to limit the amount of processing required to determine exact phonemes. For example, a node that has a vowel membership value less than 0.1 from the coarse classification stage will not have the stop-vowel matching method applied since it is very unlikely that it represents a vowel. These thresholds have been placed low enough so that no errors have resulted from missing classes while decreasing the computation time significantly.

Diphone Group  Group Members  Number of Templates
stop-vowel  /a/ /a/ /a/ /a/ /a/ /.../  24
stop-fricative  /a/ /a/ /a/ /a/ /a/ /3 contexts/  27
pause-vowel  /a/ /a/ /a/ /a/ /a/ /3 contexts/  9
vowel-pause  /a/ /a/ /a/ /a/ /a/ /8/  8
nasal-vowel  /n/ /n/ /n/ /n/ /n/ /8/  8
nasal-pause  /n/ /n/ /n/ /n/ /n/ /3/  3
vowel-fricative  /v/ /v/ /v/ /v/ /v/ /8/  24
vowel-nasal  /v/ /v/ /v/ /v/ /v/ /8/  24
vowel-pause  /v/ /v/ /v/ /v/ /v/ /8/  24

Table 2 Diphone groups supported by the recognizer.

The diphone match results are stored in diphone objects and are linked with the part-of-speech information to the nodes which have been evaluated. This diphone comparison along the active path can be seen as the right branch along the phoneme recognition strategy shown in figure 1.

Along with the diphone match a series of no-context spectral comparisons are carried out for the vowel class. Fricatives, nasals, and stops receive as their no-context-match values their corresponding coarse classification values. The pause (symbolized as [ ]) receives the same value as a stop except that it is dependent upon the loudness level following the node. This classification of phonemes in the active path with no context is seen in the left branch of figure 1.

Once all nodes in the active path are evaluated in both the diphone and no-context-match manner a voting procedure is applied that determines the node's final phoneme membership values. Figure 3 shows the nodes in the active path that have been linked with diphones which are represented by lines. The five diphone pairs with the highest membership values are indicated above these lines. The final phoneme memberships are listed inside the nodes.

Phoneme sequences are then generated but since more than one possible phoneme path may exist due to the diphone search that was carried out all generated phoneme paths are used. This may occur since some nodes may have been skipped by the diphone rules due to a node's time duration. For each possible path a set of phoneme sequences are recorded as well as an index of probability which is currently calculated as the number of phonemes in the sequence) of the product of each phoneme's membership value. Other ways of calculating the probability also exist which would not weight equally all phonemes in the set. For instance, by emphasizing vowels and fricatives more than nasals and stops a more pertinent probability value might be obtained. Once all paths have had their phoneme sequences generated they are merged and sorted according to their probability index. Table 3 shows the first ten generated phoneme sequences for the Finnish word /yksi/.

As is seen in the above example the correct form was found as the second best hypothesis. A system is needed that can select from the set of phoneme strings the correct word form.

Morphological Filtering
Finnish exhibits a high correlation between phonemes and graphemes making the mapping between the two relatively straightforward. Finnish is however a highly inflected language with many cases and possible suffixes. Since a single word may have several thousand different forms a simple spelling checker is not feasible. MORFO [12], a morphological analysis program developed by the SITRA Kielikone project, is capable of accepting word hypothesis forms and determining whether they exist in Finnish. Hypotheses are passed to MORFO in the order of highest probability effectively filtering out erroneous word forms. Thus in the above example the first form yke would be removed since no such word exists while the second form yksi would be accepted.
RESULTS

Three very preliminary recognition tests were performed with the limited set of recognition rules available. Fifty test words were chosen randomly from a Finnish dictionary and were used in their base form with no inflection and the performance of the recognizer measured. Phonemes which were not yet supported by the recognizer were not chosen (i.e. words containing /l/, /t/, /j/, /v/). Speech material from two male speakers was used: one a native speaker of Finnish who also supplied the spectral templates and another non-native speaker of Finnish. No difference in performance level was noticed: this was probably due to the small test set size (75). On the average 77% of the phonemes were correctly identified. On a word level 21% of the hypotheses had the correct grapheme sequence as the most probable choice. After passing these sequences through the morphological filter the recognition rate rose to 60%. The depth of the correct hypothesis in the sequence was on average 7.

The diphone rules exhibited a high variance in performance since some performed well while others were consistently poor and could be improved. Comparison of performance results on this system to other unlimited vocabulary recognition systems is difficult since no equivalent systems are currently known of. However, large vocabulary recognition systems such as the Tangora 20,000 word speech recognition system developed by IBM [13], achieve 95% recognition rates on a word level. It can be expected that this value would be reduced if the vocabulary was made unlimited so that the results obtained with the recognizer developed in this study are reasonable. A practical system should achieve at least a 95% rate while the long term goal for such systems is over 99%.

DISCUSSION

This paper presented a prototype speech recognition system based on knowledge-based multiple-resolution analysis principles.

In theory this experimental recognizer could be completed so that it would cover all phoneme groups if the missing diphone rules were added. The results indicated that this method holds promise but a more effective way must be used to generate the rules since their writing and tuning was difficult. Currently a set of method functions that contain speech-specific knowledge are used to evaluate phoneme classes and diphone hypotheses.

Missing from this system altogether is the desirable feature of adaptation or learning, a very important element in a commercially viable speech recognition system. Also, a more flexible method is required to represent diphone and phoneme templates, i.e. a level of parametrization would be helpful. In our future work on a connected word recognition system we plan to use a syllable based model since it seems to provide for a natural way to represent speech units.

The concept of context-propagation which can be defined by the way a local environment propagates to and influences the evaluation of another environment seems to be a promising area of research. It perhaps has similarities to neural computing but should be considerably more efficient and should represent knowledge in an explicit form.

The function of the morphological filter to accept only legal word candidates has proven to be a very useful feature of the recognition system and has decreased the error rate considerably. A more interactive communication between the acoustic, linguistic and phonetic levels would be useful. This would allow higher integration levels to exist enabling the language to constrain the search effectively.

The current prototype system operates at about 400 times real time on a Symbolics Lisp machine making the recognition process tedious. Third generation signal processors should decrease the processing time required considerably but a more efficient way of computation needs to be found.

ACKNOWLEDGEMENTS

This work has been financed by STIRA (the Finnish National Foundation for Research and Development) and the Finnish Academy.

REFERENCES