AN EVENT-BASED APPROACH TO AUDITORY MODELING OF SPEECH PERCEPTION

Toomas Altosaar and **Matti Karjalainen**

Helsinki University of Technology, Acoustics Lab.
Otakaari 5A, 02150 Espoo, Finland
Tel. (358-0) 451-2794

ABSTRACT

Auditory modeling is usually based on peripheral physiological phenomena. It is found, however, that this basis is not sufficient in all applications, e.g. in successful speech recognition. Our opinion is that more important than the details of periphery is to include higher-level functional processing in the models. This paper describes an experimental system that uses several spectral and temporal representations to create a hierarchical description of speech. The front-end processing is performed by an auditory model which is based on psychoacoustical principles. Several temporal and spectral representations are extracted from the resulting auditory spectra and are viewed under multiple time resolutions to yield reliable and flexible descriptions of the speech. Based on these spectral and temporal resolutions prominent extrema are located and are classified as objects called events. These objects are organized into event lists according to masking criteria and measures of prominence.

INTRODUCTION

The usual basis for auditory modeling is peripheral physiological phenomena. Transmission-line or filter-bank models are used for basilar membrane and neural models for the next stage, e.g. [1], [2]. This may give a detailed picture of the periphery but the models tend to become overly complicated and there is a certain lack of knowledge of how the higher levels work.

Another approach to auditory modeling is to apply psychoacoustical theory and knowledge. Here we can concentrate on wider functional properties of hearing that are not always directly related to physiological details. Surprisingly few models are explicitly based on psychoacoustics.

The limited success of auditory modeling in speech recognition shows that an auditory front end does not necessarily solve existing problems. We have to pick up the most essential peripheral features and combine them with higher-level symbolic processing. With this approach we are immediately faced with several problems, some of which we hope to solve by formalisms proposed in this paper. There is not much hope to find principles with evidence and support from concrete hearing research. Instead we, have to use hypothetical models that could be possible in the human auditory system.

The central problem for us appears to be in the transformation from a continuous-time speech signal to a discrete and symbolic representation without loosing any key information. The traditional pattern matching and decision process isolates the continuous and discrete domains in a way that makes it very hard to pay attention to the most essential features in a given context.

There are several concepts that we have found to be important. Retaining redundancy with multiple feature representation at each level of the auditory process and even multiple resolution analysis of each feature is needed. This presumes parallel processing to a large extent if such a system is to be implemented in real-time.

Other key concepts in our approach are events and event stuctures. Instead of segments with time boundaries we analyze events (time objects) with rich internal structures: time moment, effective time span, type according to several criteria, amplitude or prominence, link to a feature it is supported by, etc. The list-like data structures consisting of events form the basis for flexible representations that can be applied to rule-based processing at several levels of auditory modeling.

The prototype system to be presented in the next section reflects our approach in a preliminary form. It should be considered as a collection of examples to be developed towards a future speech recognizer that includes all phases from a peripheral auditory model to natural language processing.

SYSTEM DESCRIPTION

The system contains many different levels of processing ranging from auditory modeling of the speech input to symbol and event processing. Figure 1 shows an overview of the current and proposed system. The following sections explain how the system functions.

Auditory Front End

The system obtains auditory information from a filter bank that closely matches the human auditory system in terms of sound perception. The model [3] is based on the most important features of peripheral hearing known from the theory of psychoacoustics [4] and simulates the human's frequency selectivity and sensitivity as well as its temporal and masking properties. By the use of this model only relevant auditory spectral information is retained. Irrelevant information is efficiently removed during the early stages reducing the computation rate in later processing.

The auditory model is implemented as a filter bank and its output is represented by a 48 element spectral vector for each point in time. The vector's elements indicate approximately the amount of energy falling in 1 Bark (1 critical band) regions of the auditory spectrum and are scaled in loudness [4]. Each channel of the filter bank is separated by 0.5 Barks and this provides adequate frequency resolution over the entire 24 Bark auditory spectrum. A spectrum is calculated every 10 ms.

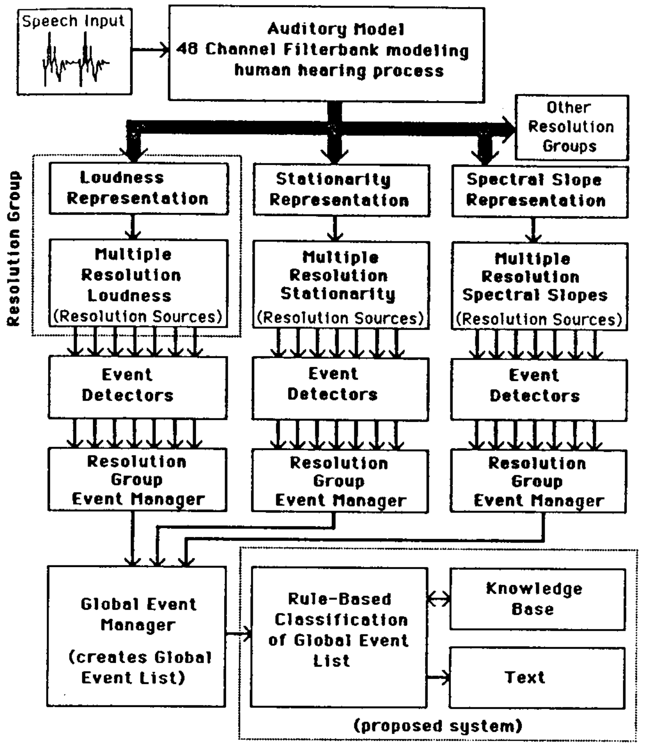

Figure 1. *Signal-to-Symbol Speech Analysis System.*

Multiple Representation Analysis

The loudness scaled auditory spectra are transformed into several parallel representations which help to identify the different speech features and events. These representations can be separated into two major groups: the frequency domain, and the time domain. These groups are described in the following sections.

Frequency Domain Processing

The frequency resolution of the hearing system to broadband signals is at best 1 Bark. For phonetic classification of speech signals different studies have shown that this can vary from 1 to 3.5 Barks. We can simulate this effect by bandpass filtering the spectrum in the *frequency* domain to emphasize the desired resolutions. This bandpass filtered spectrum representation is called the **formant spectrum.** Adequate resolution has been achieved for this system with both 1 and 2 Bark bandwidth filters. The basis for use of multiple resolutions for a single representation is explained later on. The formant spectrum can be used to identify the existence and locations of formants and formant pairs. Formant lists are created by searching for local maxima and indicate where likely formants exist as well as what their amplitudes are but no attempt is made to classify them. The Formant lists are used in an auditory spectrogram display which is shown in figure 2.

Time Domain Processing

The other category of representations are based upon information that the front end supplies in the time domain. One such representation is **total loudness** and is calculated by summing the elements of a loudness spectrum. Total loudness as a function of time reveals the temporal energy structure of the speech while being independent of the individual spectral components.

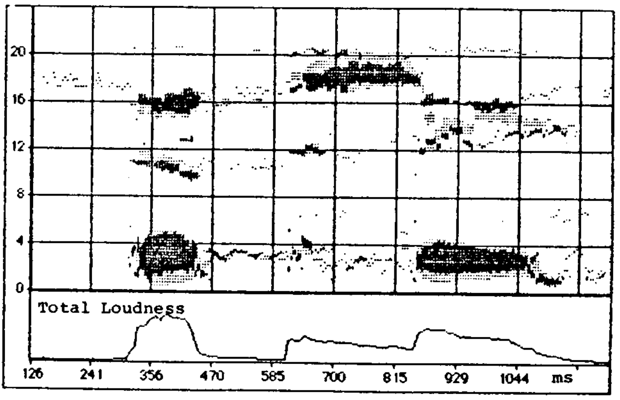

Figure 2. Auditory Spectrogram of the Finnish word /yksi/.

Stationarity is a representation that measures changes in the spectra by comparing the similarity between neighbouring spectra. Stationarity is calculated for time t_i by first finding the average spectra at time t_{i-j} and at time t_{i+j} (average computed over several spectra) and then summing the absolute difference between these averages to yield a scalar measure of distance. This representation is used to identify locations where spectral changes occur and indicates most phonemic boundaries with good reliability. Stationarity is sensitive to both spectral and amplitude changes in speech.

Another representation used in the system is **spectral slope** which indicates where the majority of the energy lies in the spectrum. Four different representations of spectral slope are used: global, formant 1, formant 2, and formant 3 slope. Global slope is a wideband locator of spectral energy while the remaining three analyze the regions of the spectrum where each formant is generally found. These functions are robust indicators of certain features such as fricatives and plosives and can also be used to detect spectral centers of gravity [5].

Time domain multiple representation analysis views the speech signal with several different but parallel perspectives. Figure 3 shows the responses of three representations to the Finnish word /yksi/.

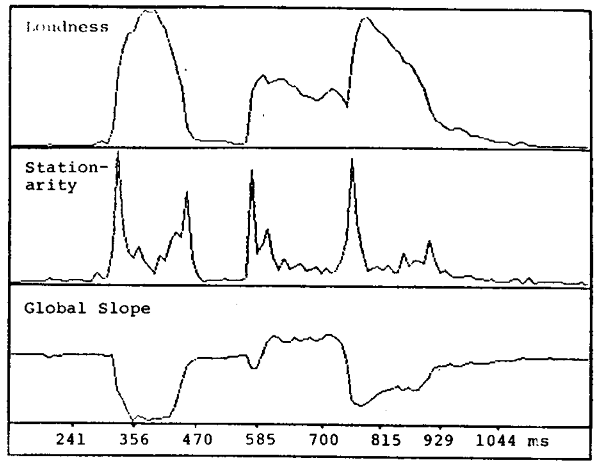

Figure 3. Multiple Representational Analysis of the word /yksi/.

Multiple Resolution Analysis

To obtain a more flexible description of each frequency and time domain representation, all representations are analyzed under several resolutions. This is performed by bandpass filtering a representation with filters having different resolutions. The impulse responses for some of these filters are shown in figure 4. For the frequency domain representation the loudness spectrum is filtered with 1 and 2 Bark resolutions as was mentioned earlier. In this case the filters are scaled in frequency. In the time domain representations the filters are scaled in time, and resolutions of the loudness, stationarity, and spectral slope representations are calculated in a similar way. This method is similar to *scale-space filtering* [6] and is used to generate qualitative descriptions of signals.

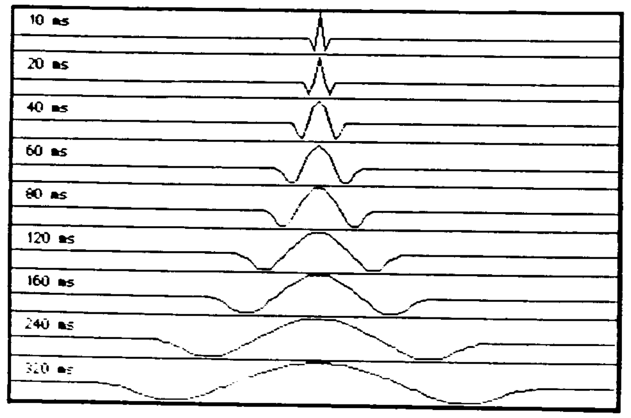

Figure 4. Impulse responses of some of the filters used in Multiple Resolution Analysis.

Each resolution of a representation is defined as a **resolution source** while the representation along with its resolutions is defined as a **resolution group**, as indicated in figure 1. Speech analysis with multiple resolutions facilitates determining event locations and their respective properties with greater ease and accuracy than would be possible with the original representations alone. The curves in figure 5 show the response of the loudness resolution group to the word /yksi/. The multiple/parallel representations and their resolutions allow for a reliable description to be created of the speech. New resolution groups may be added to the system such as pitch detection and a voiced/unvoiced indicator as is found necessary.

Event Detection and Analysis

The next phase of processing transforms a signal, in this case a resolution source, into a discrete and symbolic representation. The resolution groups are operated upon by event detectors which find local extrema and zero-crossings, depending upon which resolution group is being analyzed, and yield symbols as their outputs. Symbols are more flexible to manipulate during later stages of processing than signals since partial classification has already taken place. These symbols may contain information regarding their type, time, amplitude and formant structure. The symbols are ordered chronologically and are placed in a list for later processing.

The resolution group event manager is responsible for analyzing a resolution group and finding the most prominent areas of interest. One measure of prominence is determined by searching for the event with the largest absolute amplitude. It uses as its input the lists of symbols presented to it by the event detector. The resolution group event manager operates on these lists to produce a single list called the resolution group event list that contains the most significant events from a representation.

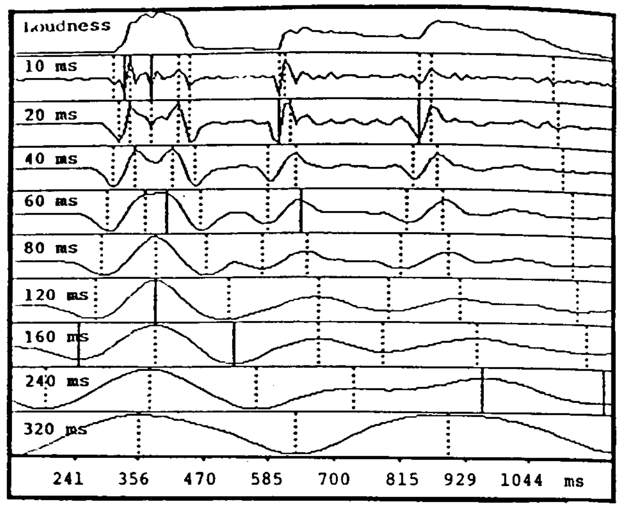

Figure 5. Multiple Resolution Analysis of the Loudness Representation for the word /yksi/. Solid lines indicate events, dashed lines indicate related events.

To avoid multiple entries of the same event in the list, all related events from different resolutions are marked as belonging to the most prominent event. Figure 5 also shows the events (solid lines) found for the multiple resolution loudness representation as well as the related events (dashed lines). Another measure of prominence that could be used is to choose the event with maximum span over the sigma axis when using scale-space filtering techniques [6] to yield a top-level descriptor.

An integrated description of the speech is constructed by the gobal event manager and it considers all the resolution group event lists created by the resolution group event managers and builds a global event list that contains the most prominent events.

The final set of symbols created by the global event manager have been proposed to be used as a primary representation of the speech in a rule-based recognition system. These symbols would describe speech in similar terms as a human would when reading a spectrogram or by listening to speech. The rule-based system would analyze these symbols until enough evidence existed to fully support a hypothesis for final classification. By deferring classification to this final stage, diverse sources of information may be viewed in a global perspective making high rates of recognition possible.

IMPLEMENTATION

The preliminary version of the model is currently implemented on a two processor system. The auditory model filterbank is realized on a TMS 320 signal processor and the remainder on an Apple Macintosh. The Macintosh is the host for the TMS and executes NEON which is an object oriented language [7]. NEON is a hybrid language with many of its features derived from Forth and Smalltalk. The next extended version of the program is being currently implemented on a Symbolics 3670 Lisp Machine including a small-scale speech recognition system.

To efficiently represent and manipulate the different resolutions, representations and symbols, object oriented programming methods are used. Object orientation is a powerful data and knowledge representation principle since knowledge regarding the object is contained within the object itself thus exhibiting object-centered control [8],[9]. Objects can communicate with each other by message passing methods. They also belong to classes and can inherit properties from other classes. This approach allows for building rule and frame-based systems.

Since each representation's analysis can be processed independantly, parallel-processing of the representations, resolutions, and events is a natural topology for the implementation of such a system. Such a concurrent system could be implemented e.g. using Transputers [10] and is one of our long-range goals.

CONCLUSION

Higher-level functional processing must be included in auditory models if the information they supply is to be of greater use. This is because peripheral physiological phenomena often does not offer a sufficient basis for applications such as speech recognition. In this paper we have described an approach to implant the higher-level processing activities into an auditory model. The conversion of speech into a loudness spectrum, the derivation of some representations, and the analysis of these under multiple scale resolutions was explained. Finally, the transformation of signals into discrete frequency/time events was described.

ACKNOWLEDGEMENTS

The authors are grateful to the Finnish Academy which has financed this research.

REFERENCES

[1] Lyon, R.F., A Computational Model of Filtering, Detection, and Compression in the Cochlea. Proc. of IEEE ICASSP-82, Paris.

[2] Lyon, R.F., Computational Models of Neural Auditory Processing. Proc. of IEEE ICASSP-84, San Diego.

[3] Karjalainen, M.A., A New Auditory Model for the Evaluation of Sound Quality of Audio Systems. Proc. of IEEE ICASSP-85, Tampa.

[4] Zwicker E., Feldtkeller R., Das Ohr als Nachrictenempfänger. S. Hirzel Verlag, Stuttgart, 1967.

[5] Chistovich L.A. et al., "Centres of Gravity" and Spectral Peaks as the Determinants of Vowel Quality, *Frontiers in Speech Communication Research,* Academic Press, 1979.

[6] Witkin, A.P., Scale-Space Filtering: A New Approach To Multi-Scale Description. Proc. of IEEE ICASSP-84, San Diego.

[7] NEON Programming Manual, Kriya Systems, Inc., Sterling, 1986.

[8] Winston, P.H. *Artificial Intelligence* (second edition), Addison-Wesley, 1984.

[9] Waterman, D.A. *A Guide to Expert Systems,* Addison-Wesley, 1986.

[10] Transputer Reference Manual, INMOS Limited, October 1986.