Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

Some current projects at KTH related to speech recognition

Blomberg, M. and Carlson, R. and Elenius, K. O. E. and Granström, B. and Hunnicutt, S.

journal: STL-QPSR
volume: 27
number: 1
year: 1986
pages: 031-040

KTH Computer Science and Communication

http://www.speech.kth.se/qpsr

II. SPEECH RECOGNITION

A. SOME CURRENT PROJECTS AT KTH RELATED TO SPEECH RECOGNITION*

M. Blomberg, R. Carlson, K. Elenius, B. Granström, and S. Hunnicutt

Abstract

Understanding and modelling the human speech understanding process requires knowledge in several domains, from auditory analysis of speech to higher linguistic processes. Integrating this knowledge into a coherent model is not the scope of this paper. Rather we want to present some projects that may add to the understanding of some components that eventually could be built into a knowledge-based speech recognition system. One project is concerned with a framework to formulate and experiment with the earlier levels of speech analysis. Others deal with different kinds of auditory representations and methods for comparing speech sounds. Still another project studies the phonetic and orthographic properties of different European languages.

Introduction

This long-term research goal at our department is to accumulate knowledge about the speech communication process and to apply this knowledge in models for, e.g., speech synthesis and recognition. Much of the research is not specifically oriented towards any single application. At present we lack a unified framework for compiling the research results into a knowledge-based speech recognition system. Also we certainly lack much of the knowledge necessary to build such a system, even approximating the human speech understanding capability. Rather than giving a full account of a single project, we have in this paper chosen to give an overview of several projects in our group related to speech recognition. Interested readers are in each case referred to more detailed reports.

A parallel speech-analyzing system

Inspired by the parallel nature of human speech perception we have set out a framework to formulate and explore speech analysis/recognition models. The speech is represented as a continuous flow of information in multiple channels. A flexible notation for describing interactions within and between channels is included. Graphic representations of the processing could be created on line to facilitate evaluations and experimentation. The system has been used to formulate the lower levels of speech analysis including spectral transformations, lateral inhibition, temporal onset/offset effects, and a variety of phonetic cue-detectors. Interactions on higher levels such as lexical access still need to be worked out.

---

*This is an overview paper presented to the International Workshop on Recent Advances and Applications of Speech Recognition, Rome, May 1986.

The work is heavily inspired by the feature notation, which has proved to be a powerful tool to use for phonetic descriptions and phonological transformations. A broader classification has also been explored in lexical search. Rather than using features in our approach, we will use continuous parameters that represent cues. The goalwith our approach is to explore the descriptive power of cues, and to use multiple cues to analyze, classify, and segment the speech wave. A successful model will help us to better understand the acoustic representation of speech and can also be implemented as part of a speech recognition system.

Parallelism is the most important motivation for our current approach. Speech recognition systems cannot be based on simple decisions involving few parameters and little or no complementary supporting cues. This model should make it possible to use diverse analysis mechanisms which can be simple but should work in a coordinated structure.

The computer program RECSYS consists of a pipe line in which a spectral representation is used as input and stored in a number of input channels. At each sample time, all data in each channel are moved like a delay line, and new input is stored in the beginning of the line. It is thus possible to study not only one but a sequence of spectral representations in a time/value matrix. Between each move in the pipe line, a number of user-defined analyses takes place. These can be regarded as new cells that connect to a certain number of elements in a matrix. Each connection can be activating or inhibiting. Such an analyzing structure is called a UNIT. The mathematics in each unit can be simple, but groups of units form complex patterns. The user or model builder designs such a unit by writing a simple definition telling which elements in the matrix the unit is connected to, and the influence of the contents of the matrix elements on the unit's result. By this definition, a new active line in the matrix corresponding to the output of the unit is created and can be connected to other units. The result stored in this line is moved in synchrony with all data in the matrix.

Each unit can be connected to all other positions in the matrix and detailed acoustic information can be combined with gross feature analysis. This will result in a system that has no explicit levels unless the user so wishes and expresses it by unit connections. A unit can be very specific or have a general function. The earliest ones might create a new spectral representation. Another kind of unit could measure the spectral balance or movements of energy in the spectrum. As the system becomes more complex, several units can measure cues such as voice onset time or degree of aspiration.

In Fig. 1 one type of analysis is illustrated. Stop cues have been chosen as examples since they are simple but not trivial. The BH1-17 units form a bark-band based spectral representation with enhancement in time and frequency. The STOP unit adds up information from several units. The stop is regarded as voiced (the VSTOP unit) if it has a voice-bar (VBAR) but the output is inhibited if the explosion has a high

level drop after it. The unvoiced stop (unit NSTOP) is simply the remaining component of the STOP unit response. The ASP unit gives response to aspiration.

A more detailed account of this project is given in Carlson, Granström, & Hunnicutt (1985).

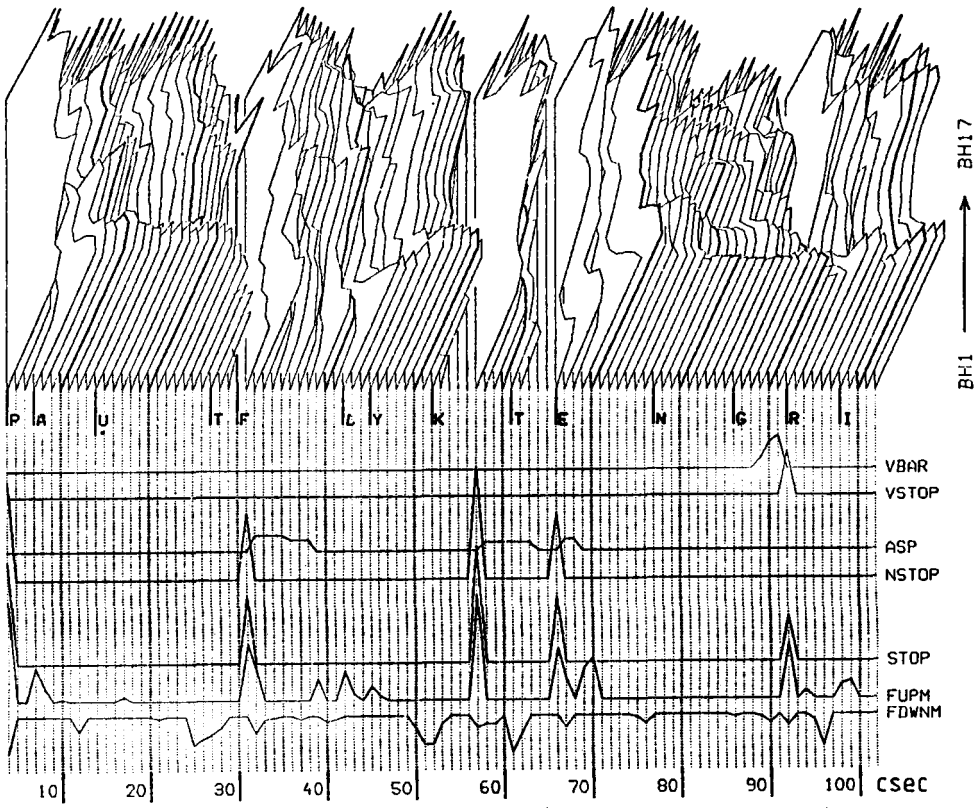

Fig. 1. Example of RECSYS output (see text for details).

Auditory models in isolated word recognition

The use of auditory models as speech recognition front ends has recently attracted a great deal of interest. The underlying assumption is that a good model of the auditory system should generate a more natural and efficient representation of speech compared to ordinary spectrum analysis. However, we have to keep in mind that only some of the peripheral processes of sound perception are included in most existing models.

A straightforward isolated word recognition system has been used to test different auditory models in acoustic front end processing. The models include BARK, PHON, and SONE. The PHONTEMP model is based on PHON but also includes temporal forward masking. We also introduce a model, DOMIN, which is intended to measure the dominating frequency at

each point along the 'basilar membrane'. All the above models were derived from an FFT analysis, and the FFT processing is also used as a reference model. One male and one female speaker were used to test the recognition performance of the different models on a difficult vocabulary consisting of 18 Swedish consonants in an /aCa/ context and nine Swedish vowels in an /hV1/ context.

MODEL	VOWELS		CONSONANTS		MEAN
	MALE	FEMALE	MALE	FEMALE	
FFT	99	96	95	97	**97**
BARK	99	95	92	95	**95**
PHON	96	91	88	91	**92**
PHONTEMP	95	85	82	88	**88**
SONE	91	93	82	83	**87**
DOMIN	99	99	90	90	**94**

Table I. Recognition accuracy in percent.

The recognition system is based on ordinary isolated word recognition techniques using pattern matching and dynamic programming. The Euclidean metric is used for calculation of distances between word patterns. The results are displayed in Table I where the recognition performance of each auditory model is shown separately for consonants and vowels for each speaker. The mean of each model is also given. The results indicate that the performance of the models decreases as they become more complex. The overall recognition accuracy of FFT is 97% while it is 87% for SONE. However, the DOMIN model which is sensitive to dominant frequencies (formants) but preserves no over-all level information performs very well for vowels.

It is obvious from our experiment that the unqualified assumption does not hold -- auditory models used as speech recognition front ends will not consistently improve performance. Several plausible explanations for our results could be mentioned.

- All models are based on the FFT analysis. The data will be smeared in frequency and time depending on the chosen approach.

- The models describe only a few selected ways in which the human auditory system processes data. They may be based on too specific experiments, capturing ways of processing the signal that is not very important for speech processing, and missing those that are.

- There is no match between the human-modelled primary analysis and the rest of the recognition system. If this match is required, the modelling could pay off only if the decision making part of the program models the way the central nervous system looks at the sensory data.

This should, of course, not hold us back from the interesting fields of auditory modelling and speech perception. It might, however, be premature to include our fragmentary knowledge of the auditory system in today's speech recognizers.

A more extensive report on these experiments can be found in Blomberg, Carlson, Elenius, & Granström (1984).

Nonlinear frequency warp for speech recognition

A usual way of measuring the distance between two spectral frames in a speech recognition system is to accumulate the absolute value of the amplitude differences along the frequency axis. However, recognition systems based on this method exhibit high sensitivity to factors with little phonetic information content such as fundamental frequency, voice source spectrum, formant bandwidths, transmission channel frequency response, etc. A possible way of reducing the influence of these effects would be the use of a formant tracking algorithm. Formant tracking is a very difficult task, though, which is why that method has not been used in practical systems.

The problem is to find a metric that is insensitive to other parameters than the formant frequencies. In this report, we have investigated a technique using frequency warping. The frequency axis of one spectrum is expanded or compressed in different regions to minimize the distance to another spectrum. This nonlinear warping is a way to align formants in the two spectra and the amount of warping might be proportional to the sum of formant frequency differences.

The frequency distance, defined as the area between the obtained warping function and the diagonal, is contributing to the spectral distance. The distance between two spectra is a weighted sum of the warped amplitude distance and the frequency distance. By changing two weights, we get a gradual shift between non-warped amplitude distance, warped amplitude distance, and frequency distance. In recognition experiments on natural and synthetic vowel spectra, a metric combining the frequency and amplitude distances gave better results than using only amplitude or frequency deviation. Analysis of the results of the synthetic vowels shows a reduced sensitivity to voice source and pitch variation. For the natural vowels, the recognition improvement is larger for the male and female speakers separately than for the combined groups.

A combined amplitude and frequency distance metric gives higher recognition accuracy than both conventional spectral distance and frequency warped amplitude difference for the separate male and female speakers. This suggests that the proposed technique manages to reduce the influence of different voice characteristics. In the combined male and female experiment, the formant frequency variation is larger and the recognition improvement is smaller. This indicates that formant frequency variation and possibly also pitch variation are not handled very

well by this method. The results conform with the original intention of emphasizing formant frequencies over spectral amplitudes. The best use of the method would probably be in speaker-dependent systems to reduce the influence of the voice source and transmission variation. Extending the method to speaker-independent recognition could be achieved by more detailed analysis of the warping function shape. This will be tried in further work as well as rules for establishing optimum weights in a given application.

This project is reported in Blomberg & Elenius (1986).

Automatic time alignment of speech with a phonatic transcription

The problem of automatic time alignment of a speech wave to a known phonetic transcription has attracted a lot of attention during the last years. It would facilitate or replace the tedious manual labeling and would be a way to make it more consistent. The development of speaker-independent large-vocabulary speech recognition systems requires very large amounts of speech data to get quantitative and qualitative measures of the influences of, e.g., coarticulation, reduction, and stress patterns on the acoustic speech signal. Several hours of speech will be necessary to cover a sufficient amount of phonetic variation to get reliable statistics of the speech data. The alignment procedure itself is an essential part of the verification component in several phonetic speech recognition systems. It serves the same function as the non-linear time warping algorithm in standard pattern matching word recognition systems.

A block diagram of the system used in this report is shown in Fig. 2. The main principle is a continuation of earlier work on word recognition systems (Blomberg & Elenius, 1978). The main features of the

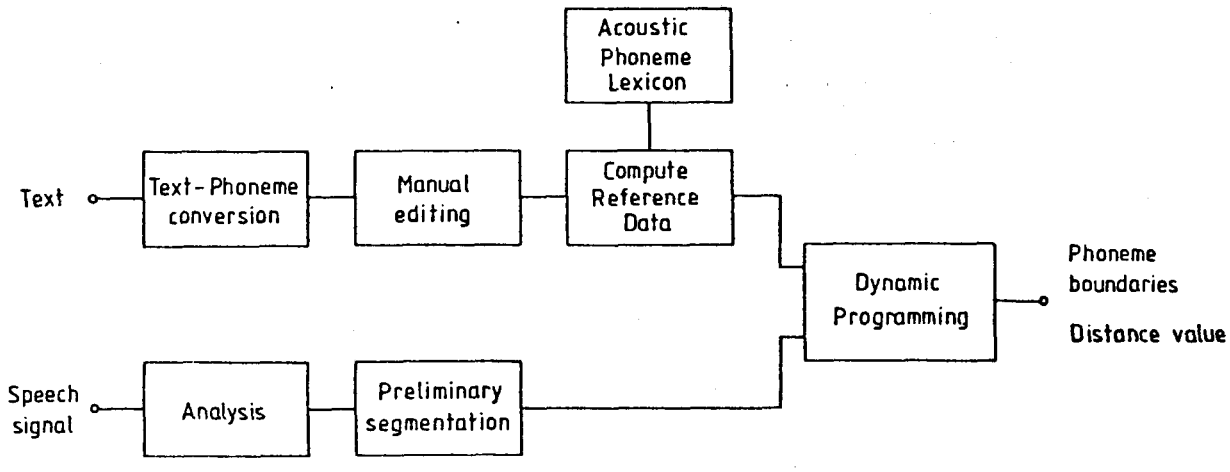

Fig. 2. Block diagram of the phonetic alignment system.

alignment process are described as follows. Phonetic transcriptions are generated from text using a module of the text-to-speech synthesis system developed by Carlson & Granström (1982). Only one transcription is generated for each sentence. It must be modified by hand to fit the actual pronunciation. General acoustic descriptions of the phonemes are retrieved from a lexicon.

Mainly for processing speed reasons, a step of preliminary segmentation is performed. Preliminary segments are put where changes in either or both of the intensity bands are detected. The sensitivity of this process must be high enough not to lose any correct boundary, since the phoneme boundaries are chosen from these preliminary boundaries. No classification is done of the preliminary segments, since they only serve for selecting boundary candidates among the time samples.

The alignment is performed by a dynamic programming algorithm at the parametric level. This means that the acoustic distances may be measured more accurately compared to doing a string comparison at the phonetic level. Time constraints on the time warp are given by durational limitations of each phoneme. A beam search algorithm prunes paths with an accumulated distance sufficiently higher than that of the best path. The processing speed is about equal to real time on a 16 bit minicomputer.

A small experiment consisted of 30 sentences spoken by one male speaker. The average sentence length was eight words. The rule-based transcription was correct for 97% of the segments. The boundaries were judged to be correct within 10 ms, the sampling interval, for 87% of the segments.

The alignment results are surprisingly accurate considering the few parameters used and the very broad acoustic information they contain. An advantage of this analysis is its robustness and speaker-independency, but it has to be combined with parameters more sensitive to certain phonetic contexts. Coarticulation rules have to be applied to predict the regular variation of some phonemes.

The most important improvement to be made is the inclusion of optional pronunciation rules in the text-to-phoneme's module and in the alignment algorithm. Until this is done, we still need an expert phonetician to inspect and modify the rule-generated transcriptions, which will increase the time needed for segmentation and labeling of speech data bases.

Incorporating the suggested improvements would form a module that could be used as the verification component of a continuous-speech recognition system. We will continue work in this area for both alignment and recognition purposes. The automatic alignment system is currently used in our speech data base project. (Blomberg & Elenius, 1985; Carlson & Granström, 1985)

Lexical structure of five European languages

Corpora of approximately 10,000 words have been examined in five languages: Swedish, English, German, Italian, and French. A 2-class and a 6-class "cohort" classification have been defined, and calculations made of the number of cohorts, the number of unique cohorts and their maximum and expected sizes. The discriminatory ability of stress is also considered.

The corpora represent a variety of West European languages: both Germanic (German, English, and Swedish) and Romance (French and Italian) with large individual differences. Swedish is a language with tonal word distinctions. German has a very rich inflectional structure and a strong tendency for compounding compared to English, for example. Italian has a simple relation between orthography and pronunciation but relatively free stress, whereas French has fixed stress but a more complicated orthographic system with many letter sequences lacking direct correspondence in pronunciation. This suggests that the optimal strategies for handling large vocabularies in speech recognition systems, for example, would be different.

The mean word length of the languages differs by one or two letters, and by two or three phonemes. French averages 1.5 letters per phoneme while the other languages average slightly more than one letter per phoneme. The French consonants which are pronounced only in liaison are not included in these data.

	Letter	Phoneme	Lett./Phon.	Vowels	C/V Ratio
English	7.09	5.96	1.19	2.47	1.41
Italian	7.39	6.94	1.06	3.27	1.12
Swedish	7.43	6.94	1.07	2.69	1.58
French	7.62	5.20	1.47	2.21	1.35
German	8.69	7.78	1.12	2.87	1.71

Table II. Mean letter and phonetic length of words, mean number of vowels per word and consonant/vowel ratio.

Mean data for the number of vowels per word, approximately equal to number of syllables, is also given in Table II. The ratio of consonants to vowels in a word is lowest for the two Romance languages and highest for German words. Italian words contain the most vowels per word, and German, the most consonants.

The corpora have been partitioned by two major classifications. The first results from mapping the phonemes of a word into a string in which all consonants are replaced by C, and all vowels by V. A variation includes stress, mapping vowels into 'V (stressed) or V (unstressed) vowel. The second classification results from mapping the words into strings of 6-valued or 7-valued elements. Consonants fall

into the five categories nasal, other sonorant, stops, strong fricatives, and weak fricatives. Vowels are categorized as above. This classification was earlier used by the speech group at Massachusetts Institute of Technology, Cambridge, MA, USA. Partitioning of the corpora into cohorts in this manner provides us with an approximation of the benefit that would be derived in a speech recognition system if all phonemes could be classified at least this well. Examination of the words in a particular cohort shows us which words we would have to make a decision among after this coarse classification.

Lang.	Stress	Words	Cohorts	Unique	Max size	Exp.size
Swedish	+	9679	6100	4871	57	5.1
	-	9679	5411	4127	57	6.5
English	+	9526	5305	4187	121	9.6
	-	9526	4725	3544	126	11.1
German	+	9219	6283	5228	29	3.7
	-	9219	5978	4825	33	4.0
Italian	+	8857	3863	2321	91	9.1
	-	8857	3400	1937	100	11.3
French	+	11388	2878	1544	170	25.1
	-	11388	2871	1540	170	25.2

Table III. 6- and 7-class cohort statistics.

One general conclusion is that the lexical space is extremely unevenly exploited. Considering a model language with a fixed word length of six phonemes and no further constraints yields (in a six-category classification) over 45,000 unique patterns (six raised to the sixth power). In the natural languages with this classification, we observe between 1540 and 4825 unique patterns, and an expected cohort size of from 4 to 25. This indicates a tendency towards more standardized patterns in natural language, a clustering within the lexical space.

Knowledge of the lexical structure of large samples of natural languages is interesting for models of lexical access. The distribution of lexical items within the lexical space, and their similarity to one another are meaningful considerations in constructing test corpora and explaining the process by which words are retrieved. Such knowledge also has immediate application in large-vocabulary speech recognition systems since it is useful to know how many actual words can be expected to correspond to a recognized pattern and how detailed the decision process must be.

A more extensive and detailed report on this study may be found in Carlson, Elenius, Granström, & Hunnicutt (1985).

References

The list contains references to publications of our group only. The referenced papers contain a fuller description of the work briefly described here and give also references to related work elsewhere.

Blomberg, M. & Elenius, K. (1978): "A phonetically based isolated word recognition system", J.Acoust.Soc.Am. 64, Suppl. No. 1, p. S181(A).

Blomberg, M. & Elenius, K. (1985): "Automatic time alignment of speech with a phonetic transcription", STL-QPSR 1/1985, pp. 37-45.

Blomberg, M. & Elenius, K. (1986): "Nonlinear frequency warp for speech recognition", Conference Record, 1986 IEEE-ICASSP, Tokyo, Japan

Blomberg, M., Carlson, R., Elenius, K., & Granström, B. (1984): "Auditory models in isolated word recognition," Conference Record, 1984 IEEE-ICASSP, San Diego, USA.

Carlson, R. & Granström, B. (Eds.) (1982): The Representation of Speech in the Peripheral Auditory System, Elsevier/North Holland Biomedical Press, London.

Carlson, R. and Granström, B. (1985): "Rule-controlled data base search", STL-QPSR 4/1985, pp.29-45.

Carlson, R., Granström, B., & Hunnicutt, S. (1982): "A Multi-language text-to-speech module", Conference Record, 1982 IEEE-ICASSP, Paris, France.

Carlson, R., Granström, B., & Hunnicutt, S. (1985): "A parallel speech analyzing system", STL-QPSR 1/1985, pp. 47-62.

Carlson, R., Elenius, K., Granström, B., & Hunnicutt, S. (1985): "Phonetic and orthographic properties of the basic vocabulary of five European languages", STL-QPSR 1/1985, pp. 63-94.

Carlson, R., Elenius, K., Granström, B., & Hunnicutt, S. (1986): "Phonetic properties of the basic vocabulary of five European Languages: Implications for speech recognition", Conference Record, 1986 IEEE-ICASSP, Tokyo, Japan