MACHINE AIDED ANALYSIS AND RECOGNITION OF SPEECH

D.C. Levinson
University of Calgary

Abstract

As one approach to automatic speech recognition currently under investigation at the University of Calgary, we are attempting to describe speech in terms of those acoustic features which experiments identify as good discriminants, as an alternative to basing a lexicon upon the segmental phoneme.

A system set up for conducting these experiments contains components for simplified display of the speech spectrum, interactive definition and evaluation of features, and resynthesis of utterances. With this system, a potential feature is defined at run-time as a function of acoustic parameters and other features. Utterances of a word are then displayed and individually normalized; points in the spectrum are chosen by inspection, and the values of the feature at these points are tabulated.

This cooperative process permits us to separate the problems of word identification and of time and intensity normalization, until automatic procedures for handling their interaction can be developed.

ANALYSE ET RECONNAISSANCE DE LA PAROLE À L'AIDE DE L'ORDINATEUR

Résumé

Dans le cadre des travaux menés actuellement à l'université de Calgary sur la reconnaissance automatique de la parole, on tente de trouver un moyen de décrire la parole en termes des dominantes acoustiques identifiées expérimentalement comme de bons discriminants, plutôt que de baser un lexique sur le phonème segmentaire.

Le système mis au point pour la réalisation de ces expériences comporte des éléments permettant un affichage simplifié du spectre de la parole, une définition et une évaluation interactive des dominantes, ainsi qu'une resynthèse des prononciations. Dans ce système, une dominante potentielle est définie pendant l'exécution comme une fonction de paramètres acoustiques et d'autres dominantes. Les prononciations d'un mot sont alors affichées et normalisées individuellement; des points du spectre sont choisis par inspection et les valeurs de la dominante à ces points sont mises en tableaux.

Ce procédé de coopération nous permet de faire la distinction entre les problèmes de l'identification des mots et ceux de la normalisation du temps et de l'intensité, jusqu'à ce que l'on ait mis au point des méthodes automatiques capables de traiter leur interaction.

MACHINE AIDED ANALYSIS AND RECOGNITION OF SPEECH

D. C. Levinson

Department of Computer Science
University of Calgary

INTRODUCTION

This paper describes a system set up for conducting experiments in speech. The purpose of these experiments is to develop an Automatic Speech Recognizer in line with one of the several approaches to the recognition of speech currently under investigation at the University of Calgary.

In this paper the requirements of a speech experimentation system, as determined by the present approach to recognition, are first considered; the design of this system is then detailed.

DESIGN CONSIDERATIONS

The immediate aims for this approach to speech recognition are modest. The attempt is being made to develop an identifier of isolated single words, utilising only acoustic, phonetic, and phonological information, and taking no account of syntax, morphology, word probabilities, etc. However even this limited project necessitates the writing of a lexicon, and a major premise of this work abrogates the simplifying assumption that words can be adequately represented using their phonetic spellings as suggested by the theory of phonology. In fact, an even less constraining assumption is being avoided: namely that speech should first be segmented in the time domain and then recognised by labelling each of the segments individually (Hill 1975). Experiments on speech must therefore be conducted, in order to discover appropriate representations for words.

An alternative approach would be to associate with each word, through the lexicon, a word verification function of unrestricted complexity to test for the presence of that word (Walker 1972). The compromise adopted in this work is to look for "features" (i.e. functions on the acoustic input) of general applicability which will distinguish between words.

The search for these features proceeds as follows: first, utilising the most effective pattern recognizers in existence, human eyes, the characteristics of a word are explored, and an apparently reliable feature is chosen for testing; a function returning a value intended to represent this feature is next specified; finally, this function is evaluated on a number of utterances (both instances of the word and otherwise), and its utility is assessed.

System Requirements

The system for conducting the requisite experiments must therefore have three principal capabilities:

1. for the display of utterances. Interactive manipulation of display parameters and examination of the underlying data should also be permitted.
2. for interactive definition of functions. These should be expressible in an interpretive, but calculation oriented, language. No ideal model is known to the author.
3. for data management. It should be possible to command the evaluation of a feature on an individual utterance or on a whole class.

The ideal system is, therefore, a cross between a graphics display processor, a language interpreter and an operating system. Some further facilities for which provision should also be made are:

4. auditory feedback (i.e. the resynthesis of utterances).
5. interaction during the evaluation of features. It should be possible for the user to supplement the algorithms already developed by means of interactive execution.
6. extension into a speech recognizer. The features chosen will form the nucleus of a speech recognition system. Provision should be made for eventual experimentation with control structures.
7. performance of statistical analyses.

A SYSTEM FOR EXPERIMENTATION

Experimentation systems have in the past been based on LISP (Bobrow & Klatt 1969) as an interpretive language, or on hybrid LISP-FORTRAN systems (Walker 1973), or have been powerful and general speech processing systems developed with much effort (Millar 1972, Reddy 1966). The author's approach has been to aim for a special purpose system, to be developed with a minimum of effort on a dedicated, but small, minicomputer facility. The resultant system has more of the flavour of a text editor than a programming language. Its development in assembly language took a matter of man-months.

Hardware

Speech input is from files of previously collected data. This data is the output of a bank of 24 filters, producing 7-bit samples every 15 msecs. The filters cover the range 90-3300Hz at approximately 150Hz intervals, 3dB down at the crossover points.

Processing is performed on a PDP/8 system with DECtape and 8K of core memory. The display and re-synthesis parts of the system also run under TSS/8 in 4K.

All interactive I/O, including the display, is in ASCII characters via the console terminal. The most noticeable drawbacks of this system have been the slow speed at which displays may be generated (the current device is limited to 300baud), and the impossibility of pointing to a position in an utterance. The next generation of the system should use a CRT for I/O, which will make interaction both faster and more versatile.

Software

Two primary considerations have governed the implementation of this system. It had first to be sufficiently simple to be developed and built by a single person, working in assembler language, in a few months at most; but it had also to possess sufficient power to make the conduct of experiments convenient. The present design, though primitive in many respects, satisfies, in part at least, the first five of the requirements discussed earlier, and the possibility of extension has been borne in mind. The system consists structurally of three components, handling respectively the display, the re-synthesis of utterances, and the interaction with the user.

Display

The initial version of the display, shown in figure 1, was simply a character density spectrogram. This yielded a general impression of the underlying structure of the speech, but it was eventually abandoned for two reasons; it generated a great quantity of detail without being commensurately informative, and furthermore no intuitively satisfying method of adjusting the parameters to achieve amplitude normalisation could be found.

During experimentation with the density spectrograms, together with an examination of the underlying data (see figure 2), some strong consistencies between utterances began to appear. The current display grew from an attempt to systematise these as characteristic features, in the light of phonetic knowledge. An example of this display is shown in figure 3. Its main body is a frequency-time diagram on which are marked characters (1, 2, 3 or F) at positions corresponding to the frequencies of significant peaks of short-term mean energy. A measure of voicing is indicated by "V" in the lowest row, and of high frequency energy content by "H" in the top row. The full list of features used in the display also includes measures of aspiration, silence and noise bursts following silence.

The interactive component allows for manipulation of the thresholds for each of these features. At present these features are "programmed in", and the PDP/8 debugging program must be invoked to make more complex modifications. Modifications to the thresholds are generally sufficient to effect satisfactory amplitude normalisation of an utterance.

Re-synthesis

Auditory feedback has also been provided through a resonance analogue speech synthesizer (PAT). The necessary synthesis parameters of amplitude of voicing, hiss and aspiration are estimated during generation of the display. Frequency values for the three first formants are obtained by simply labelling the displayed energy peaks, from the low frequency channels up, with some attempt to exclude the voice bar from consideration.

Results obtained with this procedure have been disappointing, little more than the temporal structure of the speech being evident in the re-synthesized "speech". It seems likely that effective re-synthesis cannot be achieved using the information extracted alone; rather, it is necessary to first "recognise" the speech, utilising phonological rules and other high level constraints to restore some of the missing information.

Interaction

The third component controls interaction with the user, both for feature definition and evaluation, and for data management. At the time of writing its implementation has not been completed and no results are available.

No attempt has been made to organise the data-base using indices. Instead the data is stored as sequences of utterances in files. These can be linked to the system on command, and commands are available for moving from file to file, from utterance to utterance, and from sample to sample within an utterance.

At any moment the data stream is positioned at a particular sample within an utterance. Through a set of rotating buffers of speech samples, the user can examine the underlying data by specifying a calculation directly. A calculation consists of a sequence of primitive operations and calls to user-defined functions, expressed, for ease of implementation, in postfix form.

The primitive operations currently taken into account include only the arithmetic operations and a special operator which returns the average energy contained in a region of the spectrum specified by upper and lower frequency bounds and duration. This last operator has been found very useful for accomplishing smoothing in computing the features used for display and re-synthesis. For each of these features it was possible to find a region of the spectrum within which intensities could be averaged without any serious loss of transitional information. In addition, contrasts, such as are required to detect short bursts of noise, can be expressed succinctly.

A feature is defined by associating the text describing a calculation with a name. No attempt has yet been made to incorporate parameters, except for the entry of data from the keyboard. More importantly no local storage facility is available.

Despite these limitations, the feature definition language has much of the requisite power for the applications envisaged. Implementation of the complete set of features described for LISPER (Bobrow & Klatt 1969) would require only one major change, the inclusion of a facility for hysteresis, and the implementation of various operators for boolean and comparison functions, "MAX", "ABS", and a ternary conditional expression operator. Hysteresis, by means of which a threshold for a feature depends on its history, would in any case only be useful for features which were to be evaluated on every time sample. In this system smoothing is instead achieved by integrating over more than one speech sample wherever appropriate.

Interactive Execution

It became apparent very early in this project that in order to apply even the simplest feature function, very complex tasks, such as handling absolute time variation, had first to be performed. At some later stage automatic procedures for controlling the application of functions must, of course, be developed; initially, however, it is desirable that the capabilities of the machine be supplemented once more with those of the human. Thus the system includes facilities for interaction execution.

Interaction with the user during evaluation of a feature can take place in two ways. A function may request entry of numerical parameters from the keyboard. Alternatively, a function can give control to the user by invoking him, in a manner consistent with the invocation of any other function. The user can then perform any necessary manipulations using the full power of the interactive language; for instance he can find an appropriate position in the data for evaluation to continue and then move the data stream to this point. Finally he returns to the calling function. This facility is similar to the debugging systems sometimes provided in interactive languages, except that it is oriented towards manipulations of data rather than programs.

CONCLUSION

Although the present implementation is crude in many ways, it is felt that a special purpose system, available on a dedicated machine, offers significant advantages over the alternative of working in a general purpose interactive language under a time-sharing operating system. The appropriate form of display can be an integral part of the design, rather than achieved through what would necessarily be an awkward interface, new data will be easy to obtain and use, and additional capabilities such as for re-synthesis, can be added.

REFERENCES

BOBROW, D.G. & KLATT, D.H. (1968) A limited speech recognition system. Proc. Fall Joint Computer Conference. pp.305-318.

HILL, D.R. (1975) Avoiding segmentation in speech analysis: problems and benefits. Proc. 8th International Congress of Phonetic Sciences, Leeds, England, August 1975.

MILLAR, J.B. (1972) An interactive speech processing system using a large computer. Int. J. Man-Machine Studies 4, pp.285-317.

REDDY, D.R. (1966) An approach to computer speech recognition by direct analysis of the speech wave. Technical Report CS49, Computer Science Department, Stanford University.

WALKER, D.E. (1972) Speech understanding research. Annual Report, Stanford Research Institute Project 1526.

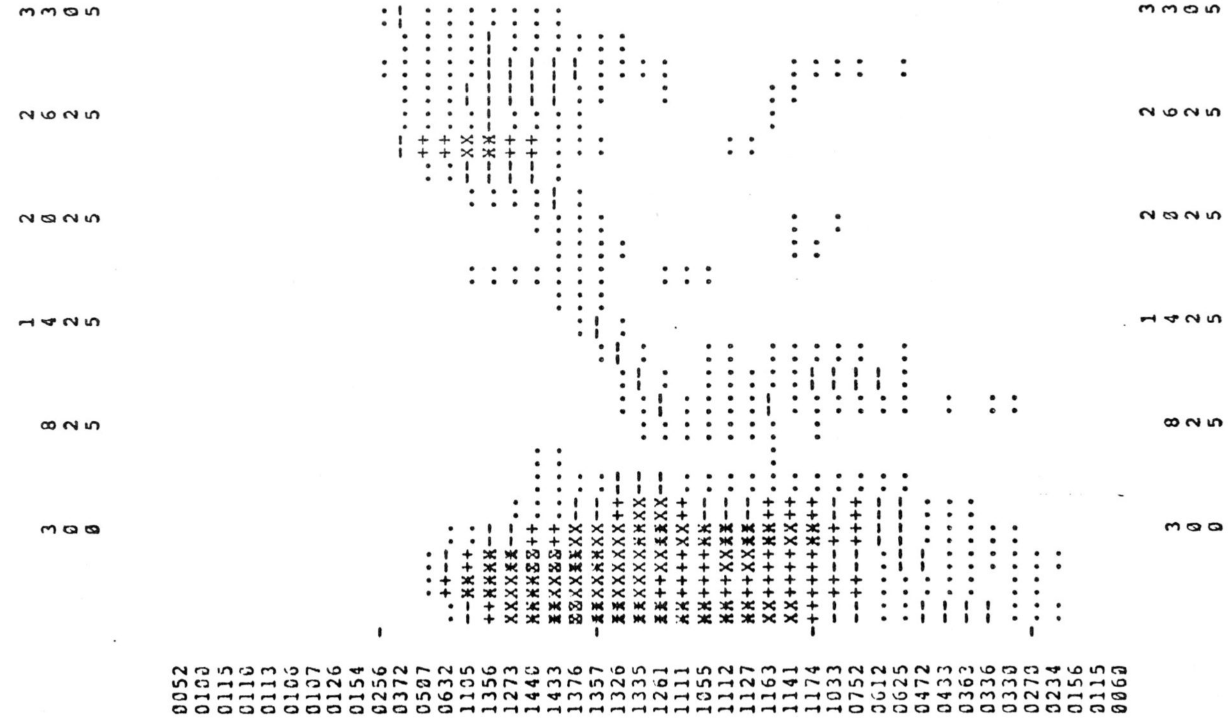

Figure 1: Character density spectrogram and total intensities for an utterance of the word "ZERO"

```
                                              1               2               2               3
              3               8               4               0               6               3
              0               2               2               2               2               0
              0               5               5               5               5               5

  4   2   2   1   2   1   1   1   1   1   3   2   1   1   1   1   2   1   2   1   2   3   3   3
  6   7   7   3   3   2   2   2   1   1   3   2   2   2   1   1   1   1   3   2   2   3   4   3
  8   8   6   2   3   2   2   2   1   1   4   3   2   2   1   2   2   2   5   3   3   5   4   4
  A   6   4   2   3   1   2   2   1   1   3   3   2   2   1   2   2   2   3   3   4   5   4   4
  9   5   3   2   3   2   2   2   2   2   3   3   2   2   2   2   2   2   3   4   5   4   4   5
  7   4   3   2   3   1   2   2   1   2   3   3   2   2   2   2   2   2   3   4   5   5   4   4
  7   3   2   2   2   1   2   2   1   1   3   2   2   2   1   2   3   3   5   5   5   6   4   5
  7   3   3   2   2   1   2   2   2   1   4   3   2   2   2   2   3   4   6   6   7   6   6   8
  8   2   2   2   3   2   2   2   2   2   4   4   2   3   3   4   4   4   5   6   9   B   A   C
  8   3   3   3   3   2   3   2   3   2   5   5   4   3   3   4   4   C   E   E   F  12   E  1B
  A   6   6   4   5   4   3   3   3   3   5   6   4   6   5   7   7   A  25  10  14  17  13  26
  A  19  11   8   6   5   4   4   4   5   7   D   8   6   5   7   9  19  32  15  17  14  10  1D
 13  38  25  10   8   7   6   5   5   5   8   A   D   A   7   9   E  19  37  13  16  13  10  13
 20  53  3F  1C   B   B   8   8   7   7   8   B   C  13   B   A  10  21  41  1A  21  1D  18  1A
 33  54  53  25   F   D   A   8   8   8   A   B   F  1D   F   E  14  27  5D  26  2E  27  21  1F
 48  4F  63  26  15   E   E   B   9   7   9   9   B  11   E   E  11  25  35  1A  23  22  1D  1E
 5D  50  74  33  19  11  11   E   D   9   B   A   B  1F   F  13  1E  24  31  1D  23  24  1C  19
 68  4B  70  33  1E  14  10   E   E   A   C   B  11  1E  12  14  24  1B  1C  1A  20  26  1F  17
 72  4D  6D  42  20  17   E   C   B   8   C  12  18  1B  13  19  19   D  11  14  1D  22  16   F
 6F  4A  5F  42  2A  1D   D   D   D   C  1A  2E  17  10  16  16   D   8  19   D  11  15  11   E
 6D  46  4F  4A  39  23   B   D  12  1A  29  1E   B   C  13   E   8   7   F   A   D  15  12   F
 69  41  44  5F  4A  26   E  13  1E  2B  19   C   9   E   C   C   A   8   C   B   F  13   D   A
 64  3C  43  60  48  20   E  14  22  16   C   C   B  11   D   C   A   7   C   D  12  12   9   8
 5F  39  3E  49  37  19   C  12  1D   F   D   B   9  14   8   8   6   6   C   E   D   A   7   7
 5A  35  3F  51  29  14   F  17  1E  13  10   D   A  10   5   6   6   6   C   9   6   6   6   5
 56  33  48  61  23  12   E  18  1A  14  12   F   9   E   8   6   7   8  13   C   9   6   6   6
 51  31  4D  63  2F  15   F  16  1E  12   F   E   A   F   9   9   8   7  10   A   8   7   6   6
 4D  36  3D  51  3C  19  12  13  20  1E  14   E   9   A   E   D   8   9   F  10  11   B   7   7
 48  3C  36  4B  34  19   C   F  18  1B  10   E   8   B  14  15   C   7   B   F  14  13   A   9
 37  38  38  5C  3B  1B   C  10  1E  22  1C   F   A   C  10   F   E   8   F   A   E  13   C   B
 2A  38  2B  33  2F  1B   A   E  18  25  1A   F   8   C   C  11   E   B   A   9   C  12   D   B
 22  30  23  32  36  1B   D   D  1B  24  10   9   9   A   C   C   9   5   8   9   C  13   D   A
 1E  1C  1F  2A  2D  14   A   C  18  24   F   7   6   7   A   9   7   4   6   7   A   D   9   6
 1E  16  23  2C  23  16   C   C  18  17  14   A   6   8   C   E   8   5   6   9   C  12   A   8
 20  15  21  1F  1C   E   A   9   E   D   9   6   4   7   D   D   7   3   5   7   A   C   7   6
 25  12  19  1A  16   D   9   8  10   B   B   7   4   7   9   A   6   3   5   7   9   A   5   5
 23  10  16  14  10   C   8   8   D   B   9   6   4   5   9   7   4   4   4   5   8   8   5   4
 20   F  13  10   D   A   7   B  10   9   8   5   4   5   8   6   4   3   5   5   7   6   4   4
 1E  10  14  11   C   8   7   C  11   8   7   5   4   5   8   5   3   2   5   5   7   6   4   3
 1A  10  17   E   9   9   7   C   B   5   5   4   3   4   6   4   3   1   4   3   4   4   4   3
 13   E  17   B   8   8   7   C   8   3   4   3   2   3   4   4   2   1   3   3   4   5   3   2
  E   A   E   8   7   6   5   6   4   2   3   2   2   3   3   3   2   1   2   2   2   4   3   2
  A   8   9   5   5   4   4   5   3   2   3   2   1   2   2   2   1   1   1   1   1   2   2   1
  6   3   3   3   3   2   3   3   2   1   3   2   1   2   1   1   1   1   1   1   1   1   2   1

                                              1               2               2               3
              3               8               4               0               6               3
              0               2               2               2               2               0
              0               5               5               5               5               5
```

Figure 2: Hexadecimal dump of the data underlying figure 1

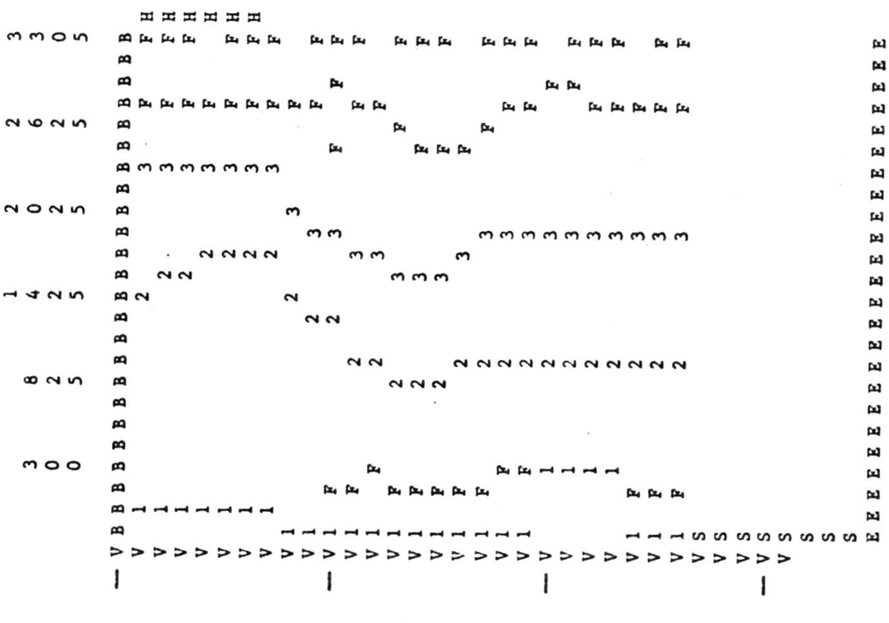

Figure 3: Current version of the display, shown for the utterance of figure 1