SPEECH PERCEPTION BASED ON LARGE-SCALE SPECTRAL INTEGRATION

Malcolm Crawford and Martin Cooke

Department of Computer Science
University of Sheffield
England

ABSTRACT - This paper presents a computational model of speech perception, within the framework of a general theory of auditory processing. We believe that large-scale spectral integration may play an important part in speech recognition, and may account for a disparate range of findings in auditory psychophysics. We present initial findings from a model of integration on an ERB scale treated as a post-streaming transformation, discuss some of the current limitations of the model, and proposals for future work.

INTRODUCTION

A major goal of speech research has been to find invariant representations of speech units. Workers in the field of Automatic Speech Recognition (ASR) have turned to auditory models in the hope that the representations they produce will be less variable than those produced by standard analysis techniques. A promising auditory process seems to be **large-scale spectral integration** (LSI), which follows from Chistovich and Lublinskaya's (1979) seminal work on the concept of a "spectral centre of gravity". They showed that if two formants are within a certain range of each other (around 3.5 Bark) they can be replaced by a single, "equivalent", formant whose frequency is approximately midway between the two, and the vowel perceived remains the same. In other words, some formants appear to be "integrated" by the auditory system to form a single auditory peak.

Other authors have proposed that LSI forms the basis for speech perception. Syrdal and Gopal (1986), for example, present a "perceptual model of vowel recognition based on the auditory representation of American English vowels". Formant frequencies, taken from Peterson and Barney's (1952, cited Syrdal and Gopal, 1986) classic data, and from LPC tracks, were transformed to a Bark scale. It was proposed that if two formants were within 3 Bark of each other they would be integrated by the auditory system. Vowel classification was based on linear discriminant analysis of the values of F1-F0, F2-F1, F3-F2, F4-F3, F4-F2, in Bark (these differences were said to correspond to binary phonetic features of American English vowels). Bladon (1986) expands on this by suggesting that LSI is a general process applied to all speech segments. He shows that LSI, and a consideration of more general auditory processing, might account for the structuring of the speech sounds used in languages. In order to propose a tenable theory of speech perception, however, we must consider LSI as a transformation within the framework of an overall theory of auditory processing. Consider the following two experimental results;

a Darwin and Gardner (1986) showed that mistuning a harmonic led to its making a reduced contribution to the percept of a vowel (it was "streamed out", cf. Bregman, 1990); and
b Darwin and Gardner (1987) showed that exciting a formant on a different fundamental also led to its being streamed out.

Two main conclusions follow from this work;

a the F1 region is resolved into harmonics; if it were not, the auditory system would not be able to detect the lack of relationship of the mistuned harmonic and so stream it out; and
b if linguistic decoding is in some way mediated by LSI, therefore, it follows that LSI should also be a post-streaming process; it only makes sense to consider those components of the signal that belong together.

For Syrdal and Gopal to call their model "perceptual", and particularly "based on the auditory representation" is misleading; simply transforming formant frequencies from Hz to Bark takes little account of the actual processing performed by the auditory system, or its true resolving power. It is now generally considered that the ERB-rate scale (hereafter referred to simply as ERB) due to Moore and Glasberg (1983)

more accurately reflects the true resolving power of the auditory periphery. Using the ERB scale the F1 region is always resolved into harmonics, as is F2 at times. The resolution of the lower frequency regions into harmonics is not taken into account by Bladon (although this may be of lesser importance to his theory), or, curiously, by Stevens (1989) who states incorrectly that the bandwidth of F1 is less than that of the auditory filters in the lower frequency region. It is also of interest to note that many systems based on auditory representations do not show the expected resolution (e.g. Seneff, 1987). Most current approaches (with the exception of Bladon, cf. Bladon 1982) seem to be expectation driven, in trying to fit auditory representations to more traditional ones. The resolution of harmonics is seen as an embarrassment to cover up, rather than a rich source of information for the auditory system. This attitude may be summed up by Klatt (1982, p 186): "... there is insufficient perceptual evidence to justify building a speech analysis system with filter bandwidths wider than a critical band, although I am sorely tempted to do so".

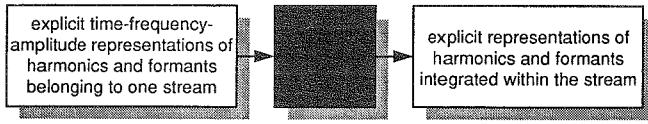
These points are addressed in our original model of integration (Crawford and Cooke, 1990, hereafter CC90), and in Crawford (1990). This paper presents initial findings from an improved implementation. As in CC90, we make three main proposals which form the basis for the current study:

a F1 estimation and higher formant integration have a common mechanism, namely, large-scale spectral integration;
b Integration is a post-streaming process; and
c Integration is a general mechanism which is applied wholesale to all streamed input (i. e. it is not restricted, say, to voiced segments, cf. Bladon, 1986).

A MODEL OF LARGE SCALE SPECTRAL INTEGRATION

Background

Following Green *et al.*'s (1990) arguments for the use of a representational approach in ASR, we consider the perception of speech as a sequence of representational transformations, using intermediate representations in the manner proposed by Marr (1982) for visual processing (an overview is given in Cooke, Crawford and Brown, 1990; cf. also Darwin, 1984, and Schwartz and Escudier, 1989). On the basis of the foregoing discussion, we can state that our goal is to model the following transformation;



The integrated representation may then serve as input to a phonemic classifier. The initial representations are currently provided in the form of **synchrony strands** by a model of the auditory periphery described in Cooke (1990, for an overview again see Cooke, Crawford and Brown, 1990; future implementations may also use information from modulation maps, cf. Brown and Cooke, 1990). Briefly, synchrony strands are explicit time-frequency-amplitude representations of synchronous auditory filter activity. Since the grouping algorithms are not yet fully implemented, the assumption is made that the strands produced by analysis of a single speaker in quiet conditions without streaming are equivalent to those that would be produced in a noisy environment after "ideal" streaming. Our original model (CC90) also made the assumption that integration at the stage of strand formation would be equivalent to integration of actual strands. The present model tests that assumption, and more closely adheres to our theory by integrating representations that could have passed though a streaming process.

Current implementation

The current model of integration is implemented as follows. The original strands representation is converted to a frame-based line spectrum. Since amplitude variation due to glottal excitation is not in phase across channels, a fact which will disrupt integration, the amplitude along each strand is smoothed with a leaky integrator (this also gives a crude estimate of temporal integration of amplitude). The height of

each line in the spectrum is determined by the smoothed amplitude of the strand during a frame (this is not directly correlated with any measure of perceived loudness - it is determined by the number of filters recruited by the spectral peak). The discrete spectrum is convolved, frame by frame, with the first derivative of a Gaussian of width N ERB (this is equivalent to smoothing and differentiation). For a range of experiments N was in the region of 2.5-3.5 (see later). The positions of positive-going zero-crossings in the convolution profile are then found. These represent the positions of peaks in the smoothed spectrum, which are then grouped to form integrated strands. An example of the results of this processing is shown in Figure 1.

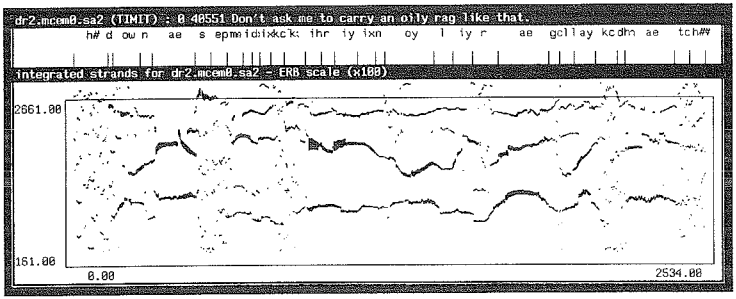


Figure 1. Integrated strands from the utterance DR2.MCEM0.SA2 (TIMIT database) on an ERB scale (50 - 4000 Hz).

INITIAL FINDINGS

The following findings were reported in CC90 using the original model;

a In the integrated representation, there are quite clear discontinuities in the integrated strands at several phoneme boundaries (as marked by the transcription) which are not made explicit in either the original strands or in spectrograms:

b the integrated strands produced for the similarly labelled segments for different talkers, male and female, are often surprisingly similar. There is also the suggestion that normalisation may be effected by subtracting the value of F0 from the integrated formant frequencies (cf. Seneff, 1987):

c from the evidence of synthesised utterances (generated using the Klatt (1980) synthesiser) there is a complicated interaction between F0 and integration. Generally, when F0 is high, formants must be closer together to be integrated than for lower F0. (This makes the choice of the integration range, N, difficult without original experimental stimuli).

It was predicted that the original model (CC90) would give a good approximation to the current one, but it was further hoped that the promises held out by the original would be realised in the current model. Our initial findings with the current model have not been consistent, for example;

a The representation of F1 is often poor; harmonics are not always integrated as might be expected, particularly in female speech; the most dominant harmonic is frequently tracked rather than F1 itself. This may be, however, a manifestation of the "harmonic efficiency criterion" (Bladon,1982). Using synthesised stimuli it was found that F1 is often very accurately tracked for F1 > about 3.5 x F0, but follows the nearest harmonic for F1 < about 3.5 x F0, concurring with Bladon's observations.

b In a similar way, instead of producing a single equivalent formant with a frequency midway between the two formants that are integrated, the more dominant one is often tracked. In synthesised stimuli where the amplitude of both formants is similar, both are represented explicitly.

c Segmental boundaries are sometimes less clear in the current implementation than in the CC90 model. There is also no greater evidence of invariant representations, although there are again some startling similarities of representation across speakers, male and female.

DISCUSSION

We must ask why the results do not live up to our (admittedly rather high) expectations. In addressing this question we will also expand upon our approach to speech perception in general.

Problems with the model

The current amplitude measure over-exaggerates the spectral peaks. Combined with the (questionable) use of a discrete spectrum, this may account for the non-integration of formants in the synthesised stimuli, and the "tracking" of the most dominant formant. This is probably the most serious fault, and the one to be most quickly remedied by relating the amplitude to a measure of perceived loudness.

Some of our assumptions may be wrong

The most interesting of these to consider is that integration may not be a post-streaming process. Consider the following experiment: Darwin (1989) has reported work by Culling in which two vowels were synthesised; each was split into two frequency regions, between the first and second formants. One fundamental was assigned to the first formant of one vowel, and to the higher formants of the other, and vice-versa. This should lead to an incorrect grouping, so that vowel identification when the two are played simultaneously should be very poor. In fact the experiments showed that identification was only slightly worse than for "correct" simultaneous vowels. This poses a challenge to the "serial" model of processing outlined above. It still does not make sense to propose that integration is not a post-streaming process; this would be in effect to propose a broad-band analysis of the signal, noise and all, and leave no explanation for the experiments outlined in the introduction. Darwin suggests that in some circumstances phonetic mechanisms can group together sounds that are otherwise treated as separate groups. What might these circumstances be?

We can suggest an answer based on the "grouping hypotheses list" outlined in Cooke and Green (1990). In their preliminary model of streaming, strands are grouped on the basis of harmonicity, and alternative groups are ranked in order of the proportion of the data that they account for. It is possible, however, if grouping algorithms based on common amplitude modulation are implemented, that substreams may be formed which would, for example, group the harmonics and formants separately. The top ranking hypotheses might then be those which combined the harmonics and formants excited by the same fundamentals. These may then be rejected, however, as they do not correspond to recognised vowels. We must recall Marr's (1982) "Principle of Least Commitment": early representations are **not** thrown away; they may be reinterpreted at a later stage by higher level processes. Hypotheses that are lower ranking might provide a better fit to expectations, and are therefore chosen in preference.

Points to note with respect to speech perception, and invariant transformations

a Beware of spectrogram envy! We are used to looking for particular features in representations of speech due to continued exposure to FFT displays. It may not be appropriate to look for similar shapes or groups of objects in auditory representations. Simply converting from Hertz to ERB frequency scale has a profound effect on the "formant" structure".

b There are many other representations which may be used by the auditory system in the process of decoding the speech signal. The current representation is not very rich; for example bandwidth could be encoded as part of a future integrated representation; onset and offset descriptions may also be utilised in decoding the speech signal.

c It is possible that there are no invariant representations *per se*. Speech units may not be encoded in absolute representations, but rather with reference to foregoing structures. This strengthens the argument for a representational approach in which temporally extensive structures are made explicit (see Green *et al.*, 1990, and below); temporal relationships are difficult to encode using a frame-based or "bacon-slicer" approach.

FUTURE WORK

The research outlined so far has been guilty of falling into the "rush to recognition" trap. A period of consolidation must now follow, in which more basic work should be carried out;

a calibration of the model, particularly the amplitude/loudness measure. An important task will be to analyse stimuli from experiments; in this way the model can be properly tested, and tuned to produce appropriate representations. In particular a review may show that experimental results are more consistently explained using an ERB rather than a Bark scale, and that there is an effect on integration range due to F0 frequency.

b We have so far been unable to resynthesise from the current representations, since the amplitude modulation information is lost in the smoothing stage. We aim to implement a resynthesis route to allow testing as outlined in CC90. We predict that if the model is functioning properly the resynthesised signal will be perceptually equivalent to the original (cf. single equivalent formant experiments).

c Once we have a "correct" implementation, we will perform an extensive study of syllables with continuous formant structures (e.g. vowel-semivowel-vowel) to test the following hypotheses;

  i some segments are marked by discontinuities (cf.also Abry, Boe and Schwartz, 1989)

  ii normalisation may be effected by F0 subtraction

d We aim to characterise the current descriptions in a similar manner to the formant characterisation detailed in Green *et al.* (1990). This would provide higher level representations such as "Peak" or "Dip", which could be used develop descriptions of relationships between parameters. This would allow the development of a recognition system founded on object-based distance metrics. In order to pursue the route toward understanding speech perception, however, we must also consider enriching the description with, for example, onset and offset markers, representations of bandwidth, and descriptions of the rate representation.

In conclusion we should point out that we are aware that much of this work is speculative. We feel, however, that our early findings have been of sufficient interest to warrant further investigation.

ACKNOWLEDGEMENTS

Thanks are due to Guy Brown and Phil Green for their support. Thanks are also due to Dave Pallet, Maxine Eskénazi, and Lori Lamel for the (continued!) use of the TIMIT database. Enfin, merci à Jean-Luc Schwartz pour une discussion qui a durée une journée entière (et un déjeuner excellent), à Maxine Eskénazi pour toute son aide (et un dîner superbe, et le merguez), à Denis Beautemps pour plusieurs enregistrements très utiles (et La Geuze), et aux trois pour leur attitude amicale.

Malcolm is supported by SERC CASE award 88501079 and by British Telecom Research Laboratories. Martin is supported by SERC award GR/E 42754

REFERENCES

Abry, C., Boe, L-J., and Schwartz, J-L. (1989) *Plateaus, catastrophes and the structuring of vowel systems, Journal of Phonetics*, **17**, pp 47-54.

Bladon, A. (1986) *Phonetics for hearers*, in: McGregor, G. (Ed.) *Language for hearers* (Pergamon Press).

Bladon, R. A. W. (1982) *Arguments against formants in the auditory representation of speech* in: Carlson, R.; Granström, B. (Ed.s) *The representation of speech in the peripheral auditory system* (Elsevier Biomedical Press).

Bregman, A. S. (1990) *Auditory scene analysis: the perceptual organization of sound* (MIT Press, Cambridge, Mass.).

Brown, G. J., Cooke, M. P. (1990) *Extracting descriptions of amplitude modulation from an auditory model: a comparative study*, Institute of Acoustics Autumn Conference, Speech and Hearing, Windermere, 22/25 November.

Chistovich, L. A., Lublinskaya, V. V. (1979) *The 'centre of gravity' effect in vowel spectra and critical distance between the formants: psychoacoustical study of the perception of vowel-like stimuli*, Hearing Research, **1** pp 185-195.

Cooke, M. P. (1990) *Synchrony strands: and early auditory time-frequency representation*, University of Sheffield Departmental Research Report, March 20, 1990.

Cooke, M. P., Crawford, M. D., Brown, G. J. (1990) *An integrated treatment of auditory knowledge in a model of speech analysis*, Third International Conference on Speech Science and Technology, SST-90 Melbourne, November 27-29,1990.

Cooke, M. P., Green, P. D. (1990) *The auditory speech sketch*, Institute of Acoustics Autumn Conference, Speech and Hearing, Windermere, 22/25 November.

Crawford, M. D. (1990) *Knowledge-based approaches to speech recognition*, in: Linkens, D. A., Nicolson, R. I. (Ed.s) *Trends in Information Technology* (Peter Peregrinus).

Crawford, M. D., Cooke, M. P. *A computational study of large scale integration*, Institute of Acoustics Autumn Conference, Speech and Hearing, Windermere, 22/25 November.

Darwin C. J. (1989) *Speech perception seen through the ear* European Conference on Speech Communication and Technology (Eurospeech), Volume 1, pp 230-234.

Darwin, C. J. (1984) *Perceiving vowels in the presence of another sound: Constraints on formant perception*, J. Acoust. Soc. America, **76** pp 1636-1647.

Darwin, C. J.; Gardner, Roy B. (1987) *Perceptual separation of speech from concurrent sounds*, in: Schouten, M. E. H. (Ed.) *The psychophysics of speech perception* (Martinus Nijhoff).

Darwin, C. J., Gardner, R. B. (1986) *Mistuning a harmonic of a vowel: grouping and phase effects on vowel quality*, J. Acoust. Soc. America, **79** (3) pp 838-845.

Green, P. D. *et al.* (1990) *Bridging the gap between signals and symbols in speech recognition*, in: W. A. Ainsworth (Ed.), *Advances in Speech, Hearing, and Language Processing*, Volume 1 (JAI Press Ltd. London).

Klatt, D. H. (1982), *Speech processing strategies based on auditory models*, in: Carlson, R., Granström, B. (Ed.s) *The representation of speech in the peripheral auditory system* (Elsevier Biomedical Press).

Klatt, D. H. (1980) *Software for a cascade/parallel synthesiser*, J. Acoust. Soc. America, **67** (3), pp 971-995.

Marr, D. (1982) *Vision* (Freeman Press, San Francisco).

Moore, B. C. J., Glasberg, B. R. (1983) *Suggested formulae for calculating auditory-filter bandwidths and excitation patterns*, J. Acoust. Soc. America, **59** pp 750-753.

Schwartz, Jean-Luc, Escudier, Pierre (1989) *A strong evidence for the existence of a large-scale integrated representation in vowel perception*, Speech Communication, **8** pp 235-259.

Seneff, S. (1987) *Vowel recognition based on 'line formants' derived from an auditory-based spectral representation*, Proc. 11th Int. Congress Phonetic Sciences, Tallinn, USSR, Paper **Se 95.1**

Stevens, K. N. (1989) *On the quantal nature of speech*, Journal of Phonetics, **17** pp 3-45.

Syrdal, Ann K., Gopal, H. S. (1986) A perceptual model of vowel recognition based on the auditory representation of American English vowels, J. Acoust. Soc. America, 79 (4) pp 1086-1100.