Proceedings of the Institute of Acoustics

A COMPUTATIONAL STUDY OF LARGE-SCALE INTEGRATION

Malcolm Crawford & Martin Cooke

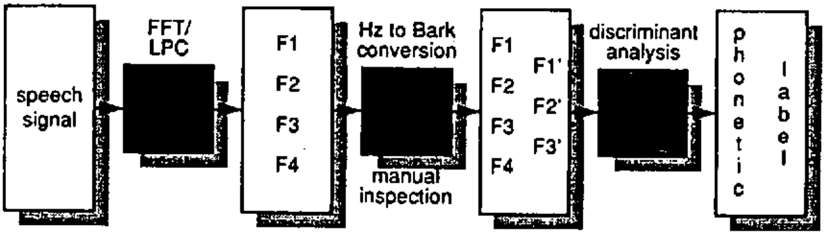
University of Sheffield, Department of Computer Science, Sheffield, England.

1. INTRODUCTION

This paper presents, in a deliberately provocative manner, a preliminary model of large scale spectral integration based on a model of the auditory periphery, and some ideas that follow from early findings.

The seminal work of Chistovich and Lublinskaya, [1] into the concept of a "spectral centre of gravity", has led several authors (most notably Bladon, [2]) to propose that 3 - 3.5 Bark integration may form the basis for much of speech recognition. In this paper we are primarily interested in what representations might be available to the auditory system and how it would compute them, rather than with the encoding of linguistic units. The results of informal experimentation using a model of large scale spectral integration, however, may have some relevance for understanding speech perception.

From casual discussions, the way many researchers seem to understand spectral integration is along the lines of the model of vowel perception proposed by Syrdal [3] and Syrdal and Gopal [4]. The model is summarised in the diagram below:



In their "perceptual model of vowel recognition based on the auditory representation of American English vowels" the formant frequencies are taken from Peterson and Barney (1952, cited Syrdal [3]), and in a second experiment (Syrdal [3]) extracted by a formant tracker based on Linear Predictive Coding of the signal (visually verified on a spectrographic display). It is assumed on the basis of evidence from a spectrogram that if two formants are within a critical distance of each other (3 Bark) they will be integrated by the auditory system. Linguistic classification is based on linear discriminant analysis of the values of F1-F0, F2-F1, F3-F2, F4-F3, F4-F2, in Bark, where these differences are said to correspond to binary phonetic features of American vowels.

To call this a perceptual model, however, is misleading. Simply transforming formant frequencies from Hz to Bark does not take account of the actual processing performed by the auditory system, or its true resolving power. For example, just because two formants, whose frequencies have been determined from an FFT, are within 3 Bark of each other may not necessarily mean that they are integrated; it is not even the case, using a more accurate auditory frequency scale than Bark, that lower formants are represented as single peaks.

It is now generally considered that the ERB-rate scale (hereafter referred to simply as ERB) due to Moore and Glasberg [5] more accurately reflects the true resolving power of the auditory system. On a Bark scale

F1 may be resolved into harmonics for utterances with high fundamental frequencies. Using the ERB scale the F1 region is always resolved into harmonics, as is sometimes F2. The problem of estimating the frequency of F1 from a series of harmonics must therefore also be addressed.

Algorithms proposed for F1 estimation include; calculation from the spectral envelope, a weighted measure of either the single or the two most prominent harmonics, and a weighted measure with exceptions for high F0 or F1 near a harmonic (Assman and Nearey [6], Carlson, Fant and Granström [7], Darwin and Gardner [8]; Javkin et al. [9]). These approaches all seem rather mechanistic and top down; a more appealing solution is that there is a single, simple, explanation for F1 estimation which has so far been overlooked.

Experimental evidence for the resolution of the lower frequency region into harmonics comes from work such as that of Darwin and Gardner [10]. They showed that a mistuned harmonic makes a reduced contribution to the quality of a vowel, it is, in effect, "streamed out". This would not be possible if the harmonic were not resolved so that its lack of relationship to other harmonics could be detected. This result suggests that the determination of F1 frequency is a **post-streaming** process.

In a second series of experiments, Darwin and Gardner [11] showed that exciting a formant by a different fundamental frequency also causes it to be streamed out, thereby altering the perception of the stimulus. This suggests that, if phonemic categorisation is mediated in some way by spectral integration, it too should be a post-streaming process.

Rather than postulate a separate mechanism for the integration of formants and harmonics, therefore, the goal of the current research is to model integration as a single process, whilst also taking into consideration other auditory constraints such as streaming. Furthermore, it does not seem to make sense to have integration as being vowel specific. Following Bladon [2], therefore, it is also suggested that integration is a general auditory mechanism, which might be seen as a data reduction process, applied wholesale to all input.

In summary, the three main proposals which form the basis for this study are:

a F1 estimation and higher formant integration have a common mechanism, namely, large-scale spectral integration;

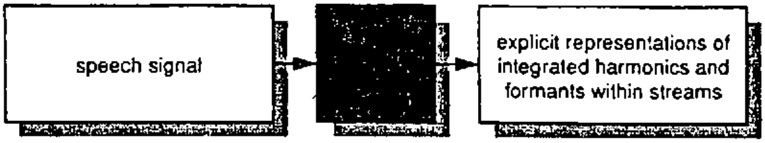
b Integration is a post-streaming process; and

c Integration is a general mechanism which is applied wholesale to all streamed input (i. e. it is not restricted to single spectral frames or to vowel classification).

2. BACKGROUND TO THE CURRENT RESEARCH

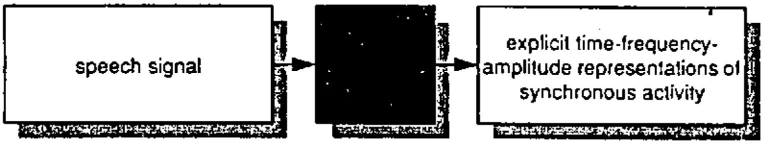
Following Green et al.'s [12] arguments for the use of a representational approach in Automatic Speech Recognition (ASR), the perception of speech is considered to proceed via a sequence of representational transformations, using intermediate representations in the manner proposed by Marr [13] for visual processing (cf. also Darwin [14], and Schwartz and Escudier [15]). Suggestions of those intermediate representations that might be used in speech perception, what transformations are made, and what constraints might apply at each level, are given in Bregman [16], and a computational approach presented by Cooke and Green [17]. Crawford [18] and Cooke, Crawford and Brown [19] outline suggestions for the applications of auditory processing to ASR.

In the broadest terms, the model of integration should effect the transformation shown below. The integrated representations should then form the input for phonemic classification. It should be clear that, due to the constraints of streaming, intermediate representations are required. These are currently provided, in the form of explicit time-frequency-amplitude representations known as **synchrony strands**, by the auditory model developed by Cooke [20]. A brief description of this model follows.


speech signal
explicit representations of integrated harmonics and formants within streams

2.1 Cooke's (1990) auditory model

Cooke's model currently performs the following transformation;


speech signal
explicit time-frequency-amplitude representations of synchronous activity

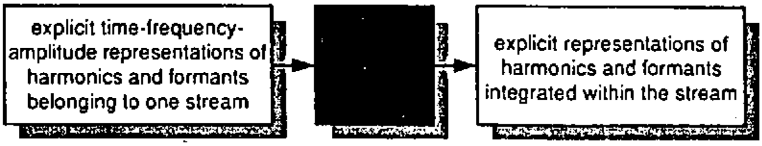
The initial analysis is made by a bank of bandpass gammatone filters, equally spaced along an ERB scale, whose characteristics closely match findings from physiological and psychophysical experimentation. The frequency of the most prominent component in the output of each filter is calculated on a frame-by-frame basis, by median-filtering the instantaneous frequency of its output.

The responses of filters tend to be determined by the most dominant local frequency component of the stimulus. Each frame of instantaneous frequency estimates, therefore, generally contains a high degree of redundancy caused by large numbers of filters responding to the same spectral peak. The third stage of processing provides a summary of the synchronous activity within a time frame by grouping channels with similar characteristics into **place-groups**. Finally, place-groups are aggregated over time to produce descriptions of auditory synchrony in the form of **synchrony strands**. These are explicit time-frequency representations of synchronous filter activity; amplitude is also encoded as part of the description.

This representation will, with future implementation of grouping algorithms such as those proposed and outlined in Cooke and Green [17], enable streaming to be modelled. An example of the output of the current model is shown in Figure 1 (top).

3. THE MODEL

The input to the model of integration will be all those representations that are deemed to have originated from the same source:


explicit time-frequency-amplitude representations of harmonics and formants belonging to one stream
explicit representations of harmonics and formants integrated within the stream

The current model of integration makes the assumption 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. The following important assumption is also made: that the effect of smoothing at the stage of the formation of synchrony strands is equivalent to integration of discrete spectrum synchrony strands.

The first task was to convert the 3 Bark integration range to ERB, since this is the frequency scale used by the auditory model. Since the relationship between ERB and Bark scales is non linear, however, a constant range of integration in ERB will not be constant when transformed to Bark. For example; at 2 kHz a range of +/- 3 Bark corresponds to about +/- 3.7 ERB; at 400 Hz, however, a range of 3.7 ERB corresponds to +2.5 to -2.2 Bark. A value of 3.8 ERB was used as this corresponds to 3 Bark for the mid-range of formant frequencies. The model was then "implemented" by altering the width of the Gaussian used to convolve the place groups in the production of "ordinary" synchrony strands to be equivalent to the chosen range of integration.

4. INITIAL EXPLORATIONS

4.1 Representation-based experiments.

A series of analyses of utterances chosen at random from three databases were made. The utterances contained all phonetic units, not just vowels. An example of one of the representations produced is shown in Figure 1, together with the original (non-integrated) strands representation.

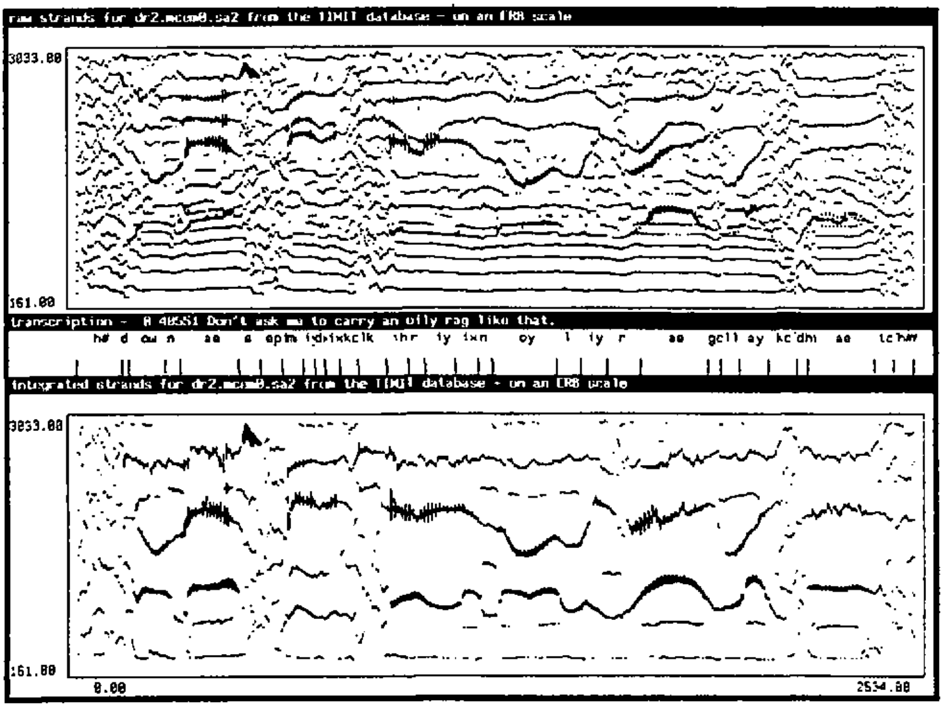


Figure 1: Non-integrated (top) and integrated (bottom) strands produced by analysis of utterance dr2.mcem0.sa2 from the TIMIT database, with transcription (centre).

The most interesting observations were that;

a In the integrated representation, there are quite clear discontinuities at several phoneme boundaries (as marked by the transcription) which do not show up on other representations; and

b the integrated strands produced for the similarly labelled segments for different talkers, male and female, are often surprisingly similar. They also appear to be "normalisable" by subtracting the value of F0 from the integrated formant frequencies (cf. Seneff [21])

c in synthesised utterances (generated using the Klatt [22] synthesiser), changing F0 has a non-linear effect on the integration of formants. For example, increasing F0 whilst keeping the formant frequencies of a stimulus the same may result in segregation of formants that were previously integrated.

4.2 Resynthesis experiments.

Concurrent with the above, a series of resynthesis experiments was performed. Ideally, if a representation still contains enough information, resynthesis should be intelligible. If this is not the case, important information may have been lost. This may, of course, not be the case for representations of increasing levels of abstraction. At the highest level of representation, synthesis from a symbol representing a phonemic category is likely to be impossible. Clearly, though, integrated strands are a lower level abstraction than phonemic categories, and, given their aim of modelling perceptual equivalence, might be expected to allow some resynthesis.

Synthesis from strands is relatively straightforward. Each strand is synthesised individually as a frequency and amplitude modulated sine-wave, and the output signal formed by summation of all the strands comprising an utterance (cf. Cooke [20]). Resynthesis from non-integrated strands is generally highly intelligible, and retains most of the speaker and prosodic characteristics in the utterance. A number of the utterances resynthesised from integrated strands were very clear, in particular those from female speakers (although they often suffered from background noise, which may havebeen due to onset transients). This included utterances containing fricatives and stops. More rigorous testing is required to determine to what extent the intelligibility is due to "top-down" processing. The main exception to the foregoing was in the resynthesis of synthesised utterances. The Klatt-generated test utterances sounded particularly poor, and were hardly recognisable. One effect of particular note in cases where F0 was changed was that they were subject to quite distinct segmental changes at points where formants ceased to be or became integrated.

5. DISCUSSION

The above observations may have a number of repercussions for current theories of speech perception, and production. Note that the propositions outlined below are not dependent on the actual accuracy and validity of the model itself; they constitute a (level 1; Marr [13]) computational theory, which is independent of the (level 2) algorithm used in the model.

It is proposed that integration is a post-streaming process.

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.

Wholescale integration results in interesting and unexpected non-linear effects in the representation of spectral dominances. Sections of speech where formants are, in physical terms, moving considerably, may be represented by relatively static strands. Some segments which may be difficult to segment in acoustic representations suddenly posses almost "categorical" distinctions.

Although there is a certain "noisiness", the outputs do suggest that segmentation of some speech sounds is a fairly trivial matter for the auditory system, and that there is an often remarkable intra- and inter-speaker similarity, in terms of temporal and frequency organisation between the representations for the same linguistic units. Whilst Stevens' [23] arguments for the Quantal Nature of Speech are concerned primarily with the production of speech, he does, suggest that there may be discontinuities in the auditory representation of speech as a result of wide-scale integration (cf. also Abry, Boe and Schwartz [24]).

5.1 An appealing theory of speech perception...

It is proposed that speech perception proceeds by a series of representational transformations similar to those outlined below:

a Speech is analysed by the peripheral auditory system to form explicit representations of the components of the auditory scene.

b These representations are assigned to streams on the basis of auditory grouping and segregation.

c The "raw" representations within a single stream determine speaker quality and characteristics.

d Within a speech stream wholesale spectral integration over a range of around 3.8 ERB, makes explicit the major auditory spectral dominances - "auditory formants" (to adopt Karjalainen's [25] terminology).

e Preliminary observations of similarly labelled phonemes suggests that speaker normalisation might be effected by "subtracting" F0 from the formants in voiced sections of speech. The concentrations of energy appear to be fairly consistent in frequency, across speakers. Fricative energy concentrations appear to remain at a constant frequency.

f Phonemic "labelling" may then proceed by reference to mental models of categories described in terms of the time and frequency relations between the integrated representations.

Further investigation and exploitation of this theory seems to hold some promise for increasing our understanding of human speech perception, and for automatic speech recognition. Part of its beauty lies in its simplicity. It is fair to ask, however, why it is that the representations of linguistic categories are not entirely uniform. There are several possibilities.

a The theory may be entirely wrong...

... in which case it will not be around for long, as it is eminently testable. It should serve its main purpose, however, in stimulating thought and discussion.

b The algorithm may be wrong.

Amongst other things, the assumption that convolution of place-groups is the same as the effect of convolving a discrete spectrum produced from synchrony strands may be false. The model should also be calibrated using data from perceptual equivalence tests. It should be the case that both the test stimulus and the subject-set signals produce essentially the same integrated representation.

There is no point c.

d The "speech" may be wrong

It is possible that there are genuine errors in production, that are quickly corrected by feedback. Under normal circumstances the auditory system may "correct" the interpretation, using a similar mechanism to that which results in the "phonemic restoration effect" (cf. Warren [26]). This would account for some of the difficulties with resynthesis. When the signal is resynthesised from such a reduced representation, the initial raw strands representation is likely to be as disjointed as the integrated which may prevent more "normal" temporal integration and smoothing.

5.2 ... and production

It is tempting to think that the "targets" of speech production are defined in auditory terms. This reverses the motor theory of speech perception, by proposing a sort of "production by analysis". It is proposed that the goals of production are defined by auditory constraints. The aim of articulation would therefore be to produce in the speaker the auditory representations appropriate for the phonemic category intended. It is clear from this hypothesis that, for some phonemic categories in particular, whose perception is dependent on formant integration at a crucial point, feedback is required for continued accurate production. It would be of great interest to test this hypothesis by examining the productions of deaf people, and of subjects in experiments where the feedback route is interrupted.

This obviously has repercussions for strategies for artificial speech synthesis; we might propose a method of synthesis by analysis. Formant values could be adjusted to aim, as in the human model, for targets defined in auditory terms, by reference to the representations produced by an ongoing analysis by an auditory model.

5.3 Summary

This paper has presented preliminary observations from model of large-scale spectral integration, and, more importantly the rationale behind it. Whilst it will be clear to the reader that a great deal more rigorous testing is required, these early results are presented in order to stimulate thought in a new direction.

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 (continuedl) use of the TIMIT database.

Malcolm is supported by SERC CASE award 88501079 and by British Telecom Research Laboratories.

Martin is supported by SERC award GR/E 42754

REFERENCES

[1] L A CHISTOVICH & V V LUBLINSKAYA "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 (1979)

[2] A BLADON "Phonetics for hearers" in: G McGREGOR (ed) "Language for Hearers", *Pergamon Press* (1986)

[3] A K SYRDAL "Aspects of a model of the auditory representation of American English vowels", *Speech Communication*, 4, pp 121-135 (1985)

[4] A K SYRDAL & H S GOPAL "A perceptual model of vowel recognition based on the auditory representation of American English vowels", *J. Acoust. Soc. Am.*, 79 (4), pp 1086-1100 (1986)

[5] B C J MOORE & B R GLASBERG "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns", *J. Acoust. Soc. Am.*, 59, pp 750-753 (1983)

[6] P F ASSMAN, T M NEAREY "Perception of fromt vowels: The role of harmonics in the first formant region", *J. Acoust. Soc. Am.*, 81 (2) pp 520-534 (1987)

[7] R CARLSON, G FANT & B GRANSTROM "Two formant models, pitch and vowel perception" in: G FANT & M A A TATHAM (eds) "Auditory analysis and perception of speech", *Academic Press, London* (1975)

[8] C J DARWIN & R B GARDNER "Which harmonics contribute to the estimation of first formant frequency?", *Speech Communication*, 4, pp 231-235 (1985)

[9] H R JAVKIN, H HYNEK & W HISASHI "Interaction between formant and harmonic peaks in vowel perception", *Proc. 11th Int. Congress Phonetic Sciences, Tallinn, USSR*, Paper Se 82.1 (1987)

[10] C J DARWIN & R B GARDNER "Mistuning a harmonic of a vowel: grouping and phase effects on vowel quality", *J. Acoust. Soc. Am.*, 79 (3), pp 838-845 (1986)

[11] C J DARWIN & R B GARDNER "Perceptual separation of speech from concurrent sounds" in: M E H SCHOUTEN (ed) "The psychophysics of speech perception", *Martinus Nijhoff* (1987)

[12] P D GREEN et al."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* (1990)

[13] D MARR "Vision", *W. H. Freeman* (1982)

[14] C J DARWIN "Perceiving vowels in the presence of another sound: constraints on formant perception", *J. Acoust. Soc. Am.*, 76 (6), pp 1636-1647 (1984)

[15] J-L SCHWARTZ & P ESCUDIER "A strong evidence for the existence of a large-scale integrated spectral representation in vowel perception", *Speech Communication*, 8, pp 235-259 (1989)

[16] A S BREGMAN " "Auditory scene analysis: the perceptual organization of sound", *MIT Press, Cambridge, Mass,* (1990)

[17] M P COOKE & P D GREEN "The auditory speech sketch", *Institute of Acoustics Autumn Conference, Speech and Hearing, Windermere, 22/25 November* (1990)

[18] M D CRAWFORD "Knowledge-based approaches to speech recognition", in: D A LINKENS & R I NICOLSON (eds) "Trends in Information Technology", *Peter Peregrinus* (1990)

[19] M P COOKE, M D CRAWFORD & G J BROWN "An integrated treatment of auditory knowledge in a model of speech analysis", *Third Intenational Conference on Speech Science and Technology, SST-90 Melbourne, November 27-29* (1990)

[20] M P COOKE "Synchrony strands: an early auditory time-frequency representation", *University of Sheffield Departmental Research Report, March 20* (1990)

[21] S SENEFF "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** (1987)

[22] D K KLATT "Software for a cascade/parallel formant synthesiser", *J. Acoust. Soc. Am.*, **67** (3), pp 971-995 (1980)

[23] K N STEVENS "On the quantal nature of speech", *Journal of Phonetics*, **17**, pp 3-45 (1989)

[24] C ABRY, L-J BOE & J-L SCHWARTZ "Plateaus, catastrophes and the structuring of vowel systems", *Journal of Phonetics*, **17**, pp 47-54 (1989)

[25] M KARJALAINEN "Auditory models for speech processing" *Proc. 11th Int. Congress Phonetic Sciences, Tallinn, USSR*, Paper Pl **2.1.1** (1987)

[26] R M WARREN "Perceptual restoration of missing speech sounds", *Science*, **167**, pp 392-393 (1970)