RESYNTHESIS AND MATCHING EXPERIMENTS ON AN AUDITORY THEORY OF MALE/FEMALE NORMALISATION

R.I. DAMPER*, R.A.W. BLADON**, R.W. HUKIN* and G.N.A. IRVINE*

* Department of Electronics and Computer Science
University of Southampton
United Kingdom

** Phonetics Laboratory
University of Oxford
United Kingdom

ABSTRACT

Variability between speakers, particularly those of different sexes, poses problems for speaker-independent speech recognition. Recently, it has been suggested that much of this variability could be minimised using a suitable computational model based on known or assumed details of human auditory processing. We are attempting to test this notion experimentally by resynthesising speech which has been processed by the model and studying its perceptual nature.

INTRODUCTION

Current approaches to speech recognition are characterised by the use of signal and pattern processing techniques which are "general" in the sense that little account is taken of the fact that the input (speech) has some very particular properties. As a consequence, spectral representations are typically used in which the coordinates are decibels (relative to some reference level) and logarithmic hertz-frequency, in spite of perceptual evidence that the human auditory system uses a loudness-density versus tonality representation. It is now widely held that the exploitation of knowledge about human speech processes (production and perception) is a pre-requisite for further, significant advances in speech technology, embracing recognition, synthesis and coding. Indeed, there have been several recent attempts to embody at least some of the current understanding of auditory perception into computational models ("auditory models"). The hope is that such models may prove to be more effective as pre-processors for recognition and coding than are traditional speech analysers.

One area where conventional signal processing and statistical pattern matching techniques have proved inadequate is in the handling of speaker variability such as arises from speaker sex and age differences. This sort of variability poses clear problems for speaker-independent recognition. Recently, Bladon and his coworkers [1] have suggested that many of these differences could be minimised (in vowel spectra at least) using a suitable "auditory normalisation" model. In Bladon's model, a perceptually-motivated "auditory spectrum" (obtained by transformations of the spectral coordinates and convolution with a filter intended to represent peripheral frequency analysis) undergoes linear shifts in the tonality (bark scaled) dimension. We believe that claims for the normalising potential of the model are, to some extent, testable by resynthesising speech direct from the bark-shifted auditory spectral representation.

For instance, resynthesising a one-bark-incremented version of a male vowel spectrum, but with voicing appropriate to a female speaker, should induce listeners to report no change in perceived vowel quality. On the other hand, playback of the "incremented" vowel with the male voicing retained should yield shifts in perceived quality. Indeed, it may even prove possible to effect an automatic transformation of male to female speech, or vice versa. We are attempting to substantiate these ideas experimentally and this paper reports on the early stages of the work.

The paper is structured as follows. First, previous work on auditory models and speaker normalisation is reviewed. The implementation of one particular model (essentially that due to Bladon **et al**) is then described. Subsequently, the resynthesis operation is described and a number of problems identified; the most important being that certain of the "forward" (acoustic-to-auditory) transformations effect a data-reduction and so are inherently non-invertible. Finally, some early results of listening experiments using the resynthesised speech are presented.

AUDITORY MODELS AND NORMALISATION

There is considerable variability in the acoustic realisations of the same speech sounds by different speakers [2]. Thus, the human auditory system has the ability to perceive as phonetically equivalent vowels of markedly different formant (and voicing) structure. This normalisation process implies an ability to make allowances for different vocal tract sizes and shapes. In attempting to mimic this ability in model systems, we might take either of two somewhat different approaches. One possibility is to adopt a speech production viewpoint whereby some dimensional scaling is effected according to supposed vocal tract characteristics. The alternative speech reception point of view leads us to search for an explanation of normalisation ability on the basis of known or assumed details of auditory processing i.e an "auditory model". For instance, the hypothesis of Potter & Steinburg [3] that a particular pattern of stimulation on the basilar membrane might be identified as a given sound, within limits independant of displacement along the membrane, is one possible mechanism for normalisation.

Auditory models are generally based, at least in part, on the concept of the auditory filter originally proposed by Fletcher [4]. He suggested that the peripheral auditory system behaves as if it contained a bank of filters, with a continuum of centre frequencies. The output of such a filter bank is usually termed an 'excitation pattern' since it is meant to represent the degree of activity (or excitation) evoked by a particular sound at some unspecified level of the auditory system. Schroeder suggests that the excitation pattern, E(z), could just as well be thought of as mean-squared amplitude of the basilar membrane motion at place z [5]. His model uses a rather broad auditory filter shape estimated from the somewhat dated masking experiments of Zwicker [6]. More recent evidence from experiments taking into account factors such as off-frequency listening suggests that filter shape should be much narrower [7, 8].

The auditory modelling approach to speaker normalisation is exemplified by the work of Bladon **et al** [1]. In this model, the spectral frequency axis is transformed from hertz to bark prior to filtering using the filter shape described by Schroeder (see [5]). Because of the broadness of these filters, there is a "smearing" of the spectrum with a substantial loss of resolution rendering different realisations of the same vowel more alike and removing much of the fine detail due to voicing. Following a conversion from intensity to loudness density to yield an "auditory spectrum", a linear shift in the bark dimension is effected. From the data presented, it is apparent that such shifts can have a normalising effect, by bringing vowel spectra for male and female speakers into reasonable coincidence. Following this work, Holmes [9] attempted to investigate the perceptual effect of bark-scaled shifts in formant frequencies using a speech synthesis-by-rule system. Preliminary results suggest that, for some vowels at least, an approximately constant bark difference between F_1 and F_2 is necessary to maintain phonetic quality.

The principal objection to the Bladon model is the use (following Schroeder) of a wideband auditory filter. Klatt [10] has observed that male and female speech can be made to look similar merely by increasing the bandwidth of the analysis filter in the spectrogram. Thus, caution must obviously be exercised to ensure that vowel identity is preserved when the variance is reduced in this way. There is little virtue in making the same vowel from different speakers appear more alike if different vowels from the same speaker also look more alike. It is only to be expected that representations preserving gross features only of the spectrum shape would be more likely to improve similarity between male and female vowel spectra, since a lot of information (whether relevant or not) has been discarded. It is important to know, therefore, what information is left in the smoothed spectrum representation. One way to discover this might be to conduct listening experiments with speech resynthesised directly from the auditory spectrum. Such resynthesis also offers a means of studying the perceptual effect of bark-scaled shifting, much as Holmes has done, but with real (rather than synthetic) speech.

One difficulty with this approach is apparent. If the auditory system really does perform a frequency smearing operation, then the resynthesised speech will naturally be subjected to this operation. I.e. the speech will be smeared "twice", hence possibly invalidating the idea of testing by resynthesis. Evidence that the smeared, auditory representation is adequate to retain vowel identity is given below. Of course, it may be that a second application of the smearing has relatively little effect, most of the data reduction being done on the first application. One early priority, therefore, must be to compare smoothed and unsmoothed speech for perceptual differences.

IMPLEMENTATION DETAILS OF THE MODEL

An auditory model based closely on that described by Bladon **et al** [1] has been implemented on a DEC MicroVAX computer. As well as "forward" acoustic-to-auditory transformations, some provisional "inverse" auditory-to-acoustic transformations have also been included to allow resynthesis.

Forward Transformations

The excitation patterns for the auditory model are computed as follows. The power spectrum S(f) for the input speech is computed over (Hamming weighted) time windows of approximately 32 ms using an FFT algorithm. The windows are advanced in steps of 8 ms for each new segment. The power spectrum (with units of V^2/Hz) is then transformed to a critical band density (with units V^2/bark) using the formula:

$$S(z) = S[f(z)] \cdot \frac{df}{dz}$$

The mapping between frequency, f, and critical band number, z, is approximated by the expression due to Traunmuller [11]:

$$f = \frac{1960(z + 0.53)}{(26.28 - z)}$$

Thus, the critical band density is computed from the spectrum by the f -> z mapping followed by multiplication with the density conversion factor.

Next, an excitation pattern is computed from the critical band density by convolution with the auditory filter frequency response. The specific filter used at this stage is Schroeder's (as described in [5]) but we intend to investigate the use of different filters. The convolution operation is equivalent to using a filter bank analysis, but is more convenient as the filter shape (as defined by Schroeder) is invariant across the bark scale, and no weighting has to be applied to account for changes in filter bandwidth.

The Bladon model differs slightly from Schroeder's in the calculation of the loudness density pattern, which is accomplished by conversion from critical band density to loudness level density in phons/bark followed by a conversion to loudness density in sones/bark. In this work, we have neglected to compute the loudness density pattern: justification for this omission in terms of resynthesis is that the phon curves are fairly flat in the region 200 - 4 kHz where the formants lie, and thus a displacement of the pattern along the bark scale would have little effect on the spectrum.

Once the excitation pattern has been calculated for the input segment, its position on the bark scale can be adjusted before resynthesis in order to investigate the perceptual effects of displacement.

Inverse Transformations

Since the filtering (convolution) operation has effected a data reduction on the original spectrum, it is impossible to recover the full spectrum for resynthesis. Some indirect evidence that the smoothed, auditory spectrum is a reasonable representation from which to resynthesise is given by certain other psychoacoustic findings. Using the relatively broad Schroeder filters, the physical formant pattern is smoothed to just two auditory peaks. This characteristic is consistent with the "centre of gravity" theory advocated by Chistovich and Lublinskaya [12] as well as with experiments in the matching of two-formant synthetic vowels to the full reference vowel - the so-called F-prime paradigm [13, 14]. Thus, the auditory spectrum should in principle be capable of retaining information concerning vowel identity. Confirmation of this notion is given in the work of Hermansky et al [15] who processed all-voiced sentences to show that a "reduced" spectrum produced by auditory filtering (18 critical band filters equispaced in the bark dimension) could yield "intelligible" speech.

The resynthesis operation involves conversion of the critical band density back to a spectral density by multiplication with the inverse density conversion factor, dz/df. However, the smearing operation removes much, if not all, of the voicing information. For the resynthesis process, therefore, two possibilities present themselves. Either the loss of voicing information could be ignored or appropriate voicing could be added. We intend to explore both of these approaches.

Finally, continuous speech output is obtained from the auditory spectra by inverse Fourier transformation using an overlap-add technique [16].

RESULTS

At this early stage, it is only possible to give initial results from some informal listening tests. The oral presentation will describe results of more extensive testing. Speech of telephone quality (low-pass filtered at 3.2 kHz and sampled at 8 kHz with 12-bits resolution) has been processed by the model. Two complete sentences have been studied: a male speaker saying "live wire should be kept covered" and a female saying "the kitten chased the dog down the street". At the resynthesis stage, no extra voicing has been added.

We wished first to examine the effect of smoothing but without shifting in the bark dimension. The speech was subjected to the forward transformations with zero bark shift and resynthesised. The speech output was slightly degraded but speaker identity was retained and the sentence was clearly intelligible. This observation lends weight to the belief that resynthesising is a valid technique for testing auditory models. If anything, the result extends the observation of Hermansky et al referred to above to speech consisting of voiced and unvoiced segments.

Subsequently, the effect of processing the male speech using the model, and including a shift of one bark, was investigated. Again, the speech was intelligible but more severely degraded. We speculate that this additional degradation is principally due to destroying the harmonic relation between voicing-frequency components when a linear bark shift follows the non-linear hertz-to-bark transformation. First impressions, however, were that speaker identity was markedly different. It was not easily possible to assign a perceived sex to the speaker with any confidence.

FUTURE WORK

The major priority is to conduct more formal matching experiments (perhaps using steady-state vowels) with a larger number of listeners.

Informal experimentation so far has not used added voicing. Further work is planned in which the speech spectrum will be deconvolved by cepstral techniques into excitation and envelope components. The envelope alone will be processed by the model and speech resynthesised with a variety of voicing components appropriate to different speakers (and including the natural voicing itself).

There are, of course, many specific details of the model which could be further tested by resynthesis. For instance, there is a good case to be made for employing auditory filters of much narrower bandwidth, such as the rounded-exponential (roex) filters described by Moore and Glasberg [8]. Arguably, in this case, equivalent rectangular bandwidth (ERB) would be a more appropriate frequency scale for shifting than the bark scale.

REFERENCES

[1] Bladon, R.A.W., Henton, C.G. & Pickering, J.B. (1984) 'Towards an auditory theory of speaker-sex normalisation', Language and Communication', 4, 59-69.

[2] Peterson, G.E. & Barney, H.L. (1952) 'Control methods used in the study of vowels', JASA, 24, 175-184.

[3] Potter, R.K. & Steinburg, J.C. (1950) 'Towards the specification of speech', JASA, 22, 807-820.

[4] Fletcher, H. (1940) 'Auditory patterns', Rev. Mod. Phys., 12, 47-65.

[5] Group Report (Fourcin, A.J. et al) (1977) 'Speech Processing by Man and Machine' in "Recognition of Complex Acoustic Signals", T.H. Bullock (Ed), Life Sciences Research Report 5, Abakon Verlag, Berlin.

[6] Zwicker, E. (1963) 'Uber die Lautheit von ungedrosselten und gedrosselten Schallen', Acustica, 13, 194-211.

[7] Patterson, R.D. (1976) 'Auditory filter shapes derived with noise stimuli', JASA, 59, 640-645.

[8] Moore, B.C.J. & Glasberg, B.R. (1983) 'Suggested formulae for calculating auditory filter bandwidths and excitation patterns', JASA, 74, 750-753.

[9] Holmes, J.N. (1985) 'Normalization in vowel perception' in "Invariance and Variability of Speech Processes", J.S. Perkell and D.H. Klatt (Eds), Lawrence Erlbaum Associates.

[10] Klatt, D.H. (1982) 'Speech processing strategies based on auditory models' in "The Representation of Speech in the Peripheral Auditory System", R. Carlson & B. Granstrom (Eds), Elsevier Biomedical.

[11] Traunmuller, H. (1983) 'Analytic expressions for the tonotopical sensory scale', Unpublished manuscript.

[12] Chistovich, L.A. & Lublinskaja, V.V. (1979) 'The "centre of gravity effect" in vowel spectra and critical distance between the formants: psycho-acoustical study of the perception of vowel-like stimuli', Hearing Research, 1, 185-195.

[13] Bladon, R.A.W. (1983) 'A study of two formant models for vowel identification', Speech Communication, 2, 295-303.

[14] Paliwal, K.K., Ainsworth, W.A. & Lindsay, D. (1983) 'A study of two-formant models for vowel identification', Speech Communication, 2, 295-303.

[15] Hermansky, H., Hanson, B.A. & Wakita, H. (1985) 'Perceptually based linear predictive analysis of speech', Proc. ICASSP 85, 509-512.

[16] Allen, J.B. & Rabiner, L.R. (1977) 'A unified approach to short-time Fourier analysis and synthesis', Proc. IEEE, 65, 1558-1564.