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 minimized using a suitable computational model based on known or assumed details of human auditory processing. We are attempting to test this notion experimentally by resynthesizing speech which has been processed by the model and studying its perceptual nature.

INTRODUCTION

Current approaches to speech recognition are characterized 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 decimals (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 prerequisite for further, significant advances in speech technology, encompassing recognition, synthesis and coding. Indeed, there have been several recent attempts to embody at least some of the currently available computer 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 recognisers.

One area where conventional signal processing and statistical pattern matching techniques have proved inadequate is in the handling of speaker variability such as that arising from sex and age differences. This sort of variability poses serious problems for speaker-independent recognition. Recently, Bladon and his coworkers [1] have suggested that many of these differences could be minimized (in vowel spectra at least) using a suitable "auditory normalisation" model. In Bladon's model, a perceptually-activated "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 directly from the shift-aided auditory spectral representation.

For instance, resynthesising a one-bark-increased 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 "incorrected" vowel with the male voicing retained should yield shifts in perceived quality. Indeed, it may even prove possible to effect an automatic correction of sex, race, or voice, or vice versa. We are attempting to substantiate these ideas experimentally and this paper reports on the early stages of such an investigation.

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 affect 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 formants (and voicing) structure. This normalisation process implies an ability to make allowances for different vocal tract sizes and shapes. In attempting to model this ability in a computational model, we might take either of two somewhat different approaches. One possibility is to adopt a speech production viewpoint whereby the supralaryngeal vocal tract is treated as the "forbidden" (acoustic-to-auditory) transformations affecting a data-reduction and so are inherently non-invertible. 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 appears that such shifts can normalise the effect, by bringing vowel spectra for sale and female speaking into reasonable coincidence. Following this work, Holmes [9] has attempted to investigate the perceptual effect of bark-shifted formants by using a speech synthesis-by-rule system. Preliminary results suggest that, for some vowels at least, an approximately constant bark difference between F1 and F2 is necessary to maintain phonetic identity.

The principal objective of 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 more alike by increasing the bandwidth of the analysis filter, in the spectrogram. Thus, caution must obviously be exercised to ensure that vowel quality will not be adversely affected. When the variance in the model 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 group 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 such resynthesised speech. Such resynthesis also offers a means of studying the perceptual effect of bark-scaled shifting, such 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 measuring operation, then the resynthesised speech will necessarily be subjected to this operation, i.e. the speech will be heard "twice", hence possibly invalidating the internal model of the resynthesiser. Evidence that the measured, auditory representation is adequate to retain vowel identity is given below. Of course, it may be that a second application of the data reduction being done on the first application. One early priority, therefore, must be to compare smoothed and unsmoothed speech for perceptual equivalence. The effectiveness of this preliminary study is given below. In attempting to mimic this effect ory) transform inverti:Idata-reduction and so are inherenzlions

IMPLEMENTATION DETAILS OF THE MODEL

An auditory model based closely on that described by Bladon et al [1] has been implemented on an 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) overlapping Fourier transform windows of approximately 30 ms using the FFT algorithm. The window size is adjustable in steps of 8 ms for each segment. The power spectrum (with units of
of \( f'(v) \) is then transferred to a critical band density (with units \( v^{-1} \)) using the formula:

\[
\text{SISI} \times \frac{f'(v)}{v} = \text{SISI}
\]

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

\[
f = 20 \log_{10} f + 1.0\text{ S}_0.5(\text{kHz})
\]

The adapted conventional Bark scale [15] processed all voiced sentences to show that a "reconstructed" 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, \( \frac{1}{\text{SISI}} \). However, the mapping 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 spectrum 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 actual presentation will describe results of more extensive tests of the model. To validate the model, two controlled experiments have been conducted by the model. Two male speakers saying "live wire should be kept covered" and a female saying "the kitten chased the dog down the street" in the resynthesis stage, no extra voicing has been added. We wished first to examine the effect of smoothening without shifting in the Bark dimension. The speech was then transferred 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 also implies that the resynthesis is a valid technique for the transformation of speech in the Bark dimension. If anything, the result indicates the observation that the Bark dimension is preserved above to speech consisting of voiced and unvoiced segments.

Subsequently, the effect of processing the male bark was investigated. Again, the speech was slightly degraded but speaker identity was retained and the sentence was clearly intelligible. This observation also implies that the Bark dimension is preserved above to speech consisting of voiced and unvoiced segments.

**FUTURE WORK**

The major priority is to conduct more formal matching experiments (perhaps using steady-state voices) 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 convolved with 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 (roc) filters described by Noice and Glaser [8].

Arguably, in this case, equivalent rectangular bandwidth (ERE) would be a more appropriate frequency scale for shifting than the Bark scale.

**REFERENCES**


