TESTING AN AUDITORY MODEL BY RESYNTHESIS

R.W. Hukin and R.I. Damper

Department of Electronics and Computer Science, University of Southampton, Highfield, Southampton SO9 5NH, U.K.

Abstract

This paper describes the use of a resynthesis strategy in testing an auditory model, specifically a version of the DOMIN model. The steps by which the original speech is processed to produce a reduced, auditory (DOMIN) spectrum are described; subsequently, the reduced representation is used to produce resynthesised speech having the same auditory spectrum. In light of this equivalence, the extent to which the original and resynthesised speech are perceptually equivalent is argued to be a good test of the model. We show that a spectral representation in which approximately two-thirds of the FFT frequency components are discarded, but the DOMIN representation is unchanged, can produce resynthesised speech of high intelligibility. We conclude that the DOMIN model retains important information pertaining to the identity of both vowels and consonants. Further, we present evidence showing that testing by resynthesis is superior to the alternative techniques for assessing auditory models.

1 Introduction

Computational models of auditory peripheral processing have become increasingly popular in recent years. The presumption is that, by embodying details of such processing, these models will implicitly emphasise perceptually important information. Because of this, auditory modelling has proved of interest from both theoretical and practical points of view. Theoretical interest centres on computational modelling as a means of exploring auditory function whereas, practically, speech technologists are concerned with better representations of the signal for applications such as recognition and voicing. However, unless we have methods at our disposal for testing a model's effectiveness in retaining perceptually important features, we cannot hope to improve such models nor choose rationally between them.

Bladon [1] outlines methods for evaluating auditory models; one common technique is to use the model as a pre-processor, or "front-end", to an automatic speech recogniser. The model is then considered to be improved when the recognition rate improves. However, the result of such testing will inevitably be influenced by the recogniser characteristics themselves, such as the particular features extracted and the distance metric employed. An alternative possibility is to calibrate the model against human perceptual data, but problems of language bias and task differences make this difficult. In earlier work, we introduced the idea of testing by resynthesis [2,3]. According to this idea, two different acoustic signals having the same auditory representation should sound "equivalent" in some important respect. We believe that the extent of the perceptual difference between two such signals is a good measure of the model's quality. Thus, evaluation is performed by listening tests without any necessity to consider the many extraneous factors due to an automatic recogniser.

Here, we describe the application of testing by resynthesis to an auditory model incorporating Carlson and Granström's DOMIN technique [4]. The DOMIN approach is based on the notion of particular frequencies dominating regions of the basilar membrane response, and was selected for its simplicity whilst retaining a degree of physiological plausibility. To date, the only test of this model has been by Blumberg et al who used it as the "front-end" to a conventional speech recogniser [5], observing the recogniser's performance for different inputs. Their conclusion was that the DOMIN representation was highly appropriate for vowels but much less so for consonants. This was felt to be a consequence of the emphasis which DOMIN gives to the frequency location of spectral peaks. Using testing by resynthesis, however, we will show that consonants are indeed well-represented in DOMIN. This result indicates the superiority of testing by resynthesis over the alterna-
The paper is structured as follows. We first outline briefly the DOMIN model before describing our complete analysis and resynthesis system. We then present the results of testing the DOMIN model by resynthesising various monosyllabic stimuli of the sort used by Blomberg et al. Finally, we draw conclusions and detail work in progress.

2 The DOMIN Representation

Any auditory model seeks to represent the coding of acoustic information at some useful level of description. Since the auditory nerve is the only link between the periphery and the brain, all the information processed by the higher levels of the system must exist in the auditory nerve firing patterns; this is, therefore, an appropriate level at which to work. Although many advances have been made towards understanding the mechanisms of the peripheral auditory system, little is known in detail about how information is coded in the auditory nerve. For example, there is still debate as to whether information is coded in the timing or rate of neural firing, or perhaps a combination of both. Results from recent animal experiments [6,7] are in favour of a temporally-based coding rather than a place-rate code as had previously been suggested, although the exact nature of the temporal coding is uncertain. The two most popular ways of summarising the information in the pattern of firing are representations based on the synchronisation index or the average localised synchronised rate (ALSR). The synchronisation index is defined as the magnitude of a particular harmonic in the Fourier transform (of the spike train) divided by the number of spikes in the record. The ALSR is defined as the average value of the response at each harmonic for a population of fibres with centre frequencies (CF's) within a certain frequency range of interest.

The DOMIN model proposed by Carlson and Granström [4] uses a representation similar to the ALSR but simpler, in that details of individual nerve firings are not explicitly considered. DOMIN assumes that certain frequency components of an input signal to the auditory system will dominate particular positions along the basilar membrane; these dominant frequencies will be reflected in the firing patterns of the auditory nerve fibres across frequency. A critical-band analysis is used to find the dominant frequencies (see below). The output of the DOMIN model is normally depicted as a plot of dominant frequency as a function of filter CF (in barks). Alternatively, it can be plotted in histogram form i.e. the number of analysis filters dominated by a certain frequency, again plotted as a function of frequency. We find the latter to be a more appropriate representation; the height of the histogram can be regarded as a measure of the number of fibres whose firing patterns are dominated by that frequency.

![Figure 1: Processing steps in implementing the DOMIN model and the corresponding resynthesis strategy.](image-url)
tween resynthesised and original speech must be due to the model.

3 Model Implementation

The auditory model used in our work is essentially that due to Bladon and Lindblom [8], with modifications to include the DOMIN processing. Figure 1 shows the processing steps involved, as well as depicting the resynthesis scheme (described below).

The input signal is segmented using a 25.6 ms Hamming window, with adjacent windows overlapping by 6.4 ms. A 256-point FFT of each speech segment is taken, followed by a transformation of the spectrum onto the bark scale as detailed in Schroeder et al. [9]. The spectrum is then processed by an auditory filter-bank comprising 128 filters (each based on a prototype proposed by Schroeder [10]), equally spaced on the bark scale. The dominant frequency in the output of each filter is found using a simple peak-picking algorithm.

Figure 2(a) shows an example spectrum from a segment of the vowel in /hid/, spoken by a male, with the corresponding DOMIN histogram pattern (Fig. 2(b)).

4 Resynthesis

In testing by resynthesis, the synthesised signal used to test the model is derived from the input signal — hence the term. The procedure for doing this, based on the DOMIN representation, is depicted in the lower part of Figure 1. The first step is to modify the input spectrum by simply zeroing those components that do not correspond to dominant frequencies. This results in a very significant reduction (by about two-thirds) in the number of non-zero components. Fig. 2(c) shows a typical modified spectrum. After bark-to-hertz scaling, the modified spectrum is then transformed into the output (time domain) signal via an inverse FFT. The resynthesised speech frames are merged into a continuous signal using an overlap-add technique.

5 Results

The first question addressed was the effectiveness of the DOMIN vowel representation. The stimuli used for testing were a series of 7 /hVd/ tokens, where V is a vowel. The tokens were spoken by one male and one female, recorded in low-noise conditions, lowpass filtered at 4.5 kHz, and sampled at 10 kHz to 12-bit resolution. Informal listening tests by the authors revealed the resynthesis intelligibility to be excellent, neither speaker nor phonetic identity were in any doubt. These tests confirm the finding of Blomberg et al that DOMIN is very good at representing vowels.

We next considered consonants, believed by Blomberg et al to be poorly represented in DOMIN. As a preliminary test, we processed the sentence The little baby sits with his daddy with the model, and found that the resynthesised speech was still of good quality, and highly intelligible with the consonants being clearly discriminable.

In a further test, and in order to make our results more readily comparable with those of Blomberg et al, we processed 7 /Cret/ stimuli (where C is a con-
sonant) from a single male speaker and conducted listening tests for both the original and resynthesised speech. The stimuli were chosen to exemplify both voicing and place distinctions. Seven listeners were played 2 original and 2 resynthesised tokens for each of the 7 stimuli (28 tokens in total) in random order, and were asked to decide which of the 7 possible consonants they heard.

Consonant recognition for the original, unprocessed tokens was 98.0% with the only confusions (2 in number) being *that* for *vat*. For the DOMIN-resynthesised tokens, however, the recognition rate remained high at 96.9%. In addition to 2 *that/vat* confusions, there was a single *bat/vat* confusion. The very minor differences in the number of errors between the original and resynthesised tokens suggest that DOMIN does adequately represent consonants.

Blomberg et al felt that a possible reason for the poor performance of their DOMIN-based speech recogniser on consonants was that "loudness information was being disregarded, which makes the discrimination between consonants problematic, especially if they have the same place of articulation" [5]. However, we would argue that loudness is encoded (albeit crudely) in the number of filters dominated by a particular frequency. For example, the relative magnitude (and hence loudness) of a frequency component in a given input spectrum will have a direct bearing on whether that frequency is dominant and if so, how many filters it dominates.

6 Discussion

Our results suggest that the reason for the poor performance of Blomberg et al's DOMIN-based speech recogniser is related to their testing strategy rather than the DOMIN representation itself. By contrast, testing by resynthesis avoids all possibility of confounding any imperfections of the representation with those of the recogniser.

In as much as highly intelligible speech can be resynthesised from the much-reduced DOMIN representation, its underlying assumptions seem to be well-founded. The testing by resynthesis approach we have employed in this work is currently being applied to a more detailed (physiologically-based) model of the peripheral auditory system [11], in which we are using an ALSR representation in place of DOMIN.

7 References