MODELLING THE RECOGNITION OF SPECTRALLY REDUCED SPEECH

Jon Barker and Martin Cooke

{j.barker,m.cooke}@dcs.shef.ac.uk
Department of Computer Science, University of Sheffield, Sheffield, UK

ABSTRACT

Progress in robust automatic speech recognition may benefit from a fuller account of the mechanisms and representations used by listeners in processing distorted speech. This paper reports on a number of studies which consider how recognisers trained on clean speech can be adapted to cope with a particular form of spectral distortion, namely reduction of clean speech to sine-wave replicas. Using the Resource Management corpus, the first set of recognition experiments confirm the high information content of sine-wave replicas by demonstrating that such tokens can be recognised at levels approaching those for natural speech if matched conditions apply during training. Further recognition tests show that sine-wave speech can be recognised using natural speech models if a spectral peak representation is employed in concert with occluded speech recognition techniques.

1. INTRODUCTION

Clean speech and speech with additive noise have been the primary conditions employed in most ASR studies. Notwithstanding progress in recent years, error rates remain significantly higher than those exhibited by listeners (one and two orders of magnitude greater for clean and noisy speech respectively [15]). Yet listeners are remarkably adept at recognising many other forms of speech-like stimuli, including those which have undergone various forms of severe spectro-temporal distortion (e.g. speech replicas employing only three time varying sinusoids [18]; speech that has been reduced to the output of two extremely narrow and widely spaced band-pass filters [20]; spectral alternation [4]; spectral smearing [1]; temporal desynchronisation across spectral bands [11]; spectro-temporal [13] and temporal attenuation [19] – see [8] for further examples). Speech under these conditions can be immediately intelligible to naive listeners who have had no previous exposure to the distorted speech. How then do these untrained listeners employ their previous experience of natural speech to interpret such heavily distorted utterances? An answer to this question is important not only for a better understanding of speech perception, but also for the light it might cast on robust representations and recognition strategies for ASR.

In this paper, we focus on a particular form of spectral distortion, namely reduction of clean speech to sine-wave replicas [18], and report on a number of ASR studies which address the question of how recognisers trained on clean speech can be adapted to cope with spectrally-reduced speech. We demonstrate results using the Resource Management corpus that suggest such distortions can be accommodated within the evolving framework of occluded speech recognition [12, 9].

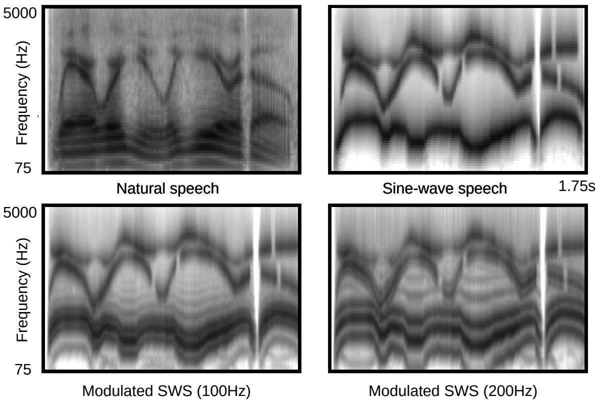

Figure 1: **'Auditory spectrograms' illustrating the differences between natural speech, sine-wave speech and modulated sine-wave speech for the utterance, "Where were you a year ago?" Frequency is an ERB-rate scale.**

2. AUTOMATIC RECOGNITION OF SINE-WAVE SPEECH

Sine-wave speech (SWS) is a spectrally reduced form of speech produced by using time-varying sinusoids to mimic the amplitude and frequency variation of the first few formants of a natural utterance (see figure 1). Using only three such sinusoids, a stimulus is produced which, although highly unnatural, can nevertheless be easily intelligible to listeners with no previous experience of the distortion [18]. This phenomenon poses questions about the perceptual organisation of speech and the cues necessary for phonetic categorisation [17, 2]. Further, SWS intelligibility can be significantly increased by applying amplitude modulation, an effect most pronounced at a modulation frequency of 100 Hz [7]. Two possible explanations of this effect have been considered. The first suggests that modulation acts as a cue to increase the perceptual coherence of the acoustic signal. The alternative proposes that the modulation side-bands make the stimulus more speech-like.

The experiments reported in this section examined recognition performance of both unmodulated SWS and SWS modulated at 100 and 200 Hz (MSWS). SWS replicas were generated for each utterance in the Resource Management (RM) task [16]. Amplitude and frequency trajectories of the first three formants, necessary for sine-wave synthesis, were extracted using the Crowe formant tracker [10]. The intelligibility of the SWS generated from this automatically extracted formant data has been confirmed in listening studies [2, 3].

Both natural and reduced speech corpora were encoded using an auditory filterbank to produce a 64 channel rate-map representation [5] (as illustrated in figure 1) and further transformed using a DCT resulting in 13 'auditory cepstral' coefficients in each 5 ms frame. HTK [21] was employed to construct a triphone-based HMM speech recogniser for each of the four stimulus conditions (natural speech, SWS, 100 Hz MSWS and 200 Hz MSWS) using cepstral coefficients along with their velocities and accelerations.

Each of the four HMM systems were used to recognise each of the 4 corresponding conditions applied to the test set (feb89 from RM) resulting in a total of 16 recognition experiments, the results of which are shown in table 1.

The main findings are:

i Spectrally-reduced speech can be recognised at levels approaching that for natural speech if matched models are used (diagonal entries in table 1).

ii Recognition in unmatched conditions gives very poor performance. By contrast, listeners typically attain word recognition rates of around 70% on sine-wave speech [2, 3, 17].

iii For models trained on natural speech, the application of modulation leads to a small increase in the accuracy of SWS recognition (with a slightly bigger effect at 100Hz than at 200Hz). This supports the findings of Carrell and Opie [7] in that the small increase in the naturalness of MSWS, as demonstrated by the ASR results, is unlikely to fully account for the greatly increased intelligibility of MSWS over SWS.

3. SPECTRAL PEAKS

In the previous experiment, high SWS recognition accuracies were achieved only for models trained on SWS data. Listeners do not require such training and hence do not appear to explicitly 'model' the novel signal, but instead are able to directly use their prior expectations of natural speech. What does this effortless mapping between SWS and natural speech say about the auditory representations and associated recognition processes employed?

	Test Condition			
Training	Natural	SWS	100 Hz	200 Hz
Natural	90.8	4.9	9.6	8.3
SWS	10.7	80.5	83.5	73.2
100 Hz	24.3	70.8	80.6	76.4
200 Hz	15.0	40.7	75.0	78.3

Table 1: **Word recognition accuracies for natural and SWS train/test conditions.**

Like most other ASR systems, the study described in section 2 used a cepstral representation of speech. Whilst such a basis is useful reducing the number of parameters to be estimated during training, it has the drawback of all non-spectral representations in suffering severe distortion under operations (such as band-pass filtering and sine-wave speech synthesis) that selectively disrupt restricted spectral regions. Human speech recognition copes with these band-limited distortions by making full use of the natural redundancy of speech (precisely the correlations in the feature vector that cepstral coding minimises). Whole-spectrum cepstral coding is unlikely to form an adequate basis for speech perception in listeners.

Vowel perception studies have demonstrated the importance of spectral peak locations relative to spectral valleys (e.g. [6]), and of course the resistance of such regions to occlusion by competing noise sources confers a degree of robustness. Could a peak-based account also explain the ease of SWS recognition? Examination of auditory spectrograms of SWS utterances (see figure 1) suggests that although sine-wave synthesis heavily distorts the spectral profile (formant bandwidths are effectively narrowed and the harmonic structure is lost), the relative heights and frequencies of the major spectral peaks are preserved.

A second set of experiments was conducted to determine whether this peak-invariance can be exploited to recognise spectrally-reduced speech from models trained on natural speech.

3.1. Method

Feature vectors were derived directly from 64 channel auditory rate-maps. However, rather than using the full 64 point profile of each frame of training data, spectral peaks alone were used i.e. the mean feature vector μ for each state of each HMM was computed from the training set according to:

$$\mu_j = \frac{\sum_{i=1}^{N} x_{i,j}.peak(i,j)}{\sum_{i=1}^{N} peak(i,j)}$$

$$peak(i,j) = \begin{cases} 1 & x_{i,j} > x_{i,j-1} \quad \& \quad x_{i,j} > x_{i,j+1} \\ 0 & otherwise \end{cases}$$

where $x_{i,j}$ is the jth channel of the ith of N training vectors. Variances were computed similarly. During testing, data was treated as being unknown except at the locations of the peaks, and missing data techniques [9] were applied to compute observation likelihoods for each HMM state.

3.2. Results and Discussion

Recognition performance for natural speech using peaks-only and whole-spectrum approaches is summarised in table 2.

Training Condition	Test Condition	
	Whole Spectrum	Peaks
Whole Spectrum	68.1	74.9
Peaks	—	75.7

Table 2: **Natural speech recognition accuracy for whole-spectrum and peaks-only representations.**

These results indicate that irrespective of how the models are trained, a significant improvement (about 8%) is gained if spectral peaks alone are used during recognition. This can be explained by the observation that whilst spectral peaks are not entirely independent, they suffer less from the inherent redundancy of processing in overlapping filter channels. This relative independence makes the spectral-peak representation better matched to the variance-only feature vectors of the HMMs employed.

A further small improvement is gained for models *trained* on peaks. This is obtained in spite of the fact that only a fraction of the training data is available (there are on average approximately 8 peaks per each 64 channel frame of test data). Two factors may contribute to this improvement. First, since harmonics in the F1 region are resolved by the narrow filters of the auditory representation, use of the entire profile leads to large variances. These arise from the movement of harmonic peaks due to F0 variation across tokens. The peaks themselves trace out the true shape of F1 and thus overcome variability due to different F0s. This effect is illustrated in figure 2 which compares feature vector variances for whole-spectrum and peaks-only models. Second, training on the full profile results in models that systematically underestimate mean *peak* values. Again, this effect is especially apparent in the first formant region where harmonic peaks are averaged with harmonic dips (see figure 3). With more training data, leading to better estimation of spectral peaks, recognition accuracy may be further improved.

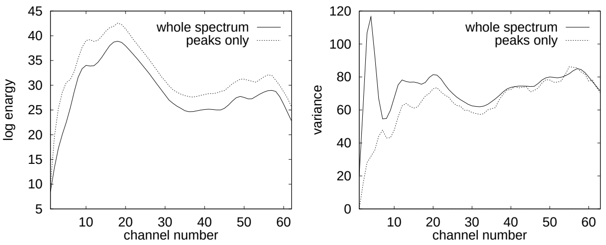

Figure 2: **Comparison of model means (left) and variances (right) when training on either the whole profile or peaks only. Graphs represent averages over all HMM triphone states.**

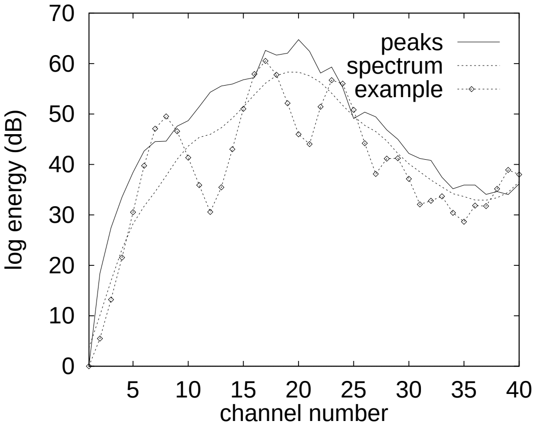

Figure 3: **Models trained on the entire spectral profile underestimate typical spectral peak values in the F1 region.**

Turning now to recognition of SWS, the first row of table 3 demonstrates performance which is significantly improved over that obtained using the cepstral representation. Thus, the combination of a spectral peak representation and application of missing data techniques allows SWS to be recognised with models trained on natural speech, with a relative performance similar to that obtained by listeners [3]. Furthermore, the technique demonstrates a certain degree of robustness to missing data: even when entire sinusoids are removed, resulting in a 2 ‘formant’ condition, significant recognition was still possible. Unsurprisingly, single sinusoid recognition accuracy dropped to chance levels. It appears that whereas one formant region alone does little to constrain recognition, significant performance is possible when two are combined. This effect echoes studies of filtered speech intelligibility [20, 14] in which intelligibility increases in a supra-additive fashion when widely separated frequency bands are combined. The relative improvement in recognition performance as the number of sinusoids present increases is also comparable with human performance as demonstrated in previous SWS listening studies [17].

4. CONCLUSIONS

These studies demonstrate the following:

F1	F2	F3	%acc
√	√	√	43
√	√		24
√		√	27
	√	√	28
√			4
	√		4
		√	3
Natural Speech			76

Table 3: **Recognition results for systems trained on natural speech and employing missing data techniques. Word accuracy figures are shown for recognition of natural speech and SWS with either 1, 2 or 3 sinusoids.**

i Although spectral reduction to sine-wave speech represents a severe distortion of natural tokens, the information content of the signal is still sufficiently high to afford good recognition in matched training conditions.

ii Poor performance in unmatched conditions suggests that cepstral coding techniques are inappropriate for dealing with drastic alterations to the shape of the spectral profile caused by spectral reduction.

iii Unmatched conditions can be successfully modelled by applying missing data techniques in conjunction with a spectral peak representation. Performance on reduced speech relative to natural speech is quantitatively similar to that attained by listeners.

Spectral reductions are just one form of distortion which listeners are capable of handling. A consideration of other modifications will be required for an adequate model of speech perception, which in turn promises to benefit robust ASR.

5. REFERENCES

[1] T. Baer and B.C.J. Moore. Effects of spectral smearing on the intelligibility of sentences in noise. *J. Acoust. Soc. Am.*, 94(3):1229–1241, 1993.

[2] J.P. Barker. *The relationship between auditory organisation and speech perception: Studies with spectrally reduced speech (in preparation).* PhD thesis, Sheffield University, U.K., 1997.

[3] J.P. Barker and M.P. Cooke. Is the sine-wave cocktail party worth attending? In *Proceedings of the 2nd Workshop on Computational Auditory Scene Analysis*, Nagoya, Japan, 1997. Int. Joint Conf. Artificial Intelligence.

[4] J.A. Bashford and R.M. Warren. Effects of spectral alternation on the intelligibility of words and sentences. *Perception and Psychophysics*, 42(5):431–438, 1987.

[5] G.J. Brown and M.P. Cooke. Computational auditory scene analysis. *Computer speech and language*, 8:297–336, 1994.

[6] R. Carlson, B. Granstrom, and D. Klatt. Vowel perception: The relative perceptual salience of selected acoustic manipulations. *STl-QPSR*, pages 3–4, 1979.

[7] T.D. Carrell and J.M. Opie. The effect of amplitude comodulation on auditory object formation in sentence perception. *Perception and Psychophysics*, 52:437–445, 1992.

[8] M.P Cooke. Auditory organisation and speech perception: Arguments for an integrated computational theory. In *ESCA ETRW on The Auditory Basis of Speech Perception*, pages 186–193, Keele, 1996.

[9] M.P. Cooke, A.C. Morris, and P.D. Green. Missing data techniques for robust speech recognition. In *Proc. ICASSP 97*, pages 863–866, Munich, 1997.

[10] A.S. Crowe. Generalised centroids: A new perspective on peak-picking and formant extraction. In W.A. Ainsworth and J.N. Holmes, editors, *Proc. 7th symposium of FASE (SPEECH '88)*, pages 683–690. Los Altos, Calif., 1988.

[11] R. Drullman, J.M. Festen, and R. Plomp. Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.*, 95(2):1053–1064, 1994.

[12] P.D. Green, M.P. Cooke, and M.D. Crawford. Auditory scene analysis and hidden markov model recognition of speech in noise. In *Proc. ICASSP 95*, pages 401–404, 1995.

[13] P.A. Howard-Jones and S. Rosen. Uncomodulated glimpsing in "checkerboard" noise. *J. Acoust. Soc. Am.*, 93(5):2915–2922, 1993.

[14] R.P. Lippmann. Accurate consonant perception without mid-frequency speech energy. *IEEE Trans. on Speech and Audio Processing*, 4(1):66–69, 1996.

[15] R.P. Lippmann. Speech perception by humans and machines. In *ESCA ETRW on The Auditory Basis of Speech Perception*, pages 309–316, Keele, 1996.

[16] P. Price, W.M. Fisher, J. Bernstein, and D.S Pallet. The DARPA 1000-word Resource Management database for continuous speech recognition. In *Proc. ICASSP 88*, pages 651–654, 1988.

[17] R.E. Remez, P.E. Rubin, S.M. Berns, J.S. Pardo, and J.M. Lang. On the perceptual organization of speech. *Psychological Review*, 101(1):129–156, 1994.

[18] R.E. Remez, P.E. Rubin, D.B. Pisoni, and T.D. Carrells. Speech perception without traditional speech cues. *Science*, 212:947–950, 1981.

[19] W. Strange, J.J. Jenkins, and T.L. Johnson. Dynamic specification of coarticulated vowels. *J. Acoust. Soc. Am*, 74(3):695–705, 1983.

[20] R.M. Warren, K.R. Riener, Bashford J.A., and B.S. Brubaker. Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits. *Perception and Psychophysics*, 57(2):175–182, 1995.

[21] S.J. Young and P.C. Woodland. *HTK Version 1.5.* Cambridge University Engineering Dept., 1993.