EXPLOITING THE AUDITORY INDUCTION EFFECT IN ROBUST SPEECH RECOGNITION

Malcolm Crawford, Martin Cooke and Phil Green
The Institute for Language Speech and Hearing, The University of Sheffield
West Court, 2 Mappin Street, Sheffield S1 1DT
{m.crawford, m.cooke, p.green}@dcs.shef.ac.uk
http://www.dcs.shef.ac.uk/research/ilash/

ABSTRACT

This paper builds on previous work on recognition of speech in noise to incorporate a model of the constraints imposed by rules governing auditory induction, implemented as a refinement of the distance metric used in a Kohonen network. We show that use of this constraint improves recognition accuracy, and points to a new understanding of some aspects of speech perception.

INTRODUCTION

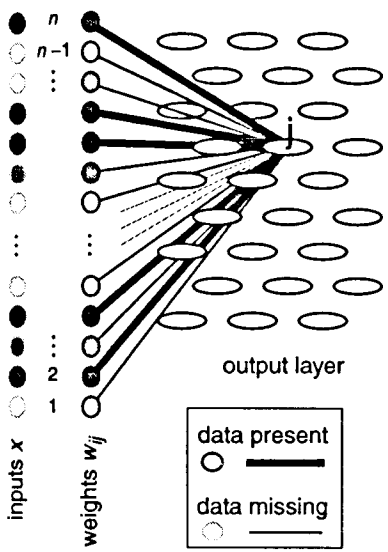
It is a matter of everyday experience that speech perception is possible in "adverse" conditions. Bregman [1] has coined the term *auditory scene analysis* to refer to listeners' ability to separate out individual sound sources in a mixture by grouping together those components of a signal which share characteristics in common (e.g. harmonicity). Recent evidence [2] has suggested, however, that in noisy environments it is not possible for the auditory system to fully recover all the evidence for, say, a speech signal. It is often the case that an intrusive noise source will completely dominate a particular time-frequency region: in these regions there are no components which can reliably be ascribed to the speech signal, so its representation will necessarily be incomplete. Few theories of speech perception, however, consider the effects that extraneous signals have on the auditory representation of speech signals, or account for listeners' abilities to induce the percept of the continuity of a speech signal through noise.

Those parts of the spectrum which cannot be attributed to the speech signal are nevertheless information-bearing. They place an upper bound on the amount of energy that the speech signal could have had. If, for example, evidence is only available for low frequencies (in the F1 region) and this suggests initially that an /i/ might be present, we would minimally expect also to find concentrations of spectral energy in the 2200-3300Hz region, corresponding to F{2,3,4} of the vowel. If this is not observed, it constitutes evidence counter to the initial /i/ hypothesis. This area has been formally investigated in studies of auditory induction: Warren and his colleagues [3] have found that the auditory system is indeed subject to such constraints. In their experiments, part of a stimulus signal is excised and replaced with a noise burst. If and only if the noise is sufficiently loud that it would have masked the original sound, had it been present, the auditory system induces the percept of the original, leading to continuity.

We now show how the auditory induction constraint can be incorporated into a recognition architecture adapted to handle missing data in a bottom-up fashion.

RECOGNITION FROM PARTIAL DATA

We have developed [4] a model of speech recognition, based on a modified Kohonen self-organising network [5], whose performance degrades gracefully as an increasing proportion of the input representation of speech is deleted (simulating the effect of increasing noise intrusions) during training and recognition. During the recognition phase, an input vector is compared with the weight vector



Normal distance metric [5]	$\sum_{i \in present} (w_{ij} - x_i)^2$
Distance metric for incomplete vectors [6]	$\sum_{i \in present} (w_{ij} - x_i)^2$
Distance metric with auditory induction constraint	$\sum_{i \in present} (w_{ij} - x_i)^2 + \sum_{i \notin present} max(0, w_{ij} - x_i)^2$

Figure 1. The input vector is compared with net weight vectors: the winning unit is chosen as that whose weight vector most closely matches the input vector, as determined by a distance metric.

associated with every node in the net. In the general case, all components of the input vector contribute to the distance measure. In the version modified for incomplete input [6] only those components present in the input vector contribute to the score.

As discussed previously, studies of perceived auditory continuity suggest that the full spectral profile places additional top-down constraints on the recognition process. Specifically, "auditory induction" operates only if sufficient energy exists in the occluded regions to account for the induced pattern. This is expressed computationally by the addition of a second factor in the distance metric which adds a penalty for any "absent" components whose level is below that in the net vector. This is illustrated in figures 1 and 2.

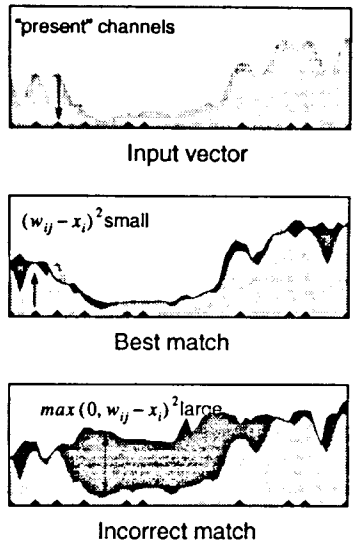


Figure 2. Illustration of the auditory induction constraint: the "best match" panel shows a correct match between the input vector (light grey) and the net weight vector (dark grey). The distance measure is determined primarily by standard metric. The "incorrect match" panel shows a match between the same input vector and a net weight vector which "expects" more energy than is present in the signal.

EXPERIMENTS

A net of dimensions 19x13 was used. The input representation was produced by a 64-channel gammatone filterbank, with channel centre-frequencies equally spaced on an ERB-rate scale between 200 and 5500 Hz, and the output of each filter

processed by a model of inner hair cell transduction, smoothed over a 10 ms window.

Training and test data were generated from utterances produced by a single male Japanese speaker from the ATR large-scale speech database [7]. The nets were trained and calibrated (i.e. one of 27 phone labels was attached to each output node) using a training set, and recognition performance measured in terms of recognition accuracy — the percentage of labels in the test set correctly identified.

Recognition performance was investigated in two series each of 10 different conditions, in which during recognition input vector components were deleted at random with a probability which varied in from 0.0 (no deletion) to 0.9 (90% deletion). In the first series the distance metric for incomplete vectors was used, in the second the metric (AIC) with auditory induction constraint was employed.

Results of these experiments are shown in figure 3. They show a clear benefit of using the AIC metric as the probability of deletion increases (using a model of auditory scene analysis the probability of deletion is likely to be around 85%).

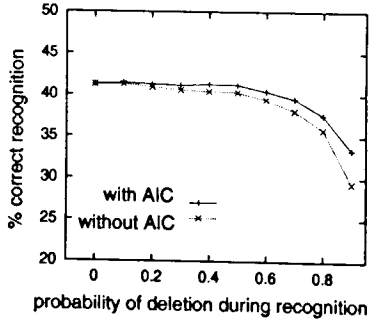


Figure 3. Recognition accuracy vs. probability of deletion for 2 recognition algorithms. The "without AIC" curve employs the distance measure for incomplete vectors, whilst "with AIC" includes modifications suggested by the continuity effect as described in the text.

IMPLICATIONS FOR MODELS OF SPEECH PERCEPTION AND FUTURE RESEARCH

These results point to a model of speech perception in which information grouped by auditory scene analysis triggers matches against stored speech schemas, which may then be verified through application of the auditory induction constraint.

This further suggests that the representations the auditory system uses to "categorise", and indeed learn about, speech sounds may be rather different to those with which the traditional phonetician or linguist familiar. If we restrict ourselves to consideration of spectrum-like representations (there is no reason why other representations such as onset, offset and frequency-transition maps should not play a part in the coding of speech), this is especially true in low-frequency regions of speech, where the first formant, and in some cases the second, and even third formants, are resolved into harmonics. When this has been addressed (cf. [8]), it has generally posed a problem for theories of speech perception, as well as automatic speech recognition systems.

A representation which includes harmonics will clearly be (even) more variable than one in which the formants are coherent, due to natural changes in fundamental frequency during the course of phonation. A model of perception based on partial matching suggests that we should regard grouped auditory primitives as indicating those frequencies at which the spectrum should be sampled, rather than as defining a pattern to be matched. This side-steps the problems associated with smoothing reconstructed spectral profiles on the basis of extracted peaks using a process of interpolation. Furthermore, a partial matching scheme might account for the centre of gravity effect, or large-scale spectral integration (cf. [8]). It gives a plausible mechanism by which the auditory system would take into account gross spectral shapes rather than individual formant frequencies. Consider for example the weights in nets *trained* (cf. [4]) using complete data vectors and data vectors with 85% of the components randomly deleted, shown in figure 4. It can be clearly seen that the deletions have little effect on the representation of the vowel spectra.

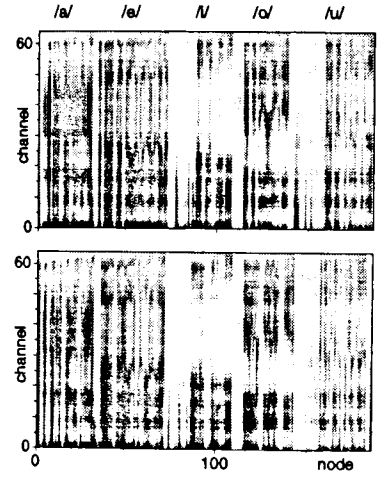


Figure 4. Spectrogram-like plots of net weight vectors (1 frame = 1 node) sorted by label (within-label order is insignificant) for nets trained using (top) complete data vectors and (bottom) data vectors with 85% of the components randomly deleted.

The model presented here may also explain aspects of the Lombard effect (cf. [9]): since harmonicity is a powerful grouping cue, and in noisy conditions likely to be more robust than common amplitude modulation, a goal of the elevation of pitch may be to make the voice more readily separated. Furthermore, studies of auditory induction [3] also show that when one sound is heard in the presence of another, the perceived loudness of the first sound is less than it would be were it heard in isolation. In other words the auditory system employs some sort of disjoint allocation of energy between the two sounds. The purpose of changes in spectral tilt observed in Lombard speech may be to raise the perceived level of high frequency regions of the spectrum to counteract the fact that otherwise, after disjoint allocation, the perceived levels would be below those expected for a particular speech sound.

Finally, the model suggests a novel methodology for investigating phonetic cues, and cue trading: it makes it possible to investigate computationally the question "can recognition be achieved without information in this region".

REFERENCES

[1] A.S. Bregman (1990), *Auditory Scene Analysis*, MIT Press.

[2] G.J. Brown & M.P. Cooke (1994), "Computational auditory scene analysis", *Computer Speech & Language*, **8**, 297-336.

[3] R.M. Warren, J.A. Bashford Jr., E.W. Healy & B.S. Brubaker (1994), "Auditory induction: Reciprocal changes in alternating sounds", *Perception & Psychophysics*, **55** (3), 313-322.

[4] M.P. Cooke, P.D. Green & M.D. Crawford (1994), "Handling missing data in speech recognition", *International Conference on Speech and Language Processing*, Yokohama.

[5] T.E. Kohonen (1984), *Self-Organisation and Associative Memory*, Springer.

[6] T. Samad & S.A. Harp (1992), "Self-organisation with partial data," *Network*, **3**, 205-212.

[7] A. Kurematsu *et al.* (1990), "ATR Japanese speech database as a tool of speech recognition and synthesis", *Speech Communication*, **9**, 357-363.

[8] A. Bladon, (1986), "Phonetics for hearers", in: *Language for Hearers*, ed: G. McGregor, Pergamon Press.

[9] J-c. Junqua (1992), "The variability of speech produced in noise", *Proc. ESCA Workshop on Speech Processing in Adverse Conditions*, Cannes, 187-190.

ACKNOWLEDGEMENTS

This work was supported by a study visit grant to ATR, Kyoto to Malcolm Crawford, and by SERC Image Interpretation Initiative Research Grant GR/H53174. Kohonen net simulations adapted the public domain SOM_PAK code (Kohonen, Kangas and Laaksonen, 1992).