This is a repository copy of *A speech recognition strategy based on making acoustic evidence and phonetic knowledge explicit*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/200769/

Version: Published Version

Proceedings Paper:
Green, P.D. orcid.org/0000-0001-9103-7287, Cooke, M.P., Lafferty, H.H. et al. (1 more author) (1987) A speech recognition strategy based on making acoustic evidence and phonetic knowledge explicit. In: Proceedings of European Conference on Speech Technology. European Conference on Speech Technology, ECST 1987, 01-04 Sep 1987, Edinburgh, Scotland, UK. International Speech Communication Association , pp. 1373-1376.

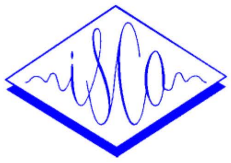
© 1987 ISCA. Reproduced in accordance with the publisher's self-archiving policy.

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

ISCA Archive
http://www.isca-speech.org/archive



European Conference on
Speech Technology
Edinburgh, Scotland, UK
September 1987

A SPEECH RECOGNITION STRATEGY BASED ON MAKING ACOUSTIC EVIDENCE AND PHONETIC KNOWLEDGE EXPLICIT

P D Green, M P Cooke, H H Lafferty and A J H Simons*

ABSTRACT

We describe a prototype implementation of a representational approach to acoustic-phonetics in knowledge-based speech recognition. Our scheme is based on the 'Speech Sketch', a structure which enables acoustic evidence and phonetic knowledge to be represented in similar ways, so that like can be compared with like. The process of building the Speech Sketch begins with spectrogram image processing and goes on to exploit elementary phonetic constraints. A multiscale approach is used throughout. The process of interpreting the Speech Sketch makes use of an object-oriented phonetic knowledge base. Objects in the knowledge base can be matched against objects in the Speech Sketch in a manner directed by the incoming evidence. This technique promises to avoid a combinatorial explosion.

1. INTRODUCTION

We have previously argued (ref 1) for an intermediate representation, the 'Speech Sketch' in the acoustic-phonetic signal-to-symbol transformation. The purpose of the Speech Sketch is to make acoustic parametric **behaviour**, rather than parameter **values**, explicit. The following sections outline the representational stages in a prototype acoustic-phonetics component for voiced speech based on this principle. Much of the software is implemented in SCOPE - a Small Common Object-based Programming Environment (ref 2).

2.1 SPECTROGRAM IMAGE PROCESSING TO PRODUCE SPECTRAL OBJECTS

A conventional spectrographic representation of speech is converted into a list of 'Fragments' describing the time-evolution of spectral peaks, corresponding to formants or harmonics. First, a multiscale representation (ref 3) of spectral edges is created (ref 4). Next, a probabilistic relaxation algorithm, employing a spectral continuity constraint, seeks out the most likely labels for each spectral peak, (from the set: onset, offset, noise, part of a track and continuation at a finer scale) and identifies the most likely links between peaks in adjacent time-frames and scales. Finally, a deterministic procedure integrates information across scales, driving from onsets in the coarsest scale, continuing at a finer scale where necessary (seeking always to return to a coarser level) until an offset is reached. Figure 1 illustrates fragments for the first part of an utterance "Our lawyer will allow your rule", by a male speaker.

* Department of Computer Science, University of Sheffield,
Sheffield S10 2TN, UK.

2.2 CHARACTERISING SPECTRAL OBJECTS TO PRODUCE THE RAW SPEECH SKETCH

Fragments, vectors of frequency, bandwidth and energy triples, are cartoonised to produce a multiscale piecewise linear description of frequency-variation in time. This characterisation scheme has been particularly influenced by refs 5, 6, 7. Curvature extrema in the frequency track of a fragment are used to identify candidate segmentation points. Granularity is associated with the total error of a piecewise linear approximation of the segment by several sub-segments, using the candidate points. The characterisation algorithm builds a tree of segments where the successors of a segment represent its most economical description at the next-finest granularity level. Each node in the tree instantiates the SCOPE class 'Bar'; Bars constitute the discrete representation of fragment data in the Raw Speech Sketch. Figure 2 illustrates coarse-granularity Bars.

2.3 UTILISING COMMON SENSE PHONETIC CONSTRAINTS TO PRODUCE THE FULL SPEECH SKETCH

Bars in the Raw Speech Sketch are interpreted as primitive phonetic objects, such that the most consistent global description is obtained using only weak local constraints (ref 8). An initial sub-set of labels, drawn from the set: harmonic or fine transition (H), formant-like resonances (F0...F4) and merged (unresolved) formant (F0F1, F1F2), is assigned to each bar, using gross frequency-region and bandwidth constraints. On each cycle, uniquely labelled bars broadcast to their neighbours in frequency-time, removing inconsistent labels, until no constraints remain to be propagated. This process degrades gracefully, such that any ambiguously labelled bars fairly reflect the data. Figure 3 illustrates.

2.4 A STRATEGY FOR BUILDING PHONETIC HYPOTHESES

Between the Speech Sketch (which makes **spectral objects** explicit) and hypotheses about phonetic identity, we envisage one or more stages of bottom-up hypothesis-refinement (ref 9) whose purpose is to make the **organisation of spectral objects** explicit: for instance we might form compound objects representing primitive objects which are roughly concurrent. Each object initiates a search for more complex objects that it could itself be a part of, until objects corresponding to phonemes are reached. In the prototype, this strategy is reduced to a single step. The coarse-granularity bars which have the F2 label are used to provide cues which drive the 'speech recognition' process of matching specific phonetic objects in the Speech Sketch. We distinguish between 'low F2 cues' (bars which start or end below 850 Hz), 'medium F2 cues' and 'high F2 cues', similarily defined. The low F2 cues trigger, and provide evidence for, a match for /w/, the high F2 cues for /j/, and so on.

2.5 REPRESENTING PHONETIC KNOWLEDGE

Phonetic Knowledge is expressed in a SCOPE class hierarchy with PhoneticObject at its head. The classes in the hierarchy have slots called components, each of which describes some pattern which will match against a single object in the Full Speech Sketch (the 'filler'). Further slots express, in an algebraic formalism, the properties that fillers must possess and the relationships that must hold between them.

There is a weighted scoring mechanism which judges the quality of component fillers and the quality of the whole match. This scheme is used, in the prototype, to express the naive knowledge that any VowelLikeObject should, ideally, manifest itself as an initial transition, a steady state and a final transition in each of F1, F2 and F3, roughly concurrently. The SCOPE inheritance mechanism is used to fit particular vowels and semivowels into this pattern. In addition, we can make explicit the notions of a 'merged form', for instance that in the evidence for a semivowel F0 may be merged with F1, and 'counterevidence', for instance that a characteristic of /l/ is that there should be no object in the Speech Sketch around 2kHz.

2.6 MATCHING PHONETIC OBJECTS AGAINST THE FULL SPEECH SKETCH

The result of matching is to create instances of the given PhoneticObject. Each instance represents a collection of Speech Sketch descriptors which consistently match a subset of the PhoneticObject's components. As a consequence of having the Speech Sketch to examine, the state-space of partial instantiations is gratifyingly small. It can, in effect, be fully explored, using the given evidence first and then looking, among neighbouring objects, for support or contradiction. Figure 4 shows some matching results: the objects which contributed to the best matches for /ɑ:/, /j/, /l/ and /w/ are displayed. Note that one object can contribute to several matches: there is no forced segmentation.

3. DISCUSSION

At the time of writing (May 1987), the prototype implementation is complete but has not been rigorously tested. Its performance on the few examples so far processed is exemplary: it is as if the Speech Sketch is asking to be recognised. Much work is required to put flesh on the bones. For instance, it remains to be seen whether our phonetic knowledge representation scheme is adequate, habitable to a phonetician and robust with respect to speaker variability. Our long term plans involve moving to an 'Auditory Speech Sketch' (ref 10) and deploying 'Causal Phonetic Knowledge' (ref 11).

ACKNOWLEDGEMENTS We wish to acknowledge the invaluable assistance of the Edinburgh CSTR Speech Input Project team, and PLESSEY SIWP. This work is supported by Alvey Grant ref MMI052.

REFERENCES

1. P D Green and A R Wood, Proc. ICASSP-86, paper 23.4 (1986).
2. P Jackson, A J H Simons and G S Watson, The SCOPE Manual, (in preparation), EUSIP (Edinburgh) and SPLASH (Sheffield).
3. A P Witkin, Proc ICASSP-84 paper 39A.1.
4. H C Leung and V W Zue, Proc ICASSP-86, paper 51.1 (1986).
5. M Allerhand, 'A Knowledge-based Approach to Speech Pattern Recognition', PhD Thesis, Engineering Laboratory, University of Cambridge, UK (1986).
6. D H Ballard, Comm. ACM 24, 5, p310 (1981).
7. D H Marimont, Proc AAAI-84 (1984).
8. D Waltz, p19-92 in 'The Psychology of Computer vision', ed P H Winston, McGraw-Hill NY 1975.
9. R Levinson, Proc AAAI-84, p203.
10. M P Cooke, Proc. Inst. Acous., 8, 7, p563 (1986).
11. A J H Simons, Proc. Inst. Acous., 8, 7, p499 (1986).

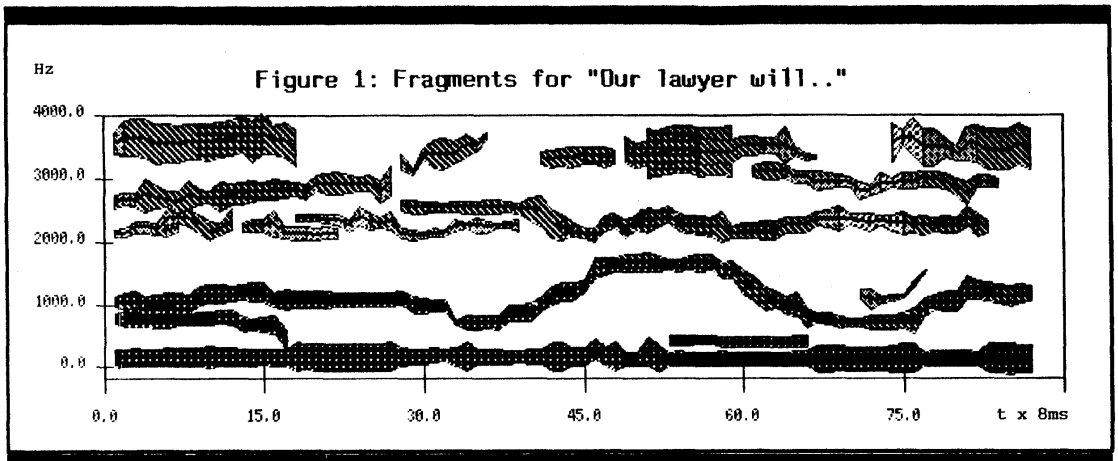


Figure 1: Fragments for "Our lawyer will.."

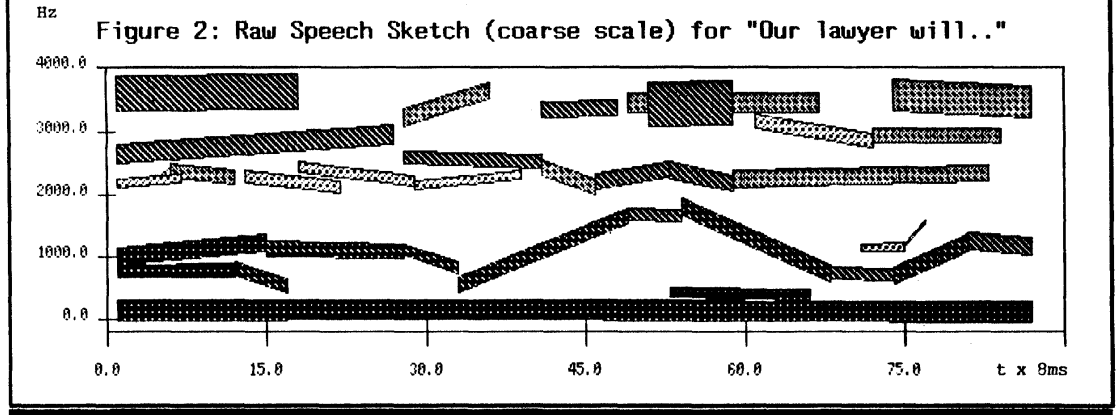


Figure 2: Raw Speech Sketch (coarse scale) for "Our lawyer will.."

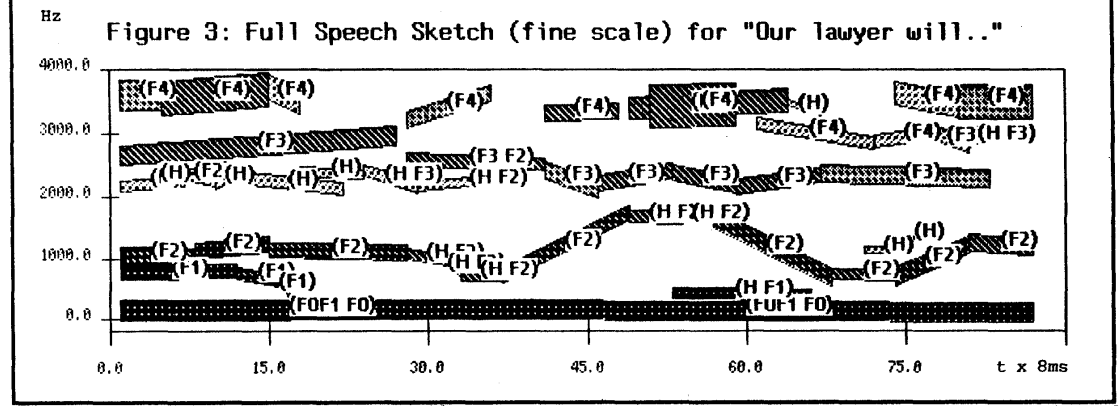


Figure 3: Full Speech Sketch (fine scale) for "Our lawyer will.."

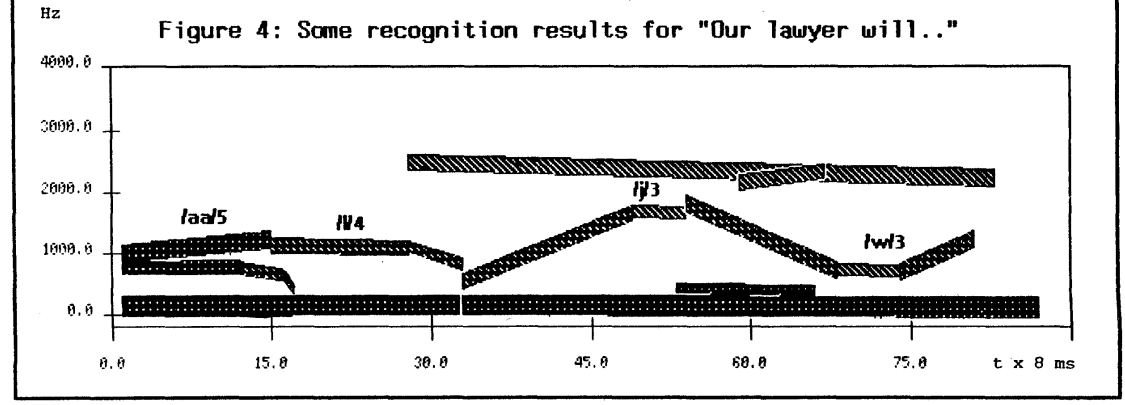


Figure 4: Some recognition results for "Our lawyer will.."