Proceedings of the Institute of Acoustics

AN INFORMATION-THEORETIC APPROACH TO MODELLING AND ASSESSMENT IN SPEECH RECOGNITION AND UNDERSTANDING IN THE SYLK PROJECT.

NR Kew (1), PD Green (2)

(1) c/o (2)
(2) Department of Computer Science, University of Sheffield.

1. INTRODUCTION

We present an information-theoretic assessment methodology for Automatic Speech Recognition (ASR) systems. We describe its application in the SYLK project, with particular reference to the measurement of Refinement performance (ie representation at multiple levels of symbolic detail). We discuss the superiority of information-based over other widely-used measures.
For present purposes, we consider an ASR to be limited to recognition of acoustic-phonetic units (APUs), although in principle these need not be conventional subword/word units, and our main discussion of ASR assessment is restricted to this level. Section 2 describes background, including the requirements of SYLK and methodologies considered. Section 3 briefly introduces the model and measures (largely summarising earlier papers), and Section 4 describes the measurement of refinement in SYLK.
In an entirely theoretical Section 5 we discuss possible wider application to modelling and evaluation of understanding in spoken language processing (SLP) systems.

2. BACKGROUND

2.1 Existing Approaches to ASR Assessment.

ASR systems are conventionally assessed by matching recogniser outputs to a correct transcription, which may be a hand-labelling, and measuring the quality of a best match. All the measures discussed below fall into this category.

2.1.1 Counts. The most widely used measures of ASR performance are those defined by the NIST (Pallett [16]), including %Correct and Accuracy applied to subwords, words or sentences. These are based on counting the proportion of units that have been correctly classified by a recogniser.
In using such measures, there is ample scope for meaningless results; for example, a system which always hypothesises every recognition unit scores 100%Correct! Accuracy is introduced to avoid this particular behaviour, but is also a simple count with no particular merit beyond that of widespread acceptance. Whilst the NIST scoring software also offers sophisticated diagnostic outputs, these are also open to the same criticism, that they take no account of the statistical behaviour of the actual data (language model). Examples are considered in §2.1.3 below.

2.1.2 Information. Rather than counting APUs, we may measure the *information* in a recogniser, expressed in terms of *entropy* (uncertainty, or disorder). Entropy-based performance measurement may be formulated in different ways:

2.1.2.1 Source Entropy (Absolute Information).
An utterance is modelled as a lattice in which all the APUs are hypothesised at every point. The APUs, and hence possible utterances, are assigned different probabilities in the lattice, determined *a priori* by the language model. Recognition is then a re-assignment of probabilities according to the acoustic evidence. The information in the recogniser is then the reduction in the entropy of the lattice, which will lie between zero and the prior entropy. Selection of *any* unique path reduces the entropy to zero, so the measure is only useful with reference to a lattice recognition, and the assignment of probabilities is of critical importance.

A variant on this approach is to make reference to the correct path in a lattice, and measure the change in its log likelihood in recognition. This is used by CSTR (McKelvie [15]).

2.1.2.2 Relative Information Transmitted (RIT). An information-theoretic measure RIT for single-path recognition is described by Smith [19]. RIT is the ratio of channel information in the recogniser to the entropy of the language data. It is easily computed, wherever the NIST measures can be computed. Smith also gives simple examples to show that RIT gives a better comparison between systems working on different alphabets than %Correct/Accuracy.

2.1.2.3 Higher-Order Measures. The power of the information-theoretic analysis enables us to define higher-order performance measures, reflecting not only relative frequencies but also context-dependencies in the actual behaviour of the data. Higher order generalisations of source entropy and RIT are given by McInnes [14] and Kew [11].

2.1.3 Discussion. To see the difference between count- and entropy-based measures, we cite a well-known analogy from written text. If every vowel is removed and replaced by a single symbol, it is reasonably easy to identify most of the words. This is because there is redundancy in the text, so that relatively little *information* is lost in spite of considerable degradation in %Correct. Although this has not (to the best of our knowledge) been quantified, a related study has been applied to SYLK (§3.3). An analysis of the (implicit) use of redundancy made by speech recognisers is given by Kew [11].
Another class of example encountered regularly in statistical classifiers (including SYLK) is that of tests biased towards the largest class, which will give unduly high %Correct scores (consider the extreme case of a classifier *always* returning the most likely outcome - giving no information)!

2.2 Requirements of the SYLK Project
SYLK (Green [5, 6], Roach [18]) is an ASR front-end adopting an unconventional approach to the symbolic representation of speech. The principal recognition unit is the syllable, for which we use a structured model based on Allerhand [1], whereby information is expressed at several levels of detail. Speech is represented in terms of syllabic *SYLK Symbols* (Roach) representing Onsets and Codas rather than phonetic units. The SYLK architecture in part follows the syllable model, and in particular involves a syllabic HMM-based first pass followed by a process of refinement, or specialisation, in which the data is described in finer detail.
This approach has two important consequences in the assessment of SYLK:-

- *It is necessary to express and compare system performance at more than one level of symbolic detail. We need a measure of refinement performance.*
- *We need some means of comparing SYLK with other systems, which typically quote results in terms of phone units such as TIMIT (Fisher [4]) or Reduced TIMIT (Lee [12]) phone labels.*

The first point concerns measurement of refinement performance in SYLK, and is discussed in §4 below. The ability to compare SYLK to other ASRs, and particularly phoneme-based systems (where both lexicon size and statistical behaviour of the language model differs from SYLK) is discussed in some detail by Kew [10].

2.3. Assessment in SYLK
The naïve count-based measures are meaningless as a means of comparing SYLK to other systems. Although they can be used to give some measure of refinement (eg by quoting %Correct figures for refinement tests, in isolation from the first pass results) this reveals little about system performance as a whole. In particular, specialisation (§4.1) must necessarily reduce %Correct, and whilst this is not an insurmountable problem it does serve further to render it unsatisfactory as a performance measure.
SYLK expresses results as a lattice, measurement of entropy in which has much to recommend it. However, the assignment of probabilities in the lattice presents a problem (HMM first-pass Viterbi recognition probabilities are useless for this purpose). In using such a measure, CSTR have devoted considerable effort to this problem, including floor probabilities in the lattice and

probability post-processing, which are indeed optimised by reference to the entropy measure (McInnes [13]). A similar approach in SYLK would be further complicated by the syllable model.

A Relative Information measure following Smith [19] does not measure lattice quality, but has the advantage of simplicity and ease of use. It has considerable functional advantages, including versatility, and it meets the particular requirements of SYLK enumerated above. Kew [10] describes developments of RIT during the course of SYLK, and further details of actual use are given below.

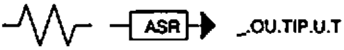
A further advantage in the use of RIT is that it may alleviate the problems caused by syllabification, in that possible *systematic* errors in the "correct" labels (Green [7]) will have relatively little effect.

3. THE INFORMATION-THEORETIC MODEL AND PERFORMANCE MEASURES

In this section, we use notation:

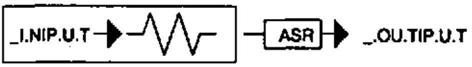
Syllabic: Onset.Peak.Coda|<next syll>
Phonetic: a/b/c/d/...

3.1 Speech Recognition as Information Channel.

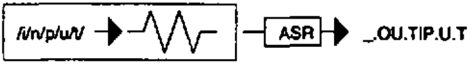


The information-theoretic model simply views an ASR as a noisy information channel. This formulation is not particularly useful, as the problem of measuring information here is equivalent to the original recognition problem!

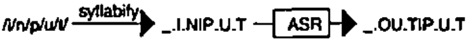
Smith [19] observes that the speech is a representation of the correct transcription, and models the "correct" path as the input, so that information may be computed directly from a confusion matrix:



In SYLK, the recognition is given in terms of syllabic *SYLK Symbols*, but the 'input' comprises phonetic TIMIT labels:



Kew [10] shows how we consider this as a compound channel comprising syllabification and recognition, and measure the latter by factoring out information loss due to the former:



3.2. Basic Performance Measure

The basic computation of entropy and RIT in given in Smith [19] or Kew[10]. In brief, the entropy H in

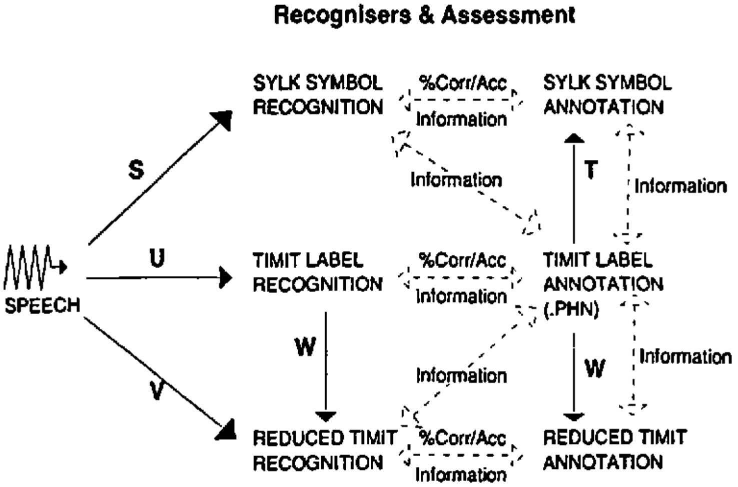


FIGURE 1 Using the model several recognisers are defined, and can be measured with reference to different 'inputs'. The Figure shows recognisers, and the Information measurements we make. Of particular interest is that of SYLK recognition against TIMIT labels, and that of SYLK labels vs. TIMIT labels against which it is normalised.

a discrete random variable X is given by:

$$H(X) = -\Sigma_{x\epsilon X}P(x)\log_2\{P(x)\} \quad (1)$$

The entropy in a relation S:[X->Y] is given by

$$H(S) = -\Sigma_{x\epsilon X, y\epsilon Y} P(x)P(y|_S x)\log_2\{P(x)P(y|_S x)\} \quad (2)$$

The mutual information in S is given by

$$MI(S) = H(X) + H(Y) - H(S) \quad (3)$$

and the relative information by

$$RIT(S) = MI(S) / H(X) \quad (4)$$

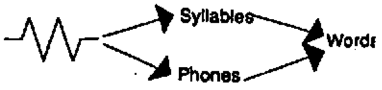
We use this with X representing TIMIT phone labels and Y SYLK symbol recognition, and compute a normalised measure by factoring out information loss in syllabification.

RIT takes values in [0,1], with 0.0 representing chance-level performance and 1.0 a reversible function. Boucher [2] shows that for several variants of a SYLK recogniser, RIT varies approximately linearly with %Correct (although this result is confined to a small range of values).

3.3 Evaluating the Syllable as Recognition Unit.

We note that all of these measures necessarily involve some compromise. Measuring SYLK recognition vs SYLK labels builds in errors of imperfect syllabification, and measuring against TIMIT labels is essentially a projection, explicitly factoring out syllabic information.

It is nevertheless possible to evaluate the syllable model, and two studies give it statistical support. The first (Kew [10b]) assumes RIT for serial information channels to be multiplicative, and combines our results with an analysis by Carter [3] of phonetic dictionary access (note that this assumption will be invalidated if it is possible for redundancy at some stage subsequently to be exploited). In spite of unfavourable assumptions, syllabic recognition was found to outperform a similar phonemic system by about 4-20% overall.



The second study (Kew [11]) analyses the effect of the syllable model structure, and finds it reduces entropy in the data by capturing redundancy.

4. MEASURING REFINEMENT IN SYLK

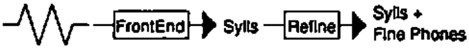
4.1 Refinement Tests

The SYLK *refinement test* (Kew [9]) is the unit of post processing. A *process* acts on a fragment of waveform or other representation, to produce a *feature vector*, which is used by a trained classifier to assign or revise symbol probabilities in a lattice via evidence combination:

Prior Output + Fragment $\xrightarrow{\text{refine}}$ Refined Output

Two distinct classes of refinement test are identified: those which aim to revise existing probabilities in a lattice, and those which make a distinction at a new level of detail (specialisation). Assessment of the first type of refinement is straightforward (and SYLK has not been able to achieve improvements using such tests). In this section, we consider specialisations.

The main case of interest in SYLK is that of fine phonetic distinctions:

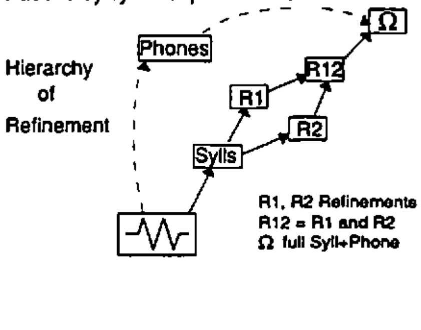


eg. D -> { b, d, g } (ie determine place of articulation in a voiced plosive Onset).

4.2 Refinement Model

The model of refinement is straightforward. A SYLK symbol is replaced by a pair, comprising itself with a specialisation comprising one or more phone labels. We use Ω to denote the set of all such pairs.

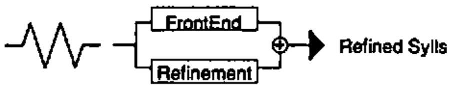
4.2.1 Refinement Hierarchy. Refinement determines a partial order on the speech representations, which is described by a lattice bounded below by syllabic detail (first-pass output) and above by syllabic+phonetic detail Ω:



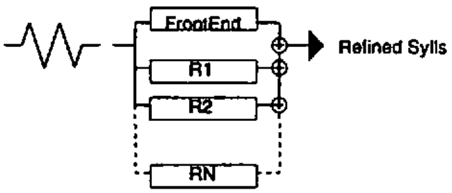
4.2.2 Orthogonal Refinement. Specialisation is orthogonal, in the sense that such refinement is independent to first pass results (and provided the new information is always added, and not considered to supersede existing information).

4.2.3 Parallel Information Channels. As a consequence of the independence of the stages of recognition in specialisation, the information channels presented by the first pass and refinements may be treated as parallel. This applies equally whether refinement is modelled as a single entity or as a series of different channels (representing for example refinement of different syllabic units) provided the refinements themselves are mutually independent.

Either



Or

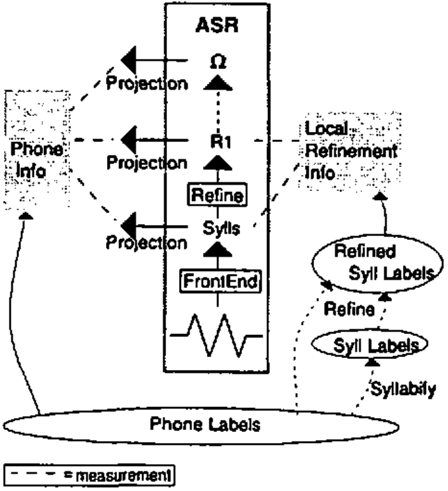


4.3 Measurement of Refinement

Measurement of the performance of a refinement test in isolation is straightforward: we may count correct outcomes, or measure information in a confusion matrix. We consider it more satisfactory to measure whole-system performance before and after refinement, and note that we are able to do so using information.

Ideally we should measure a system in terms of its own recognition units, at the most detailed level applicable. In the SYLK system described, this would imply using the full syllabic+phonetic description Ω. However, as no such full labels were available to SYLK, some modification was required. We describe two variants of the approach:

Measuring Refinement



Ideally we should measure a system in terms of its own recognition units, at the most detailed level applicable. In the SYLK system described, this would imply using the full labels Ω. However, as no such full labels were available to SYLK, some modification was required. We describe two variants of the approach:

4.3.1 Phone Information. Kew [10] measures refinement performance by measuring the phonetic information in a SYLK recognition. This projects onto a phone-unit level, which is considered appropriate to refinement tests where the new information introduced is phonetic. It makes best use of the available data (TIMIT phonetic labelling).

The drawback is that syllabic information is lost from the measurement. First-pass (syllabic) performance appears to under-perform in relation to phonetic refinement unless normalised using equation (5) above.

4.3.2 Local Refinement Information. A variant of the above is used in Green *et al* [6]. The SYLK architecture is mirrored in processing the 'correct' reference data. The phone labels are syllabified, using a rule-based syllabification program, and the phonetic information appropriate to a particular refinement re-introduced to the syllabic labels. Recognition before and after refinement is then measured in terms of the refined syllabic labels.

In traditional terms, this is arguably a more correct use of the information-theoretic model. Syllabic information is not lost in measurement so the bias is removed from front-end evaluation, although in practice the problem is merely transferred to a loss of information in imperfect syllabification. An additional drawback is that the measures are local, so the refinement hierarchy lattice structure is not preserved in the measurements. Each refinement is measured differently, so that performances of tests making different distinctions are not directly comparable.

4.4 Bounds to Refinement Performance

As noted above, RIT takes values between 0.0 and 1.0, the latter figure being achieved by a perfect recogniser. Phonetic information in a syllabic recognition is likewise bounded above by that in a perfect syllabification.

The information in a refinement test is similarly bounded above by that in a perfect refinement. We may estimate this for any given refinement by substituting correct labels for refinement outputs wherever the test is applied. This requires a strategy for cases where first-pass output is wrong so that the correct refinement may be undefined: our 'perfect' refinement leaves first-pass output unchanged in such cases.

As information cannot be lost in refinement, information in a refinement is bounded below by that prior to refinement. These bounds are used in Green [6,7] to define a percentage measure of information transmitted by an actual refinement test, given by

$$\%\text{I-trans} = 100 * \frac{\text{RefinedRIT} - \text{PriorRIT}}{\text{PerfectRIT} - \text{PriorRIT}} \qquad (6)$$

4.5 Results

It is not the purpose of this paper to report results, and the interested reader is referred to a companion paper in this volume (Green [7]) or the final report of the SYLK project (Green [6]). However, it is worth reporting here that some that, using a small number of refinement tests, we have been able to achieve some encouraging refinement performance figures, with %I-trans values often in the range 70 - 90%, and some individual tests enhancing total front-end information by up to about 21%.

4.6 Conclusions

By means of an information-theoretic description of the SYLK ASR system we are able to make meaningful measurements of the whole and any of its constituent parts. We can evaluate the performances of both the front-end and refinement tests, in isolation or in the context of the whole. We can determine what scope for improvement exists for the whole system or any given component, as well as the whole-system improvement potentially attainable by a new refinement test making any particular distinction.

As RIT results are not widely quoted, we have no figures with which to compare SYLK to other, phone-based ASR systems. Whilst a number of %Correct and Accuracy figures are given in Green [6, 7], these cannot be meaningfully compared with results quoted for other systems, for reasons discussed above. We are nevertheless able to support two main principles of SYLK: the syllable model is of value as a unit of structure (§3.3), and further information may be gained by refinement of an existing recognition (§4.5).

5. DISCUSSION AND FURTHER WORK: AN INFORMATION-ORIENTED APPROACH TO SPOKEN LANGUAGE UNDERSTANDING

5.1 Limitations on the Information Model of SYLK

The information-theoretic measurement of SYLK shares with other measures the drawback of 'tunnel vision': just as a high APU-count score may be difficult to make constructive use of in, for example, a morphological, syntactic or semantic analysis, so too the RIT measure does not ascribe a usefulness value to the information measured.

It is therefore appropriate only to the assessment of front-end ASR performance in isolation, unless coupled with a further analysis. We can only claim that information is the best measure of performance *in the recognition of a given alphabet.*

5.2 Task-Oriented Measures.

This limitation is due to the task definition and implementation. The approach of §3 may in principle be arbitrarily extended to the measurement and analysis of information transfer in any communication task, subject to an adequate mathematical description being available. The key to more general application lies in the use of task-oriented measures, and the design of interfaces between system components which are 'clean' in the sense that information is presented in terms of the task.

Front-end APU recognition is of course a valid task (though not particularly useful in isolation), so the measure discussed above is task oriented. Another task is phonetic dictionary access as analysed by Carter [3]. In estimating system performances in this task, Kew [10b] presents an argument of the form *"SYLK gives X% of syllabic information, and syllabic information gives Y% of task (word) information, so SYLK will give XY% of task information"*. It is a first-order approximation based on no usable redundancy in the syllabic recognition (contrast the unreliable zero-order approximation of applying a similar argument to %Correctl) but requires experimental investigation.

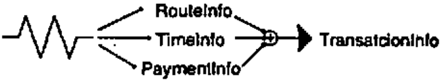
5.3 Information-Oriented Spoken Language Processing Models.

The task of spoken language understanding may be viewed as a communication channel, wherein a listener receives the information a speaker intends to convey. This is of course a difficult task, wherein the major difficulty is adequate linguistic modelling. To formulate it in terms amenable to analysis requires a suitable mathematical description of the information being conveyed.

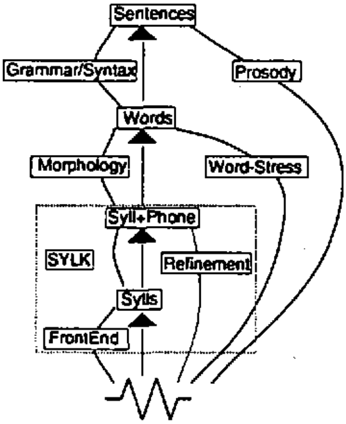
This is not within the conventional scope of information theory, and to avoid confusion with a more limited domain we will speak of an *information-oriented approach.* The change in terminology does not imply any change to basic mathematical concepts and measures.

Whilst is not clear whether and to what extent either linguistic models or statistical corpus analyses may serve this purpose, we feel that certain ideas deserve investigation:

5.3.1 Task-driven models. Within a limited domain, significant (information-bearing) ideas may be explicitly enumerated. For example, in a booking system for public transport these include times & dates, destinations and means of payment. Such cases may admit a strong information-oriented model, incorporating the different ideas as separate (parallel) channels.



5.3.2 Bottom-up approach. This involves extending the SYLK (or comparable APU-based) front-end to encompass the recognition of words, phrases and sentences. By starting from an information-oriented model, we may consider possible sources of information, and their interdependencies. A possible system architecture might be based on:



in which we have effectively moved the 'great divide' between Speech and Language outside the system portrayed. The various channels used here being precisely identified, we can use the information-based analysis to determine their exact effectiveness. This facilitates optimisation of the architecture, as well as measurement and improvement of the various subsystems.

We note the double-use of prosodic information in two separate channels: we might justify this by reference to studies distinguishing word/sentence-level stress (eg Hieronymous [8]). By building prosody into the architecture, we potentially make full use of available information, gaining a major advantage over its more typical piecemeal use as an accessory to a 'pipeline' (eg Price [17]).

5.4 Conclusions.

The information-theoretic approach to SYLK has been undertaken in a very limited timescale, but has nevertheless yielded new and rich insights into the the behaviour of the system. Likewise the assessment methodology, extended from the basic idea presented by Smith [19] has been found to be a powerful tool.

We have discussed in principle the extension of the front-end ASR to a much more comprehensive spoken language understanding system, and how the information-oriented approach might provide the requisite mathematics. We now require only the resources with which to investigate this approach, and create a prototype system.

6. REFERENCES

[1] MH Allerhand: 'Knowledge-Based Speech Recognition', *Kogan-Page* (1987).

[2] LA Boucher, NR Kew & PD Green, 'SYLK Assessment', *SYLK Working Paper #15, University of Sheffield, Dept. Computer Science* (1991).

[3] DM Carter, 'An Information-Theoretic Analysis of Phonetic Dictionary Access', *Computer Speech and Language 2, 1-11* (1987).

[4] W Fisher et al: An Acoustic-Phonetic Database, *JASA Suppl(A), 81, s92* (1987).

[5] PD Green, PJ Roach & AJH Simons 'The SYLK Project: Foundations and Overview', *Proc.IOA* (1990).

[6] PD Green, LA Boucher, NR Kew & AJH Simons 'The SYLK Project - Final Report', *University of Sheffield, Department of Computer Science, Research Report CS-92-18* (1992).

[7] PD Green, NR Kew & LA Boucher, 'Experiments with the SYLK Speech Recognition System', *Proc.IOA* (1992).

[8] JL Hieronymous, D McKelvie & FR McInnes, 'Use of Acoustic Sentence Level and Lexical Stress in HSMM Speech Recognition', *Proc.ICASSP, San Francisco* (1992).

[9] NR Kew, 'Towards a Voiceless Speech Sketch', *Proc.IOA* (1990).

[10a] NR Kew, 'Information-Theoretic Measures for SYLK Assessment', *SYLK Working Paper #16, Univ. Sheffield, Dept. of Comp. Sci.*, (1991).

[10b] NR Kew, 'An Information-Theoretic Methodology for the Assessment of Automatic Speech Recognition Systems', *submitted to Computer Speech & Language, or available from Dept. Comp. Sci., Univ. Sheffield* (1992).

[11] NR Kew, 'An Analysis of Entropy and Algorithmic Information in Phonetic Structure', *submitted to Computer Speech & Language, or from Dept. Comp. Sci., Univ. Sheffield* (1992).

[12] K-F Lee, 'Automatic Speech Recognition', Kluwer Academic Publishers (1989).

[13] FR McInnes, Y Ariki & AA Wrench, 'Enhancement and Optimisation of a Speech Recognition Front End Based on Hidden Markov Models', *Proc.Eurospeech* (1989).

[14] FR McInnes 'Context-Sensitive Phoneme Lattice Generation Using Interpolated Demi-Diphone and Triphone Models', *Proc.Eurospeech* (1991).

[15] D McKelvie & FR McInnes, 'Using Entropy as a Measure of Phoneme Lattice Quality and to Evaluate Lexical Access Mechanisms', *Proc.Eurospeech* (1989).

[16] D Pallett, 'The role of speech corpora and standards in the DARPA spoken language program', *Dublin Workshop 'Integrating Speech and Natural Language'* (1992).

[17] P Price 'Prosody in Syntactic Disambiguation', *Dublin Workshop Integrating Speech & NL* (1992).

[18] PJ Roach, D Miller, PD Green & AJH Simons, 'The SYLK Project: Syllable Structures as a Basis for Evidential Reasoning with Phonetic Knowledge', *Proc.International Congress of Phonetic Sciences* (1991).

[19] AM Smith, 'On the Use of Relative Information Transmitted (RIT) Measure for the Assessment of Performance in the Evaluation of Automated Speech Recognition (ASR) Devices', *Proc.SST* (1990).