080107
200404

ISCA ITRW, Speech Analysis and Processing for Knowledge Discovery, ISBN: 978-87-92328-00-7

Innovative acoustic probes to test predictions of wider utterance context

D. R. L. Davies[1], *J. B. Millar*[2]

[1] Information Sciences and Engineering, University of Canberra, Australia
[2] Research School of Information Sciences and Engineering, ANU, Australia

dave.davies@canberra.edu.au, bruce.millar@anu.edu.au

Abstract

Innovative measures that are targeted to specific regions of the acoustic stream of speech are described as part of a predictive speech recognition system comprising multiple dimensions. Each dimension generates its own constraint on the next stage of interpretation of an unknown utterance and together they suggest targeted questions to be asked of the acoustic stream. Acoustic probes that address distinctions between vocalic nuclei and between stop consonants are presented as illustrations of the technique. A novel parametric stability level measure providing segmentation of the acoustic stream is applied alongside more conventional measures and their relative performance is noted.

Index Terms: speech recognition, acoustic modelling, source synchronous, natural language

1. Introduction

The work described here reflects interest in the automatic speech recognition (ASR) community for new approaches that might bring the order of magnitude improvements in ASR performance necessary for automated systems to approach the skill of human listeners [e.g. 1, 2]. It acknowledges the human ability to draw on a diverse array of information ranging from shared culture and language to detailed acoustic cues.

Information theory tells us that the minimum bandwidth required for a communication act is inversely related to the degree of concordance between the relevant world models of the sender and receiver at the time of the act. We can view the act as requiring sufficient cues to trigger the desired response in the receiving system. When the communication channel is noisy or the model concordance is uncertain we need a degree of redundancy in the signal.

Applying this to a specific example: if we look at the partially recognised speech act *'She closed the X'*, where *X* represents an unrecognised segment of the utterance, how might we proceed? We do not know if we are dealing with a single word but we can recognise a possible sentence structure as 'noun verb noun-phrase'. We might look for nouns that can be linked to the verb 'closed', leading to words such as 'book', 'door' 'window' and 'account'.

Leaving aside extreme situations we will have some information in the acoustic signal that is accessible – maybe the number of syllables. If we have two syllables we can reduce our options to 'window' or 'account' as single words and 'book' or 'door' with single syllable adjectives. Probing our memory of the sound further, perhaps around the vowel sounds, we might eliminate all but 'door' and have 'tar', 'bar', 'far', 'car', 'jar' etc. as possible adjectives preceding it. Drawing on some general semantic constraints we may be able to reduce these options further.

In previous reports [3-6] we have described an approach to acoustic signal analysis designed to generate cues or hints to the phonemic structure of the signal as part of an iterative, exploratory approach to ASR. Here we report results for examples of such acoustic cues and acoustic probes designed to detect them.

2. System Overview

The system described here combines simple acoustic processing with a more general production (rule-based) system capable of combining procedural and declarative statements in a hierarchical task scheduler. Sequences of procedural commands can be executed that draw on the inference logic of the declarative rules. The production system is capable of performing high-level strategic control functions for exploratory processes at the acoustic level.

The requirements for this approach are, broadly, that low-level acoustic-phonetic cues can be combined with high-level grammatical and semantic cues in a bottom-up and top-down interchange of information iteratively refining recognition hypotheses. In addition, at each level in this hierarchy the temporal context, both backward and forward, can be mapped and iteratively revised. Throughout this vertical and horizontal analysis multiple hypotheses are accumulated along with associated likelihoods that can be fed into Bayesian decision trees – the output of which will provide a choice for the next iterative step.

Elements of the existing production system relevant to unknown utterance modelling are: (a) a dictionary of word base forms with part-of-speech, tense, number, word frequencies and that allows ambiguity and multiple phonemic representations; (b) morpheme generation from rules or from tables of irregular morphemes; (c) a parser providing generalised symbol stream pattern matching and transforms to and from canonical forms; (d) contextual specification (partially implemented) and (e) an inference engine providing logical chaining.

2.1. Constraints on Speech Recognition

A system that maximizes constraints on recognition options will seek to mine evidence from the multiple dimensions that influence the construction of a speech utterance. This paper focuses on one small area of evidence as a sample of a much wider process. In Table 1 we illustrate this evidential context and highlight the area in which we present data arising from innovative acoustic signal processing. As each area of influence is addressed the options for correct recognition (lexical perplexity) are

reduced. Ideally all constraints are evaluated in parallel providing both bottom-up and top-down limitations on recognition options. The strength of evidence is likely to build unevenly such that in certain circumstances where economy of effort is paramount the evidence in some areas may be undetectable.

Table 1. Global Picture and Area of Focus on constraints

Constraint	Evidence to be Mined
....	....
Topic	domain specific words and phrases
Dialogue	continuity between utterances
Intent	history of key terms within dialogue
Syntax	speaker's grammatical style
Lexicon	speaker's local vocabulary
Prosody	acoustic options for word hypotheses
Phonetics	**acoustic options for phonemes**
Speaker size	normalisation of static phonetics
Speaker style	normalisation of phonetic dynamics
Speaker quality	normalisation of acoustic balance
....	....

In this paper we focus on probing for evidence of patterns in speech sounds at appropriate locations in the acoustic stream. This highlighted area in Table 1 is intimately linked to proximate constraint areas and indeed, less intimately, to all. The proximate area of prosody will indicate the likely presence of utterance syllables and the area of speaker size will help normalise the options between adjacent areas of the speaker's vowel space. However key consonantal information is present in a variety of acoustic forms and it is this area that we will address in this paper.

2.2. Acoustic Modelling

The acoustic modelling package generates a wide range of acoustic parameters. We have evaluated many of the acoustic parameters that have been reported in the ASR literature and derived some novel temporal combinations. The 25 base parameters tested include: frame energy, glottal period, formant frequencies and bandwidths, formant frequency and energy ratios, nasal energy, frequency band energies and ratios. In addition, a set of Mel frequency domain energy moments that approximate cepstral coefficients are available. While, for the clean speech data used to date, the band energy based parameters do not perform as well as the formant based parameters they are retained in the expectation that future tests with noisy data may show some of them to be more robust to noise.

Parameter Similarity Length (PSL) temporal parameters for each of the base parameters listed above were also generated in addition to, and in combination with, the usual first and second order parameter time differentials. The PSL parameters (.Sim) can provide a proxy for segmental duration and are described in more detail and evaluated in [3-5]. The PSL is defined as the time over which the parameter value, or its derivatives, remain within a pre-determined range – the range being selected empirically for a particular acoustic-phonetic (AP) context.

Source synchronous (glottal epoch) framing of voiced segments [6] is used to provide optimal time and frequency domain information from the signal. The Goertzel algorithm, rather than the usual FFT, is used to facilitate the processing of variable length frames and provide flexibility in frequency sampling - e.g. linear or Mel scale. The time penalty relative to the FFT is less than 30% due to the short frame lengths.

Parameter values are scaled to bandwidths of 4 or 8 bits using a monotonic nonlinear transform to produce an approximately even distribution of values across the range for aggregated AP contexts. The transform is implemented as a 1 dimensional lookup table generated from aggregated training data.

In order to test the concepts of acoustic modelling and the AP association strengths of generated acoustic parameters we use phonemically labelled data from the ANDOSL speech data corpus [7].

A mechanism for selecting and evaluating the acoustic information generated in the acoustic processing module is still to be implemented. It is proposed to use Bayesian decision trees that can be directly applied to AP association likelihoods and are capable of the high performance necessary for real-time speech recognition. Their design and performance have been extensively documented in the literature [e.g. 8,9,10].

2.3. Acoustic-Phonetic Association

AP association strength based on absolute likelihoods or a relative discrimination probability measure is used as a measure of the AP information content of a parameter and is our prime method for evaluating acoustic evidence to support lexical prediction.

2.3.1. Association Likelihoods

AP association likelihoods are generated from a 2D table with phonemes as columns and parameter values as rows. Each cell counts the association of a particular parameter value with a particular phoneme in the training data. Normalisation of the table provides the association likelihoods. A ranked list of phoneme hypotheses can be generated from each row with absolute likelihoods providing a truncation point for each list.

Glottal epochs can be assigned a ranked phoneme list for each of a range of probes in an iterative exploration of phoneme hypotheses. Adjacent epochs can be compared for aggregation into segmental hypotheses that can then be applied to the pattern matching parser to generate word level hypotheses. This mechanism provides a 'first-cut' that can be refined as information from other, more computationally complex processes, is accumulated from sources indicated in Table 1 via the Bayesian process in an iterative application of constraints applied through frame to utterance levels.

2.3.2. Discrimination Probabilities

After an initial generation of a phoneme hypotheses contrastive methods can be employed. The discriminatory or contrastive potential of a parameter for any pair of phonemes (or phonemic classes) is assessed from the degree of overlap of parameter histograms accumulated, in 16 or 256 bins, for the parameter values associated with each phoneme label.

This discrimination potential (D), calculated for all possible decision boundaries is illustrated in Figure 1 for five duration parameters for the phonemic-class pair of long and short vowels aggregated over six speakers. Distinct peaks significantly above the chance level of

50%, that are more pronounced in the single speaker data, provide useful information for input to Bayesian decision-making.

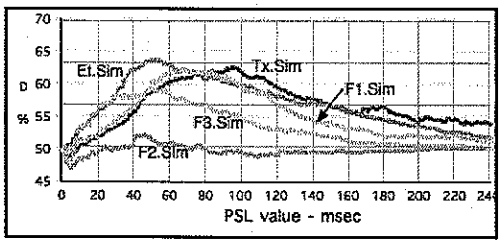

Figure 1: Six speaker aggregate discrimination potentials (D) in the Long-Short Vowel decision for 5 ".Sim" PSLs. (Tx = glottal period, Et = frame energy, F1 F2 F3 = formants).

The maximum value of D for each parameter gives us the discrimination probability (DP) for that parameter in a specific AP context as expressed (as a percentage) in Equation 1 where the H are the histogram distributions, L and S, and M is the maximum parameter value.

$$DP = Max_b\left(50\left(\frac{1}{N_L}\sum_{i=0}^{b} H_L(t) + \frac{1}{N_S}\sum_{i=b+1}^{M} H_S(t)\right)\right)$$

Equation 1: The Discrimination Probability

2.3.3. Multi Dimensional Parameter Vectors

Acoustic cues for ASR are well known to be multi-dimensional - the F1+F2 vowel triangle being a textbook example. Tabular look-up techniques used here to access AP associations are suited not just to multiple, but variable, data dimensions.

3. Evaluating Acoustic Probes

In this section, as an example of our approach to probing acoustic level information, we present results for primary cues for vocalic colour and then focus on the identification of broad stop consonant classification at the margins of vocalic nuclei.

The instantaneous status of the system will present a number of predictions regarding the range of communicative dimensions illustrated in Table 1. A forward scan of the acoustic stream will include a probe for the most likely vocalic colours predicted by lexical probabilities. A typical result of such a scan for an open back vowel is shown in Figure 2. The likelihood is generated from a 2D vector F1+F2.

3.1. Probing Vocalic Nucleus Cues

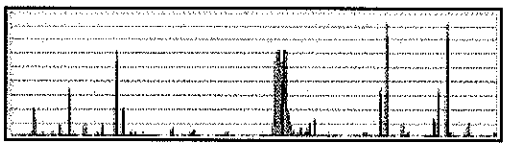

Figure 2: F1+F2 likelihoods for /ɔ/. The X axis is time for the full utterance *'The price range is smaller than any of us expected'*. The Y axis is likelihood (10% per grid line).

The broad peak coincides with the realisation of the phoneme /ɔ/ in 'smaller' and can be readily distinguished from other 'noise' peaks. The same can be done for other vowels, or the schwa that they are often reduced to in normal speech. In some cases a simple voiced-unvoiced segment decision will be all that is readily available on a first scan but at the word or utterance levels even the voicing patterns can provide valuable distinctive information as input to a Bayesian decision.

3.2. Probing Stop Consonant Cues

The classification of stops provides a good example of a diversity of acoustic cues that vary in prominence both between speakers and even, for a given speaker, within an utterance. The voicing and place decisions draw on a range of cues in both temporal and frequency domains. Their relatively brief duration has lead to a temporal spreading of cues characterised by Bell-Berti and Harris [11] as an intrinsic articulatory act that extends into preceding and following segments.

As temporal cues to stop distinction we have prior vowel durational variation in post-vocalic stops and the voice-onset timing for complete (non-terminal) stops. Segmental duration, in all AP contexts, is strongly influenced by local rate of speech, stress and by subsequent consonant clusters.

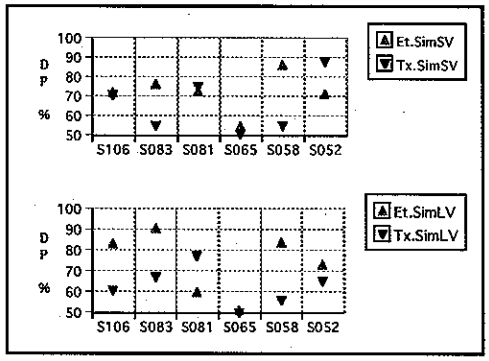

Figure 3: Discrimination probabilities for six speakers (S106 etc.) in the Voiced-Unvoiced Stop decision using two duration measures for the preceding (a) short and (b) long vowels

To access articulatory place information from the vowel duration more accurately than demonstrated here it will be necessary to model these confounding variances. It is however possible to model up to 90% of segmental durational variance [e.g. 12,13,14]. In the frequency domain we have the spectral shape of the stop burst [15,16] and formant transitions of the following vocalisation [16,17] - either or both of which may be missing.

The data in Figure 3 is derived from the peak values of distributions like those in Figure 1 yielding a discrimination probability from Equation 1. Short and long vowels are modelled separately because their durational behaviour differs. The single speaker data illustrates the high degree of inter-speaker variation. Speaker numbers are from the ANDOSL database. The first 3 are male, the last 3 female. Results such as these can be used to select the most appropriate probes for each speaker-AP combination.

Many combinations of temporal parameters combining differentials with the PSL are possible. Figure 4 shows that the use of our data-driven duration segments using the PSL mechanism that probe for consistency in value (F2.Sim), consistency in change (F2.Dif.Sim), and consistency in rate of change (F2.Dif.Dif.Sim) outperform simple time derivative based measures (F2.Dif and F2.Dif.Dif).

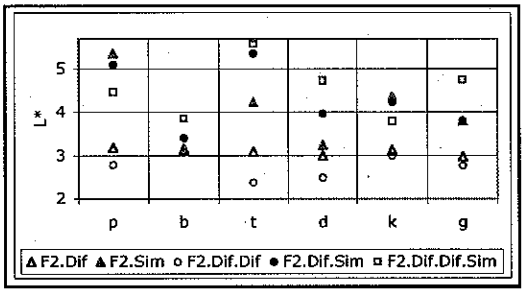

Figure 4: AP association strength for five different temporal functions of the second formant F2. The measure L* is a positive, monotonic nonlinear function of likelihoods.

The PSL F2.Dif.Dif.Sim gives the strongest value for /b/, /t/, /d/ and /g/ for our specific data. The integration of such measures within our armoury of probes designed to verify or exclude phonemic hypotheses seems to be advantageous.

Table 2: Discrimination Probabilities for the Voiced-Unvoiced stop decision using the PSL Tx.Sim for six speakers.

Speaker	S106	S083	S081	S065	S058	S058
DP%	72	69	72	68	71	68

The final example illustrated in Table 2 shows DPs for the PSLs of the glottal period (Tx) generated for all frames starting under a stop alignment label for voiced and unvoiced stops. None of the other PSLs used in Figure 2 (Et or formants) show useful DPs in this context for these speakers.

4. Conclusions

This paper has sought to illustrate the role of innovative probes of acoustic evidence to support decisions between higher order distinctions in an ASR framework. Their integration within a multidimensional predictive system shows great promise. Once all components of this system are operational, further results incorporating these probes will show whether this promise is realised.

5. Acknowledgements

The ANDOSL data corpus was prepared at the University of Sydney, the National Acoustics Laboratory, Macquarie University and the Australian National University under the auspices of the Australasian Speech Science and Technology Association, with major funding from the Australian Research Council.

Helpful comments from anonymous referees have contributed to improvements to this paper.

6. References

[1] Bourlard, H., Hermansky H., Morgan N., "Towards increasing speech recognition error rates", Speech Communication 18(3), 205-231, 1996.

[2] Moore, R.K., "Spoken language processing: piecing together the puzzle", Speech Communication 49(5), 418-435, 2007.

[3] Davies, D.R.L., Representing Time in Automated Speech Recognition, PhD thesis, 2002, Australian National University.

[4] Millar, J.B., Davies D.R.L., "A Reassessment of Temporal Information in Speech Processing", Proceedings of Workshop on Innovative Speech Processing, Stratford-on-Avon, UK, p 247-258. 2001.

[5] Davies, D., Millar, J.B., "Evaluating Representations of Segment Level Dynamics in Acoustic-Phonetic Mapping", Proceedings ICPhS'99, San Francisco, Vol. 2, p .1105-1108. 1999.

[6] Davies, D.R.L., Millar, J.B., "The evaluation of a computationally efficient method for generating a voiced-source synchronised timing signal", Proceedings 6th Australian International Conference on Speech Science and Technology, pp 527-532, 1996.

[7] Millar, J.B., Harrington, J.M., Vonwiller, J.P., "Spoken Language Resources for Australian Speech Technology", Journal of Electrical and Electronic Engineers (Australia) 17(1), 13-23, 1997.

[8] Ney, H., "Stochastic Modelling: From Pattern Classification to Speech Recognition and Translation", Proceedings International Conference on Pattern Recognition, Vol.3, p 3025-3032, 2000.

[9] Hu, G., Wang, D., "Separation of Stop Consonants", Proceedings ICASSP'03, Vol.2 p 749-752, 2003.

[10] Cesa-Bianchi, N., Conconi, A., Gentile, C., A Bayesian Framework for Hierarchical Classification, Learning Methods for Text Understanding and Mining, 2004, Grenoble, France

[11] Bell-Berti, F., Harris, K.S., "A temporal model of speech production", Phonetica, 38, 9-20, 1981.

[12] Luce, P.A., Charles-Luce, J., "Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production", JASA, 78(6), 1949-1957, 1985.

[13] Crystal, T.H., House, A.S., "Segmental durations in connected-speech signals: current results", JASA 83(4), 1553-1573, 1988.

[14] Santen, J.P.H. van, Olive J.P., "The analysis of contextual effects on segmental duration", Computer Speech and Language, 4, 359-390, 1990.

[15] Fischer, R.M., Ohde, R.N., "Spectral and durational properties of front vowels as cues to final stop-consonant voicing", JASA 88(3), 1250-1259, 1990

[16] Neagu, A., Bailly, G., "Relative contributions of noise burst and vocalic transitions to the perceptual identification of stop consonants", Proceedings EUROSPEECH-97, p 2175-2178, 1997

[17] Blumstein, S.E., Isaacs E., Mertus J., "The role of the gross spectral shape as a perceptual cue to place of articulation in initial stop consonants", JASA 72(1), 43-50, 1982.