ROBUST TECHNIQUES FOR RECOGNITION OF NEW KNOWLEDGE-BASED SPEECH PRIMITIVES

Michael Ingleby*, Wiebke Brockhaus** and Carl Chalfont*

*School of Computing & Mathematics, University of Huddersfield

**School of Music & Humanities, University of Huddersfield

ABSTRACT - A knowledge-based approach to speech recognition based on relatively recent, post-*SPE* (Chomsky & Halle, 1968), non-phonemic speech patterns is outlined. The approach emphasises the power of new phonological theories (exemplified by Government Phonology) to model the coarticulation phenomena which make continuous speech hard to recognise by machine, and proposes a set of speaker-independent features (a ***signature***) which map acoustic signal segments to the primitives of this chosen phonological theory. The features are shown to be suitable for 'coarse-to-fine' matching of a speech signal to possible parses, through invocation of a succession of cues about speaker intention imbedded in a signal.

1. INTRODUCTION

Although speaker-dependent automatic speech recognition (ASR) of isolated or weakly-connected words at lexicon sizes measured in hundreds of words has become commonplace,[1] very large vocabularies, speaker independence, and continuous speech - separately or in combination - continue to be elusive. Large vocabularies could be built up with ease - IF ONLY the words and sentences involved could be broken up into a small number of more basic segments. Speaker independence would be simply achieved, IF ONLY these segments had truly invariant acoustic signatures. Continuous speech recognition[2] would be child's play - IF ONLY the phonological processes, such as elision (*'potato'* → *'p'tato'*), assimilation (*'input'* → *'imput'*) and lenition (*'water'* → *'wa'er'*), which segments undergo in real speech production could be integrated into an ASR system. In this paper we abandon phonemes as basic segments: they are too numerous, their acoustic signatures vary in confusing ways with context, speaker and speed of articulation, and they lack phonological 'coordinates'. We argue instead for ***subsegmental*** units, called ***elements***, of which there are no more than 10, and which have phonological coordinates such as place of articulation. The phonological theory based on these elements, called Government Phonology (GP), accommodates both universal principles of spoken language and language-specific parameters (hence is applicable equally to established vocabulary, loan words, different dialects and neologisms - after setting appropriate parameters). GP is sufficiently detailed in its modelling power to codify many of the phonological processes which occur in continuous speech production. A brief outline of this phonological theory is given in Section 2 below.

Human recognition is very robust. Defects and ambiguities in a speech signal are repaired and disambiguated by a knowledgable listener, who thereby gains lexical access to speaker intentions at high success rates. We focus in Section 3 below on a variety of means for achieving comparable robustness in an automatic recogniser, treating this latter as a statistical decision-maker subject to type I (incorrect acceptance) and type II (incorrect rejection) decision errors. The decision processes involved concern the segmenting of a speech signal into a sequence and the testing of hypotheses about the GP elements within each segment of the sequence. An accepted sequence of such segment hypotheses constitutes a GP parse of the signal. When testing segment hypotheses, we operate (Section 4) with partial decision procedures driven by limited cues (manner of articulation, voicing state, and place of articulation of a segment). The decision boundaries in these procedures may be selected to minimise type II and to tolerate type I errors, with the result that the set of tolerated GP parses of a signal forms a ***segment lattice*** reminiscent of phoneme lattices (e.g. Shikano, 1980).

The cues which determine the GP composition of speech segments (see Ingleby *et al.*, 1994) are defined in terms of a standard short-time fast Fourier transform (FFT) of the acoustic speech signal. GP elements are recognised from their energy distribution, formant amplitudes and frequencies of successive short-time frames. The tracking of formants is vulnerable to signal-processing errors which arise from the fact that the peaks of an FFT amplitude spectrum do not coincide with the formant frequencies determined by a speaker's vocal tract. In pursuit of robustness we outline a means of fusing FFT amplitude and phase data to give better formant tracking than peak-picking algorithms can deliver. These techniques originated in manual form amongst engineers engaged in resonance testing (Bishop & Gladwell, 1963), and were eventually automated (Ingleby & Ronval, 1990).

2. GOVERNMENT PHONOLOGY

It is widely agreed by phonologists that the phoneme has no role to play in the study of pronunciation. For example, the standard phonemic view of nasal consonants in English is that there are three, /m/, /n/ and /ŋ/. In a word such as *'input'*, either [n] or [m] may be pronounced. In order to avoid a type II error in a recogniser expecting [ɪnput], but receiving [ɪmput], a statement to the effect that /n/ can be replaced by /m/ before /p/ would have to be included in a knowledge base accessible to the recogniser. Several similar statements covering other nasal + plosive sequences would have to appear, but they would all be manifestations of a general law: that nasals are highly likely to undergo place assimilation - that is, to share the place of articulation with a following tautosyllabic plosive. This law is an example of a GP parameter - it applies to many languages, including English, but is not universal. The phonemic approach provides no means of formalising this and other parameters and even universal principles because phonemes do not have coordinates referring to place of articulation. GP (Kaye *et al.*, 1990), like other current phonological theories which take subsegmental structure seriously, adopts the view that the phoneme is no more than 'an illusion' (Kaye, 1989) and fails to advance the understanding of either pronunciation or the sources of its variation (as outlined above).

Because ASR is concerned with pronunciation, it requires a theory of spoken language patterns which refers directly to place of articulation (and other coordinates known to participate in phonological events). The segmental primitives employed in GP (known as ***elements***) provide such a system. The most appropriate number of GP elements is still a matter for debate, but recent work (e.g. Harris & Lindsey, in press) suggests that no more than ten elements are needed. These - which are cognitive entities having both articulatory and acoustic/auditory aspects - are shown, together with their independent realisations, in Table 1. Elements, like phonemes, can be realised in isolation, but, unlike phonemes, can compound with each other to model a range of phonological segments covering the majority of known languages. For the purposes of the present example of [ɪnput] vs. [ɪmput], all that would have to be said is that a nasal may share the place element of the following plosive, in this case the labial element **U**. Other place elements are **R**, **I**, **A** and **v**, while **h**, **?** and **N** can be classified as manner elements, and **L** and **H** as source elements.

In GP, compounding of elements results in ***expressions*** in which one dominant element assumes the role of ***head***, while the remaining subordinates function as ***operator(s)***. In this paper, heads are doubly underlined. Experiments, as reported in Ingleby *et al.* (1994), have shown that compounding is equivalent to forming convex mixtures of element signatures in a feature space (described below), with the head carrying the greatest weight in the mixture. An example of an expression is the tense rounded high vowel [y] (as in the French word *'tu'* [ty] 'you', familiar form, sg.). It is composed of the elements **U** (the head, realised in isolation roughly as in English *'moon'*) and **I** (the operator, realised in isolation as in *'heed'*). Some expressions such as mixtures of **L** and **H** are universally impossible, while others, like that in *'tu'* above, are excluded in many dialects of English. An ASR system searching for signatures of individual elements in a signal can use a GP knowledge base to direct search away from impossible expressions.

Element		Articulatory		Acoustic	
		Salient property	**Unmarked properties**	**Label**	**Manifestation**
U	[u]	labial	back, high, tense ...	Falling	Fall in spectral amplitude
R	[ɾ]	coronal	tap, ...	Rising	Rise in spectral amplitude
I	[i]	palatal	non-labial, high, tense ...	Extremity	Large spectral gap
A	[ɑ]	non-high	non-labial, tense, ...	Mass	Convergence of F1 & F2
v	[ɨ]	*none*	non-labial, back, lax ...	Centrality	Central spectral activity
h	[h]	narrowed	glottal, ...	Noise	Aperiodic spectral change
ʔ	[ʔ]	occluded	glottal, ...	Glottal	Abrupt spectral change
N	[ŋ]	nasal	non-labial, back ...	Murmur	Low frequency intensity
L	L	slack vocal cords		Voicing	Fall in pitch (F0)
H	H	stiff vocal cords		Voiceless	Rise in pitch (F0)

Table 1. GP elements

A second use of a GP knowledge base is in the repair of lenition (e.g. glottalling, spirantisation or tapping). For example, London English is characterised *inter alia* by glottalling /t/ to [ʔ] in certain contexts (e.g. *'water' - 'wa'er'*). A phoneme-based recogniser would need a separate template for the recognition of [ʔ] and would then have to "convert" [ʔ] to /t/ in order to access the correct forms in its lexicon. A GP-based system, by contrast, would identify the element ʔ present in [ʔ] and, using phonological knowledge, automatically supply the additional elements h, H and R to arrive at /t/.

Both the above benefits of a GP-based approach to ASR lie at the subsegmental level. There is a third benefit to be derived at a syllabic level of language, known in GP as the ***constituent*** level. Here, the constituents of a syllable are modelled as *onset*, *nucleus* and *rhyme*. According to a universal principle (the ***Onset Licensing Principle***; Harris, 1992), every onset has to be followed by a nucleus which sanctions (or ***licenses***) the presence of that onset. The three constituents may be ***simple*** - as in the constituents of *'cat'* - or ***branching*** - as in the onset of *'trip'*, the nucleus of *'beat'* and the rhyme of *'hilt'*. Governing relations - strictly local to a constituent and having strict left-to-right (or temporally forward) directionality within constituents - effectively bar from adjacent positions certain pairs of expressions. Governing relations are captured in the ***Complexity Condition*** (e.g. Harris,1990). *Inter alia*, this stipulates that the first part of a branching onset must be more complex or compounded than the second part. Thus, an expression sequence such as [pt] cannot form a branching onset, since [t] is too complex to be governed by [p]. If a recogniser furnished with GP knowledge were to detect a tautosyllabic [pt] sequence, it would assign the [p] to an onset position. As this onset must be non-branching (due to [t] being too complex to form a branching onset with [p]), the only option is for the next position to be a nucleus, licensing the onset position to which the initial [p] is attached. Although the nuclear position is inaudible - usually as a result of rapid speech elision; e.g. in an utterance such as [pteɪtəʊ] - the recogniser would assume it to be present and undo the effects of elision, which would enable it to eventually access the word *'potato'*. Such knowledge-based repair of elision is an example of using GP knowledge to avoid type II errors - and achieve more robust segmentation of a signal affected by phonological processes.

3. ROBUSTNESS AND INVARIANCE

It has long been felt (e.g. Zue, 1985) that automatic speech recognition involves much more than pure pattern recognition, and that knowledge-based components are essential to the disentangling of speech intention from acoustic data. Information about speech intention is tangled with other information in various ways:

(i) it may be imperfectly expressed because of phonological processes such as elision, assimilation and lenition {as described in Section 2}
(ii) it is distorted by irrelevant information {relating, for example, to the speaker's vocal identity}
(iii) it is manifested through diagnostic features which are noisy {suffer statistical variation, within and between speakers}

In Section 2, the use of phonological knowledge in GP form was shown to be capable of repairing the phonological processes in (i) to yield segment lattices free of type II recognition errors. In this section, we continue earlier work on the signatures of GP elements and expressions (Chalfont, 1994). Using cluster analysis and FFT spectral data, signatures having a large degree of speaker- and context-invariance have been defined, thereby reducing (ii) above. Such signatures offer the hope of segmenting a continuous speech signal and matching it to a GP segment lattice with quite low rates of type I error (***equivocation***).

The formant frequency separation ratio and formant amplitude ratio features

$$\Phi_1 = (\omega(F2) - \omega(F1))/(\omega(F3) - \omega(F2)) \qquad \Phi_2 = A(F1)/(A(F2) + A(F3))$$

illustrate how invariance has been 'designed into' signatures. Both are dimensionless and depend on relative positioning of formants. Though formant frequencies ω(F1), ω(F2) and ω(F3) vary greatly from speaker to speaker, the frequency ratio Φ_1 remains constant for a wide variety of speakers, taking values close to 2, 1 and 1/2 during respective articulation of front/high [i], mid [ə] and back/low [ɑ] vowels. Amplitude ratio Φ_2 is high during the articulation of back/high vowel sounds such as [u], which are characterised by a decrease of amplitude with increasing frequency - and is lower for other vowels. These features together separate the vertices ([i], [ɑ] and [u]) of the well-known vowel triangle (Ladefoged, 1993), and are, therefore, good discriminators of **vowel quality**.

Vowels constitute one class of sound distinguished from others by **manner of articulation**. The six manner classes are shown as column heads in Table 2 below. If one examines the relative energy distribution (ED) between the three frequency ranges ED_{lo} in interval 0-2.0 KH, ED_{mid} in 1.5-3.5 KH and ED_{hi} in 3.0-5.0 KH, and the magnitude of the frame-to-frame fractional rate of change (ROC) of FFT amplitude averaged over all frequencies, as shown in the table, it is clear that these can be used to discriminate between manner classes. In GP terms, feature Φ_3, measuring interframe ROC, detects **h**, and the pair (Φ_4, Φ_5) measuring energy density ratios (ED_{lo}/ED_{hi}, ED_{mid}/ED_{hi}) detects ʔ. Affricates and plosives have a temporal structure, denoted by '+' in the table, with a plosive consisting of occlusion followed by plosion and an affricate composed of occlusion followed by frication.

Feature	Manner Class					
	Vowel	Plosive	Fricative	Affricate	Nasal	Approximant
ED_{lo}	H	L + L	L	L + L	H	H
ED_{mid}	H	L + L	L	L + L	L	M
ED_{hi}	L	L + L	H	L + H	L	L
ROC	L	L + M	H	L + H	L	L
Elements present	--	ʔ + h	h	ʔ + h	ʔ	--

Table 2. Manner classes and their features

Fine distinctions within the ***obstruent*** classes (fricatives, plosives and affricates) can be made with the further voicing state feature Φ_6 defined below in terms of formant trajectories. Another feature Φ_7, also concerned with formant trajectories, is used to discriminate place of articulation in obstruents, as is the ED feature Φ_4. These formant trajectory features (Φ_6 and Φ_7) are concerned with the effect of obstruents on an adjacent vowel segment. In the case of the vowel following the obstruent, the *onset* formant frequencies at the temporal boundary of the vowel (shown with subscript 'bound' below) are affected by the nearby obstruent and therefore differ from those of the steady-state frequencies; in the case of a preceding vowel, the *offset* boundary frequencies differ similarly from steady-state values.

$$\Phi_6 = F1_{bound}/F1_{steady} \qquad \Phi_7 = \{(F3_{steady} - F3_{bound}) \div (F2_{steady} - F2_{bound})\}/(F2_{steady} - F1_{steady})$$

The Φ-invariants outlined above involve formant frequency differences and frame-to-frame formant trajectory data. Such features will only determine constituent structure reliably if the formants are tracked accurately in successive frames of FFT data. Many popular formant tracking methods involve simple peak-picking algorithms which are at best accurate to within ± 100 H (e.g. Monsen & Engebretson, 1983), and historical research on resonance detection (Bishop & Gladwell, 1963) shows that peak picking is a very poor method of estimating resonant mode frequencies and amplitudes of resonant response - compared to methods of modal analysis (Ewings, 1986) which fuse amplitude and phase data. We have automated such modal analysis (Ingleby & Ronval, 1990), using a fusion technique related to the Hough transform method of shape extraction in image processing, and apply it here to formant tracking. The data fusion reduces the effects of noise on the captured signal: an indication of its success is that the estimated formant frequencies of vowels unencumbered by nearby obstruents remain constant (± 60 H) over all FFT frames from vowel onset to vowel offset, whereas frequencies estimated by peak picking are significantly more variable.

In the next section, the detection of manner classes and their refinement using the above simple Φ-invariants of an acoustic signal forms the basis of a staged approach to continuous speech recognition, beginning with autosegmentation into manner classes.

4. STAGES IN SPEECH RECOGNITION

In pursuit of robustness, a control strategy which allows early integration of linguistic knowledge into the recognition process is proposed. It is essentially a three-stage, coarse-to-fine matching of a segment lattice to acoustic data, each stage being driven by different features extracted from data, and using different phonotactic constraints from a GP knowledge base (Fig.1). In the first, coarsest stage the input string is segmented into broad classes corresponding to manner of articulation - using as cues the ED and ROC features which identify two prominent elements ʔ and h in the speech signal. The result may be conceptualised as a segment lattice whose segments are whole manner classes. At the end of this

first stage, phonotactic constraints are used to repair segmentation errors caused by elision. For example, if an utterance such as *'potato'* were segmented as *'p'tato'* at our first stage, the tautosyllabic plosive + plosive sequence would be ruled out (by a universal principle of GP). The GP recogniser would avoid violation of this principle by assuming the presence of an empty segment between the two plosives - thereby correcting a type II recognition error and a missed segment boundary.

The second stage focuses on the segmented obstruent classes, using an invariant discriminating ***voicing state***. At the end of the second stage, further phonotactic constraints can be applied. A GP parameter for English and many other languages is that there must be agreement in voicing state in tautosyllabic (in the traditional sense) plosive + fricative sequences. For example, the final *'s'* in *'dogs'* must be voiced to match the *'g'* while the *'s'* in *'cats'* remains unvoiced to match the *'t'*. An utterance such as [spɪn] recognised as [zpin] because of a type II error, could be corrected using this parameter.

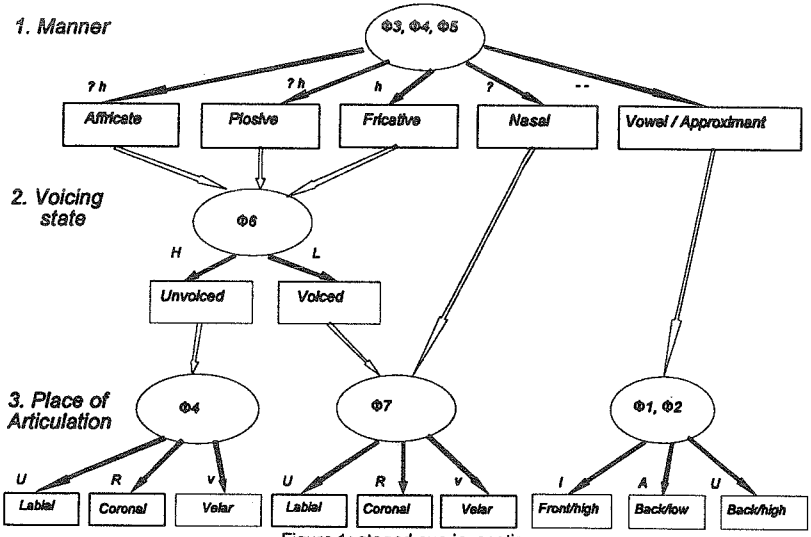

Figure 1: staged cue invocation

The third stage refines the manner and source class lattice from previous stages using invariant cues which discriminate place of articulation. As Figure 1 shows, the cues are different for different classes in the lattice, but in all cases a GP expression is matched to each segment. For example, a segment classed as ʔ h (plosive) in Stage 1 and H (unvoiced) in Stage 2, could be classed as U (labial) in Stage 3 after examination of its nearby vowel segment. This would lead to the hypothesis of expression [p]. Prior to lexical access, further phonotactic constraints can be applied at this finest stage but for brevity's sake we omit such details.

We have illustrated how the early integration of phonological knowledge into (error-prone) pattern recognition can add robustness to ASR, particularly with regard to the autosegmentation of continuous speech. Our work is still experimental, but we are sure that our staged approach gives segment lattices of low equivocation - a *sine qua non* for large-vocabulary lexical access. We do not rule out the use of traditional pattern-matching techniques such as DTW and HMM, but feel these should be used to choose between rival GP parses which remain *after* full use of phonological knowledge in the above stages.