A concurrent approach to the automatic extraction of subsegmental primes and phonological constituents from speech

Michael INGLEBY
School of Computing and Mathematics,
University of Huddersfield, Queensgate,
Huddersfield HD1 3DH, UK
M.Ingleby@hud.ac.uk

Wiebke BROCKHAUS
Department of German,
University of Manchester,
Oxford Rd, Manchester M13 9PL,
UK
Wiebke.Brockhaus@man.ac.uk

Abstract

We demonstrate the feasibility of using unary primes in speech-driven language processing. Proponents of Government Phonology (one of several phonological frameworks in which speech segments are represented as combinations of relatively few subsegmental primes) claim that primes are acoustically realisable. This claim is examined critically searching out signatures for primes in multi-speaker speech signal data. In response to a wide variation in the ease of detection of primes, it is proposed that the computational approach to phonology-based, speech-driven software should be organised in stages. After each stage, computational processes like segmentation and lexical access can be launched to run concurrently with later stages of prime detection.

Introduction and overview

In § 1, the subsegmental primes and phonological constituents used in Government Phonology (GP) are described, and the acoustic realisability claims which make GP primes seem particularly attractive to developers of speech-driven software are summarised. We then outline an approach to defining identification signatures for primes (§ 2). Our approach is based on cluster analysis using a set of acoustic cues chosen to reflect familiar events in spectrograms: plosion, frication, excitation, resonance... We note that cues indicating manner of articulation, which change abruptly at segment boundaries, are computationally simple, while those for voicing state and resonance quality are complex and calculable only after signal segmentation. Also, the regions of cue space where the primes cluster (and which serve as their signatures) are disconnected, with separate sub-regions corresponding to the occurrence of a prime in nuclear or non-nuclear segmental positions.

A further complication is that GP primes combine *asymmetrically* in segments: one prime - the HEAD - of the combination being more dominant, while the other element(s) - the OPERATORS(S) - tend to be recessive. This is handled by establishing in cue space a central location and within-cluster variance for each prime. The training sample needed for this consists of segments in which the prime suffers modification only by minimal combination with others, i.e on its own, or with as few other primes as possible. Then, when a segment containing the prime in less than minimal combination is presented for identification, its location in cue space lies within a restricted number of units of within-cluster variance of the central location of the prime cluster. The number of such distance units determines headedness in the segment, with separate thresholds for occurrence as head and as operator.

In § 3 we describe in more detail the stagewise procedure for identifying *via* quadratic discriminants the primes present in segments. At each stage, we detail the computational processes which are driven by the partial identification achieved by the end of the stage. The processes include segmentation, selection of lexical cohort by manner class, detection of constituent structure, detection and repair of the effects of phonological processes on the speech signal. The prototype, speaker-independent, isolated-word automatic speech recognition (ASR) system is described in § 4. Called 'PhonMaster', it is

implemented in C++ using objects which perform separate stages of lexical access and process repair concurrently.

1 Phonological primes and constituents

Much of the phonological research work of the past twenty years has focussed on phonological representations: on the make-up of individual segments and on the prosodic hierarchy binding skeletal positions together.

Some researchers (e.g. Anderson and Ewen 1987 and Kaye *et al.* 1985) have proposed a small set of subsegmental primes which may occur in isolation but can also be compounded to model the many phonologically significant sounds of the world's languages. To give an example, in one version of GP (see Brockhaus *et al.* 1996), nine primes or ELEMENTS are recognised, viz. the manner elements **h** (noise) and ? (occlusion), the source elements **H** (voicelessness), **L** (non-spontaneous voicing) and **N** (nasality), and the resonance elements **A** (low), **I** (palatal), **U** (labial) and **R** (coronal). These elements are phonologically active - they can spread to neighbouring segments, be lenited etc..

The skeletal positions to which elements may be attached (alone or in combination) enter into asymmetric binary relations with each other, so-called GOVERNING relations. A CONSTITUENT is defined as an ordered pair, governor first on the left and governee second on the right. Words are composed of well-formed sequences of constituents. Which skeletal positions may enter into governing relations with each other is mainly determined by the elements which occupy a particular skeletal slot, so elemental make-up is an important factor in the construction of phonological constituents.

GP proponents have claimed that elements, which were originally described in articulatory terms, have audible acoustic identities. As we shall see in § 2, it is possible to define the acoustic signatures of individual elements, so that the presence of an element can be detected by analysis of the speech signal.

Picking out elements from the signal is much more straightforward than identifying phonemes. Firstly, elements are subject to less variation due the contextual effects (e.g. place assimilation) of preceding and following segments than phonemes. Secondly, elements are much smaller in number than phonemes (nine elements compared to c. 44 phonemes in English) and, thirdly, elements, unlike phonemes, have been shown to participate in the kind of phonological processes which lead to variation in pronunciation (see references in Harris 1994). Fourthly, although there is much variation of phoneme inventory from language to language, the element inventory is universal.

These four characteristics of its elements, plus the availability of reliable element detection, make a phonological framework such as GP a highly attractive basis for multi-speaker speech-driven software. This includes not only traditional ASR applications (e.g. dictation, database access), but also embraces multilingual speech input, medical (speech therapy) and teaching (computer-assisted language learning) applications.

2 Signatures of GP elements

Table 1 below details the acoustic cues used in PhonMaster. Using training data from five speakers, male and female, synthetic and real with different regional accents, these cues discriminate between the simplest speech segments containing an element in a minimal combination with others. In the case of a resonance element, say, **U**, the minimal state of combination corresponds to isolated occurrence in a vowel such as [U], as in RP English *hood* or German *Bus*.

The accuracy of cues such as those in Table 1 for discrimination of simplest speech segments has been tested by different researchers using ratios of within-class to between-class variance-covariance and dendrograms (Brockhaus *et al.* 1996, Williams 1997), as described in PhonMaster's documentation.

The cues are calculated from fast Fourier transforms (FFTs) of speech signals in terms of total amplitude or energy distribution ED across low, middle and high frequency parts of the vocal range and the angular frequencies $\omega(F)$ and amplitudes $a(F)$ of formants. The first four cues ϕ_1 to ϕ_4 are properties of a single spectral slice, and the change in these four from slice to slice is logged as ϕ_5, which peaks at segment boundaries. The duration cue ϕ_6 is segment-based, computable only after segmentation from the length in slices from boundary to boundary,

normalising this length using the JSRU database of the relative durations of segments in different manner classes (see Chalfont 97). The normalisation is a simple form of time-warping without the computational complexity of dynamic time-warping or Hidden Markov Models (HMMs).

Cue	Label	Definition
ϕ_1	*Energy ratio$_1$*	$\phi_1 \equiv ED_{lo} / ED_{hi}$
ϕ_2	*Energy ratio$_2$*	$\phi_2 \equiv ED_{mid} / ED_{hi}$
ϕ_3	*Width*	$\phi_3 \equiv (\omega(F2) - \omega(F1)) / (\omega(F3) - \omega(F2))$
ϕ_4	*Fall*	$\phi_4 \equiv a(F1) / (a(F3)+a(F2))$
ϕ_5	*Change*	If $\delta\phi = \phi_{next\text{-}slice} - \phi_{current\text{-}slice}$, $\phi_5 \equiv \delta\phi_1 + \delta\phi_2 + \delta\phi_3 + \delta\phi_4$
ϕ_6	*Duration*	ϕ_6 operates with reference to a durations database
ϕ_7	*F1 Trajectory*	$\phi_7 \equiv \omega(F1)_{bound} / \omega(F1)_{steady}$
ϕ_8	*Formant Transition*	If $\Delta\phi = \phi_{steady} - \phi_{bound}$, $\phi_8 \equiv (\Delta\omega(F_3) + \Delta\omega(F_2)) / (\omega(F_2)_{steady} - \omega(F_1)_{steady})$

Table 1. Cues used to define signatures

The other segment-based cues contrast steady-state formant values at the centre of a segment with values at entrance and exit boundary. They describe the context of a segment without going to the computational complexity of triphone HMMs (e.g. Young 1996). The PhonMaster approach is not tied to a particular set of cues, so long as the members of the set are concerned with ratios which vary much less from speaker to speaker than absolute frequencies and intensities. Nor is the approach bound to FFTs - linear predictive coding would extract energy density and formants just as well.

Signatures are defined from cues by locating in cue space cluster centres and defining a quadratic discriminant based on the variance-covariance matrix of the cluster. When elements occur in higher degrees of combination than those selected for the training sample, separate detection thresholds for distance from cluster centre are set for occurrence as head and occurrence as operator.

3 Stagewise element recognition

The detection of elements in the signal proceeds in three stages, with concurrent processes (lexical access, phonological process repair...) being launched after each stage and before the full identity of a segment has been established.

The overall architecture of the recognition task is shown in Figure 1. At Stage 1, the recogniser checks for the presence of the manner elements **h** and ?.

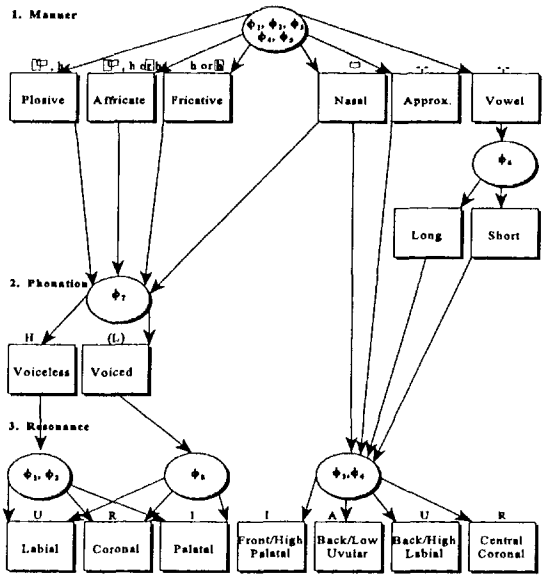

Figure 1. Stagewise cue invocation strategy

This launches the calculation of cues ϕ_5 (for the automatic segmentation process) and ϕ_6 (to distinguish vowels from approximants, and to determine vowel length). The ensuing manner class assignment process produces the classes:

- Occ Occlusion (i.e. ? present as head, as in plosives and affricates)
- Sfr Strong fricative (i.e. **h** present as head, as in [s], [z], [S] and [Z])
- Wfr Weak fricative (i.e. **h** present as operator, as in plosives and non-sibilant fricatives)

Plo	Plosion (as for Wfr, but interpreted as plosion when directly following Occ - except word-initially)
Nas	Nasal (i.e. ? present as operator)
App	Approximant
Svo	Short vowel
LVo	Long vowel or diphthong
Vow	Vowel (not readily identifiable as being either long or short).

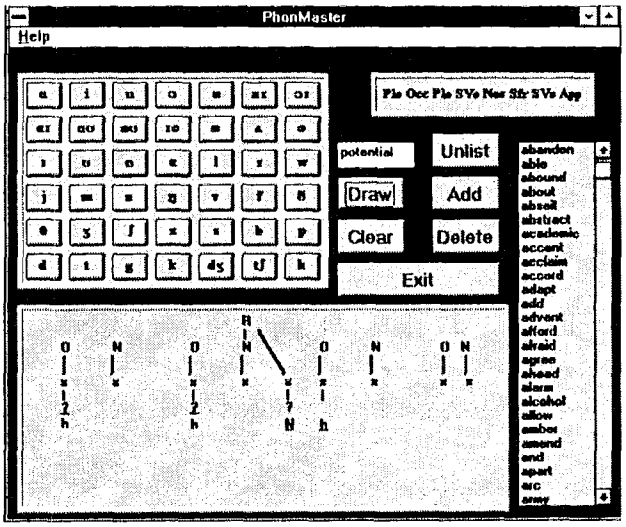

Figure 2. Representation of *potential* after Stage 1

As soon as such a sequence of manner classes becomes available, repair processes and lexical searches can be launched concurrently. The repair object refers to the constituent structure which can be built on the basis of manner-class information alone and checks its conformance to the universal principles of grammar in GP as well as to language-specific constraints. In cases of conflict with either, a new structure is created to resolve the conflict

For example, the word *potential* is often realised without a vowel between the first two consonants. This elided vowel would be restored automatically by the repair object, as illustrated in Figure 2, where a nuclear position (N) has been inserted between the two onset (O) positions occupied by the plosives.

Constituent structure is less specific than manner classes (in certain cases, different manner-class sequences are assigned the same constituent structure), so manner classes form the key for lexical access at Stage 1. Zue (1985) reports that, even in a large lexicon of c. 20, 000 words, around a third of the words can be identified uniquely by manner class alone. This is the case for languages such as English, German, French and Italian, so the accessing of an individual word may be successful as early as Stage 1, and no further data processing need be carried out.

If, however, as in Figure 3, the manner-class sequence identified is a common one, shared by several words, then the recognition process moves

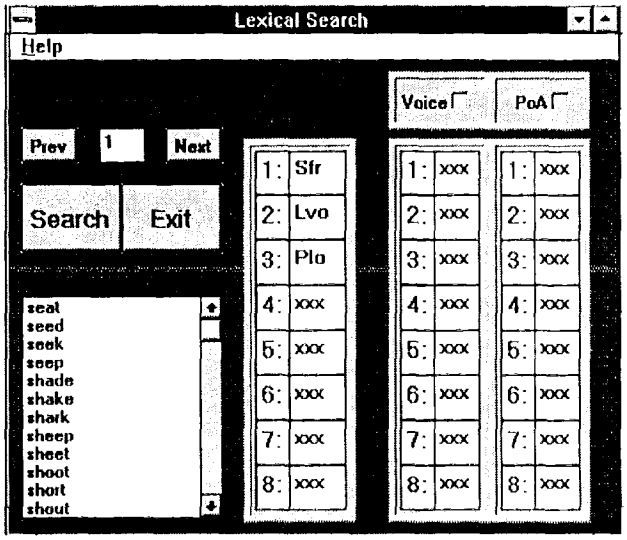

Figure 3. Lexical search screen for a common manner class sequence (Stage 1)

on to Stage 2, where the phonatory properties of the segments identified at Stage 1 are determined.

Continuing with the example in Figure 3, the lexical access object would now discard words such as *seed* or *shade*, as neither of them contains the element **H** (voicelessness in obstruents), whose presence has been detected in both the initial fricative and the final plosive at Stage 2. Again, it may be possible to identify a unique word candidate at the end of Stage 2, but if several candidates are available, recognition moves on to Stage 3.

Here, the focus is on the four resonance elements. As the manifestations of **U**, **R**, **I** and **A** vary between voiced vs. voiceless obstruents vs. sonorants, appropriate cues are invoked for each of these three broad classes (some of the cues reusing information gathered at Stage 1). The detection of certain resonance elements then provides all the necessary information for a final lexical search. In our example, only one word, *seep*, contains all the elements detected at Stages 1 to 3, as illustrated in

Figure 4. Only in cases of homophony will more than one word be accessed at Stage 3.

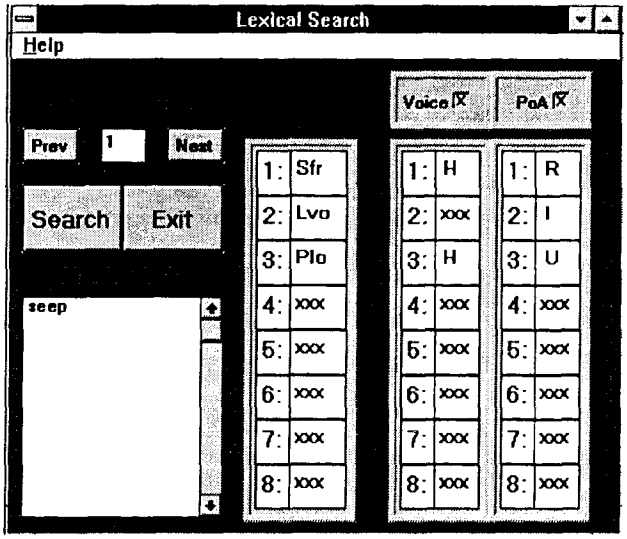

Figure 4. Lexical search screen for a common manner class sequence (Stage 3)

Concurrently with this lexical search, repair processes check for the effects of assimilation, allowing for adjacent segments (especially in clusters involving nasals and plosives) to share one or more resonance elements, thus resolving possible access problems arising from words such as *input* /'InpUt/ being realised as ['ImpUt].

4 PhonMaster and its successors

The PhonMaster prototype was implemented in C++ by a PhD student educated in object-oriented design and Windows application programming. It uses standard object-class libraries for screen management, standard relational database tools for control of the lexicon and standard code for FFT as in a spectrogram display object. Users may add words using a keypad labelled with IPA symbols. Manner class sequences and constituent structure are generated automatically. The objects concerned with the extraction of cues from spectra, segmentation, manner-class sequencing and display of constituent structure, repairing effects of lenition and assimilation are custom built.

PhonMaster does not use corpus trigram statistics (e.g. Young 1996) to disambiguate word lattices, and there is no speaker-adaptation. Without these standard ways of enhancing pure pattern-recognition accuracy, its success rate for pure word recognition is around 75%. We are contemplating the addition of pitch cues, which, with duration, would allow detection of stress, which may further increase accuracy.

Object orientation makes the task of incorporating currently popular pattern recognition methods fairly straightforward. HMMs whose hidden states have cues like ours as observables are obvious things to try. Artificial Neural Nets (ANNs) also fit into the task architecture in various places. Vector quantisation ANNs could be used to learn the best choice of thresholds for head-operator detection and discrimination. ANNs with output nodes based on our quadratic discriminants in place of the more common linear discriminants are also an option, and their output node strengths would be direct measures of presence of elements.

References

Anderson J.M. and Ewen C.J. (1987) *Principles of Dependency Phonology.* Cambridge University Press, Cambridge, England, 312 pp.

Brockhaus W.G., Ingleby M. and Chalfont C.R. (1996) Acoustic signatures of phonological primes. Internal report. Universities of Manchester and Huddersfield, England.

Chalfont C.R. (1997) University of Huddersfield PhD Dissertation 'Automatic Speech Recognition: a Government Phonology perspective'

Harris J. (1994) *English Sound Structure.* Blackwell, Oxford, England.

Kaye J.D., Lowenstamm J. and Vergnaud J.-R. (1985) *The internal structure of phonological elements: a theory of charm and government.* Phonology Yearbook, 2, pp. 305-328.

Williams G. (1997) *A pattern recognition model for the phonetic interpretation of elements.* SOAS Working Papers in Linguistics and Phonetics, 7, pp. 275-297.

Young S. (1997) A Review of Large-vocabulary Continuous-speech Recognition. IEEE Signal Processing Magazine, September Issue.

Zue V.W. (1985) *The use of speech knowledge in Automatic Speech Recognition,* Proc. ICASSP, 73/11, pp. 1602-1615.