PHONOLOGICAL ELEMENTS AS A BASIS FOR LANGUAGE-INDEPENDENT ASR

Geoff Williams
Mark Terry
Jonathan Kaye

SOAS, University of London
RMS Inc.
HAL Trust LLC

ABSTRACT

This paper proposes a novel architecture for language-independent ASR based on government phonology. The performance of phoneme-based recognisers is generally poor on languages other than the original target, which renders these systems inadequate as a platform for multi-lingual speech recognition. In this paper we pursue an alternative, non-phonemic approach, to determine whether both high accuracy and language-independence are achievable within a single system. Our approach employs a linguistic model known as Government Phonology, which we argue has clear advantages over a phone(me)-based approach. The recognition targets are a small set of sub-segmental primes, known as elements, which occur in all languages, implying that detectors can be trained once-and-for-all from a multi-lingual database of adequate size and scope. Further, the specification of well-formed structures is captured by constraints which can be relatively simply encoded as rules and applied as top-down constraints in a speech recogniser. Given a set of trained element detectors then, a recogniser for any given language can in principle be rapidly built by selection of the appropriate lexicon and constraints, given a phonological analysis of the language. We present experimental results demonstrating the problems of transferring phone-based recognisers to other languages, and describe some experimental architectures for our GP-based system.

1. INTRODUCTION

It is generally accepted that current phoneme-based ASR systems achieve high performance on large vocabularies for a single target language, but performance on other languages is invariably inferior, even where there is considerable overlap in the segmental inventory. Thus a separate set of acoustic models must be built for each target language to achieve usable performance levels, involving the construction of a large training database for each language. This seems to stem from a number of factors, most of which can be traced back to the problems in the definition of the phoneme and its limitations as a phonological unit (Section 2).

Because phonemes are defined contrastively they are inherently language-dependent, which means that a given symbol, say /b/ in English, cannot automatically be assumed to represent the same acoustic object in a different language. In fact there is a considerable phonetic difference between what is called /b/ in English and French (or Spanish, Greek etc), the respective absence versus presence of pre-voicing being the most obvious. But, these differences aside, there remains the problem of the diversity of phoneme inventories across languages, from which it follows that in order to train models for the full set of human speech sounds, data from a large number of languages would need to be collected and labelled.

Further, in order to handle the contextual variation in the acoustic realisation of phonemes within a single language, speech recognition engineers have devised ways of capturing local context-dependencies between phones, the most widely adopted being the triphone model. This technique succeeds in increasing recognition accuracy, but has the disadvantage of not only vastly increasing the number of required models, but also compounding rather than reducing the differences between the resulting sets of models needed for different languages. As a result, inherent in the current technology is a trade-off between recognition accuracy and portability between languages. In section 2 we present some results obtained with experimental small vocabulary recognisers which illustrate the point.

A possible solution to this problem is to apply a different phonological model which employs sub-segmental units typically known as features, such that any given set of segments can be built by forming appropriate compounds out of the set of features. In this type of framework, in which the vast majority of work in modern theoretical phonology is done, the segment is a derived unit rather than a prime and has a very limited role in the explanation of phonological phenomena. Feature-based systems are widely accepted as being a much more accurate model of phonological systems, and therefore of speech, than is the phoneme model. The specific phonological model we apply in this paper is that of government phonology (GP) [1,2,4], in which the atomic units are a set of seven elements which can occur either alone or in combination. There is abundant evidence that sub-segmental features are easier to recognise from the signal than phonemes [5,6,7], including our own previous work which has demonstrated this result for elements [3,7]. GP further claims that phonological variation (allowed constituent structures and the like) is limited to a small set of parameters, each having only a small range of settings (typically 2). Thus all relevant constraints can be encoded by rule. The specific advantages of GP over other autosegmental frameworks are that:

(i) constraints are parameterised such that cross-linguistic differences are trivial to state and to encode in a high-level fashion.

(ii) differences between segmental inventories are simply expressible in terms of constraints on allowed combinations of the primes.

The main challenge for this type of approach to speech recognition though is in integrating the recognition of features with the demonstrated power of HMMs in extracting the most likely utterance from the signal subject to sequencing constraints at various levels, from the task grammar right down to sub-word phonotactics. Since the HMM tools currently available are based around phone models, the most obvious initial approach is to apply a hybrid architecture whereby the outputs of element detectors are used as inputs to a segment-based HMM. Top-down constraints referring to possible simultaneous combinations of elements and phonotactics are applied at the decoding stage. We propose two different implementations of such a system in Section 4.

2. PROBLEMS WITH PHONE(ME) BASED SYSTEMS

Strictly speaking, current systems are actually based not on phonemes but on segments, or 'phones' as they tend to be known in speech recognition. Segments include the set of phonemes augmented by a subset of contextual variants ('allophones') which is usually taken from a traditional description of the phonology. In general if a phoneme has a contextual variant sufficiently different phonetically from the canonical form, it warrants having a separate phone model. So for instance in British English where an intervocalic /t/ is often realised as a glottal stop, we would want to build a model for this segment although it is not distinctive in the language. In German we would distinguish between (i.e. build separate models for) the velar fricative [x] and its realisation as a palatal fricative [ç] following front vowels. But we do not tend to build a separate model for finer phonetic distinctions, such as the realisation of /m/ as labiodental rather than labial before an [f], or to distinguish between dental and alveolar [t]. This type of variation is simply built into the basic phone models for /m/ and /t/ respectively.

A good example of why phoneme models do not transfer well between languages comes from Japanese. In Japanese, which does not distinguish /r/ and /l/ phonologically, an /r/ phoneme can be realised in a number of ways ranging from a voiced alveolar plosive [d], through [l] to something like an approximant [r] as in English. Following a strict phoneme approach all these sounds would be put into a single model, clearly one which would not perform well in other languages where the relevant distinctions were important. If this process seems rather ad-hoc and arbitrary it is no more than could be expected when using such an ill-defined unit as the segment.

To demonstrate the problems with phoneme recognisers with respect to transfer between languages, both monophone and tied-triphone isolated-word recognisers were built using the TIMIT database (dialect regions 1–8), to an accuracy of over 96% on a 20-word vocabulary. Their performance was then tested on an in-house foreign language (FL) database. The database contains single words and short phrases from a total vocabulary of around 350 words, spoken by a minimum of twenty speakers from British English, Spanish, German and Japanese in a studio environment using a close-talking microphone at 16kHz sampling rate. The recognition dictionary entries for each language were converted into appropriate Timit phone labels, and where no direct equivalent label existed, the nearest phonetic equivalent (in our judgment) was used. In Spanish for example, the following IPA → TIMIT phone mappings were used:

Vowels: a → ae; e → eh; i → ih; o→ oh; u → uh;
Consonants: x → hh; λ → y; ñ → n+y; rr → r; β → v; γ → hh

Results of the transfer tests are shown in Table 1. Figures are percent correct, using the HTK scoring system.

	Timit	British English	Spanish	German
Monophones:	96.5	70.1	71.3	79.8
Triphones:	97.7	75.4	62.0	67.75

Table 1: Transfer performance of TIMIT-trained recogniser

The most surprising point about the results is that the performance on British English shows the greatest decrement. This can be explained by the fact that the vocabularies for Spanish and German are translations of the corresponding TIMIT vocabulary, hence the difficulty of the recognition task is not identical in each case. The triphone results are more in line with our expectations, showing a performance drop of between 20–30%, with British English performing best of the three other languages. Some of the increased error will be accounted for by the different recording conditions between the two databases, however a language-related effect seems to be clearly manifested as well.

In a further experiment, a monophone recogniser was trained on the British English database and tested on passages of speech containing around 30 words spoken with short pauses. Here the recording conditions are identical, hence any difference in performance can only be due to the transferability of the phone models, plus the perplexity of the tasks as noted above. Results in percentage accuracy are shown in Table 2.

English	Spanish	German	Japanese
85.0	51.5	52.9	41.4

Table 2: Performance of English monophone recogniser on other languages

Note that a slightly different scoring scheme was used here which ignores insertions, but does not count correct silence tokens as hits either. This tends to produce lower scores than the HTK scheme, but we feel it is a more accurate reflection of the recogniser performance on this type of utterance. These results clearly demonstrate the point that phone models which work well on the training language are unusable in other languages.

3. GOVERNMENT PHONOLOGY

The GP framework posits a set of seven universal sub-segmental units, known as elements, as the fundamental components of speech sounds. Unlike the sub-segmental features proposed in other frameworks which are typically articulation-based, elements are postulated to have a direct encoding in the acoustic signal. Previous theoretical and experimental work [4,7] has provided a great deal of evidence for this hypothesis. A list of the elements and examples of the segments where they occur is given in Table 3 below.

Element	Vowels:	Consonants:
A	a,e	t,r
I	i,e,u,y	sh,ch
U	u,o	w,m,b
H	high-tones	t,s
L	low-tones	b,d (French), m,ng
?	—	p,g,l,n
N	nasal vowels	m,n,ng

Table 3: Elements in GP together with example segments

Speech segments are either empty (zero elements), or consist of single elements or compounds. For example, elements A I and U in isolation represent the 'corner' vowels [a], [i] and [u] respectively. Empty vowel expressions correspond to the high back unrounded vowel, a default vowel in many languages, and also found lexically in e.g Turkish and Russian. Mid vowels are represented as compounds of A plus either I or U, similarly the high front vowel is a compound of I and U, shown below. In compound expressions, one element is the *head* and the other(s) are *operator(s)*. Each element can occur only once, hence there is an upper limit on the number of expressions that are generated. Compounding is asymmetric, so that X.Y is not the same expression as Y.X.

Compound	IPA	Example
A.U	[o]	coat
A.I	[e]	they
U.I	[y]	tu (French), über (German)

The head position does not need to be filled, hence A._ is a valid expression. A correlation has been established between headedness and the traditional ATR feature, such that headed vowels are tense and headless vowels are lax. Headship is established on the basis of phonological analysis, not acoustic similarity, but it turns out that at least in the case of vowels the phonetic realisation of a compound is closer in feature space to the head element and so distinctions between segments which differ only in the head-operator roles can be made in practice. Languages differ with respect to the element combinations they allow in a single expression, leading to the observed differences in segmental inventories. For instance, English does not allow I and U to combine (hence the absence of the vowel ü), whereas French and German do. These combinational restrictions, known as licensing constraints, are simple statements which, under the assumption that every expression is possible unless ruled out by a constraint, generate the system of well-formed expressions in a given language.

Phonological expressions are attached to a special tier of representation known as the skeleton, which encodes sequencing information. Skeletal positions are in turn attached to constituent nodes, of which there are only 3 types: Onset, Nucleus and Rhyme. In general expressions attached to nuclear positions are vowels, those attached to the onset and rhyme are consonants.

Languages differ with respect to the number of positions they allow in each constituent, with a maximum of two. English and German allow branching nuclei (both long and short vowels), branching rhymes (closed as well as open syllables) and branching onsets (onset clusters such as br, tr, fl,...). Hence all constituents may branch in these languages. Japanese on the other hand allows no constituents to branch (the so-called long vowels are sequences of single nuclei). Spanish allows branching onsets and rhymes but no branching nuclei. A further parameter controls whether the language allows final consonants or not. Thus the syllable structure of each language is fully specified by four binary-valued parameters.

4. PROPOSED ARCHITECTURE

In previous work [3], we reported 85–90% classification accuracy for a subset of elements across a range of languages using MLP classifiers. Average transfer rates from recognisers trained on the TIMIT database to the other languages were on the order of -6%, demonstrating the language-independence of element targets. The inputs to the MLPs are combinations of spectral-based parameters. At the time of writing this transfer rate has been reduced to -4.4% by further experimentation with front-end parameters.

For word recognition we are evaluating two types of interim architecture in which the outputs of element detectors are used as inputs to a segment-based HMM. In the simplest scheme, phone models are built from features derived from the outputs of neural-net element detectors. The outputs can be either continuously valued or quantised to represent a decision as to the presence vs. absence of the element in the given frame of speech. As this model requires phonetically labelled training data in the target language, it does not allow new segments to be modelled. The intention is merely to determine whether the language-independence demonstrated for element recognition can be carried over into more robust segment models. This would be demonstrated by for example a certain level of accuracy being achieved with a smaller training set, or by improved transfer of recognition results to a new language in the experiments described above.

If a clear advantage can be demonstrated using this model, for example in terms of increased robustness or reduced training data, our next goal is to build a recogniser based purely on elements without an intermediate segmental level. In the more complex

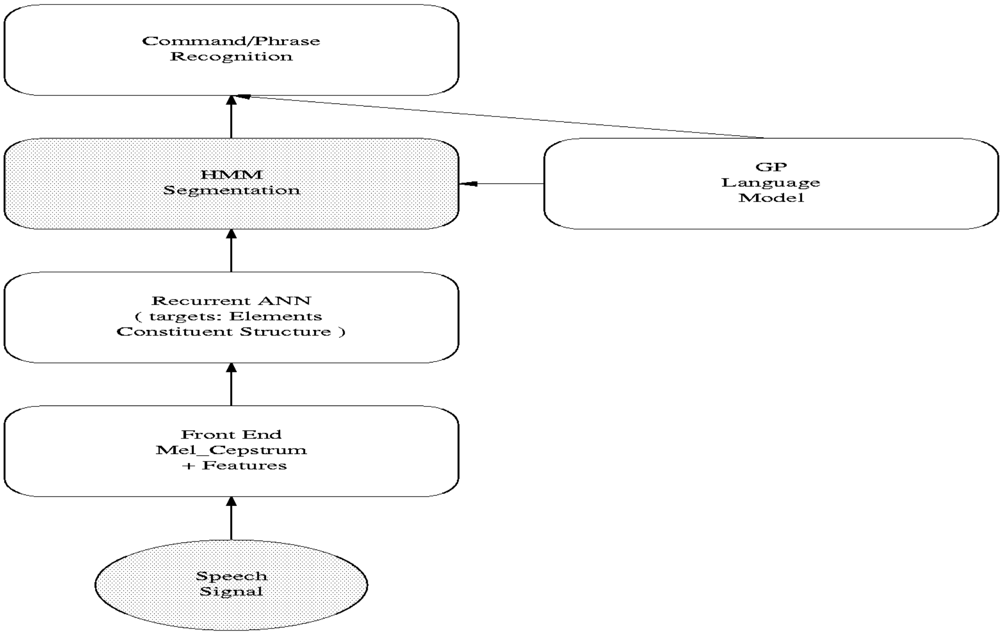

Figure 1: Hybrid Neural Net Element detection HMM segmentation-alignment

scheme, no phone models are trained at all. The speech signal is transformed into a frame-by-frame sequence of vectors each corresponding to the outputs of element detectors as before. According to the values of the sequence of vectors, supplemented by knowledge on the allowed combinations of elements in the language concerned, various hypotheses are generated as to the identity of the sequence of segments present in the speech stream. These outputs are then fed into a Viterbi decoder in order to generate the best hypothesis given the task grammar and dictionary. Constituent structure parameters as described in section 3 are applied at this decoding stage. Although only a partial implementation of a GP recogniser, this architecture has the practical advantage of allowing a standard phone-based dictionary to be used, thus keeping the other components of the recogniser within the scope of standard HMM tools. If this approach is successful then it will be possible to build a recogniser for a new language which has never been trained on any data specific to that language.

5. CURRENT STATUS

Current work is aimed at evaluating element-based phone recognisers on medium-sized vocabulary tasks in a set of diverse languages, from English to Mandarin Chinese. Once the architecture described above is in place we will be in a position to perform detailed comparisons between the current GP-based system and standard phoneme-based recognisers and provide further validation of the approach. Readers are referred to our web-site at `http://www.rmsq.com` for current recognition results in a range of languages, test vocabularies and related practical details.

6. REFERENCES

1. Kaye, J.D., Lowenstamm, J., and Vergnaud, J.-R., "Constituent structure and government in phonology," *Phonology* 7.2: 193–231, 1990.

2. J.D. Kaye, "Derivations and Interfaces," in J. Durand and F. Katamba, *Frontiers of Phonology*. Harlow: Longman 1995.

3. Kaye, J.D., Martindale, G.J., Terry, A.M., and Williams, G., "Multilingual speech recognition using phonological elements," *Proceedings of SPECOM 1996*, St. Petersburg, October 1996.

4. J. Harris and G. Lindsey, "The elements of phonological representation," in J. Durand and F. Katamba, 1995

5. K. Hübener and J. Carson-Berndsen, "Phoneme recognition using Acoustic Events," *Proceedings of ICSLP, 1994.*

6. K. Kirchhoff, "Phonetic features in speech recognition: a delayed synchronisation approach." Masters thesis, University of Bielefeld, 1995.

7. G. Williams, "The phonological basis of speech recognition." PhD Thesis, SOAS: University of London, 1998.