Development of a multi-tiered speech annotation system for
Modeling Accented English

John Ingram and Thu Nguyen
School of English, Media Studies and Art History
University of Queensland,
Australia

Abstract

This paper discusses methodological issues in the development of a multi-tiered, phonetic annotation system, intended to capture pronunciation variation in the speech of second language learners and to serve in construction of a data base for training ASR models to recognize major pronunciation variants in the assessment of accented English.

1. Introduction

Concomitant with globalization and the emergence of English as a world language has been the growth of regional varieties of English used by speakers whose native or first language may not be English. At the same time, increased workforce mobility has contributed to local diversity of spoken English in major urban centers that are linked into the global economy. Pronunciation variability has proven to be a major stumbling block for ASR systems operating in an environment where linguistically driven phonetic variability between varieties of spoken English is increasing. This is not just a problem for human-machine communication, but also for human interlocutors themselves, whether communicating locally or globally. There is therefore strong motivation to develop good linguistic and engineering models of pronunciation variability; linguistic models, so that we can better understand processes of second language adaptation and accent amelioration, that speakers of international 'Englishes' may communicate more effectively with one another; and better engineering models, so that ASR systems can cope more effectively with linguistically driven pronunciation diversity in spoken English.

Early attempts at modeling pronunciation variation in the form of connected speech processes (CSPs) relied on explicit rules encoded in expert systems (Zue, 1983). Subsequently, with the growing popularity of the hidden Markov model (HMM) framework, explicit attempts to encode linguistic rules for contextually variable phonetic forms into recognition schemes fell from favor. Context dependency could be implicitly incorporated into context-dependent acoustic models, such as the triphone model. While these statistical models initially outperformed phonetic expert systems, they have proven inadequate in the face of the extreme phonetic variability presented by accented or L2 English. Recently, hybrid approaches that attempt to narrow the search space with a combination of explicit contextual rules that acknowledge multiple sources of linguistic variation, combined with the power of context sensitive acoustic models for sources of acoustic variability that generally go undetected in phonetic transcriptions have been proposed (Hazen et al. 2002). Quite apart from any performance gains in terms of enhanced recognition rates that may be obtained by combining explicit with implicit modeling of context dependent phonetic variation, there is the important consideration that only *explicitly* identified sources of phonetic variability can be useful as source of feedback for the user interested in modifying or 'correcting' their pronunciation.

The major goals of our research are a) to develop a linguistic model of 'foreign accented' English pronunciation and pronunciation change in second language learning, and b) to develop a test of pronunciation accuracy, using ASR techniques to identify pronunciation variants and provide user feedback on pronunciation accuracy. Our general strategy is to elicit and closely phonetically annotate a substantial corpus of accented English, elicited via LanguageMAP^{TM}, a java based web browser for on-line spoken language assessment (Harrington & Ingram, 2003). The speech corpus will provide an empirical base for developing the linguistic model of accented English and for training ASR models to recognize pronunciation variants. In this report, we 1) describe in general terms the linguistic model of L2 pronunciation and pronunciation change currently under development 2) outline our methodology of pronunciation assessment, and 3) discuss options for its implementation in ASR models for accented English. A companion paper (Nguyen & Ingram, 2004) offers a detailed analysis of phonetic transfer effects and connected speech processes in a sample of 11 Vietnamese learners of Australian English and a control sample of native speakers.

Accepted after abstract only review

2. Linguistic model

The speech of English learners is characterized by a great deal of phonetic variation attributable to phonetic and phonological transfer effects from L1 and connected speech processes (CSPs) that are partly learned (language specific) and partly inherent in the biomechanics of speech production. The operation of language specific transfer effects and CSPs yields an unstable equilibrium of competing phonetic forces that resolves itself in different output states depending on the speaking style. All languages sanction departures from canonical phonological form in fundamentally similar ways that are under the ultimate control of the competing dictates of sufficient clarity and ease of articulation, but mediated by a prosodic hierarchy. The prosodic hierarchy of supra-segmental features, ranging from syllables, stressed feet or pitch accent bearing units, minor and major phrasal constituents, acts a control structure for the application of local segmental phonetic processes of assimilation, gesture overlap and lenition. Thus at the level of the syllable, consonantal deletion and lenition processes are much more prevalent in coda consonants than in onsets (Greenberg, 1999). Accented syllable nuclei resist vowel reduction, shortening or deletion. Segments that occupy left or right edges of phrasal prosodic constituents either resist or attract the application of certain CSP's.

Also, although the operation of CSPs yield multiple phonetic variants across speaking styles for any given phonological target, the array of possible realizations is quite lawfully governed in the sense that successive degrees of lenition, assimilation or loss of segmental contrasts are tightly constrained by local relations of mutual dependency and positional strength on the prosodic hierarchy. For example, [g] dropping in the phrase: *I'm **g**oing to leave* [aɪŋənəliːv] may not apply unless [t] deletion has already applied *[aɪŋən**t**əliːv]. Similarly, native speaker phonetic intuitions (falsifiable by acoustic observation) tell us that [g] deletion will not take place without leaving its calling card in terms of a velar place assimilation of the preceding nasal: *[aɪ**m**ənəliːv].

Thus, a control structure, encoded in a hierarchy of positional strength relations in the prosodic organization of the utterance regulates the application of CSP's. (This will be exemplified more fully in the oral presentation of the paper.) Furthermore, although CSPs represent natural phonological processes that 'are likely to be explicable in terms of vocal tract characteristics and the motor control mechanism, as well as being influenced by speaking rate and articulatory care' (Kerswill, 1987), they are also constrained in their application in language or dialect specific ways. Thus, Durham English has a regressive voicing assimilation rule that operates across word boundaries (*like* [g] *bairns, like* [g] *me, this* [z] *village, whats* [dz] *gone in man*), not found in other regional varieties of English (Kerswill, 1987).

There are also substantial differences across languages and dialects in prosodic structures themselves. For example, Vietnamese a strongly syllable oriented tonal language, which places a distinctive tonal marking on each syllable, may have no equivalent prosodic representation for English foot structure to act as a placeholder for vowel reduction and other lenition processes. Hence, the prosodic structure assigned by Vietnamese phonology to our English example:

Ph Ph = major phrase
(F) (F) F = foot
σ σ́ σ σ́ σ= syllable with tone
[aɪ go na lip^{ʔ}]
I gonna leave.

At least in the initial stages of L2 exposure, second language learners impose the prosodic structures of their native language on the segmental structure of the speech signal. But with prolonged exposure, presumably there is adaptation towards the prosodic structures of the target language. Herein lies a big problem. We have no direct access to speaker's phonological constructs. We may infer progressive adaptation of the underlying system of co-ordinating speech gestures to that of the target language as the learners' pronunciations take on more native-like characteristics. But how prosodic structures of a first language are modified to accommodate to those of a second language and the extent to which such modifications can take place is something we know very little about at the present time.

But in assigning phonetic annotations to speech data one is forced to make decisions or working hypotheses about the nature of phonological representations of the speaker, or to assign phonetic observations within the framework of an explicit data structure. In practice we take citation forms in Australian English as they would be represented in the Macquarie dictionary and mark phonetic departures from such standardized pronunciations using the phonetic descriptors and diacritics of the I.P.A. This means, for example, that vowel epenthesis, a frequently occurring phonetic process, motivated by an aspect of prosodic structure in L1, but something which has no phonological status in the target language (English) is represented in our hierarchical annotation as an intrusive phonetic segment with no systematic phonological status. Clearly epenthetic segments have phonetic status. Whether they should be given phonological status is debatable.

Epenthetic vowels do not encode any phonological contrasts. They merely serve to render a segment pronounceable.

3. Method

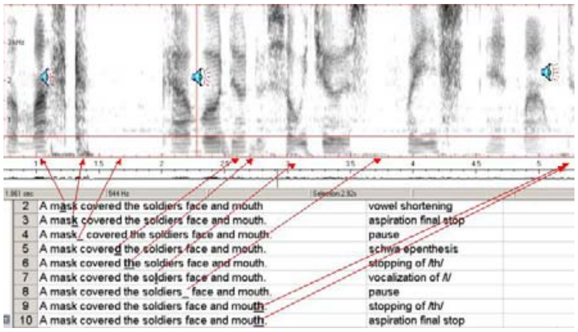
The kind of speech that we seek to elicit for pronunciation testing may be characterized as a ‘careful but unguarded style of speaking’ of the kind that subjects might use in a formal interview, where the premium on clarity of communication is high but the performance aspects of speaking itself are back-grounded.

3.1. Speech elicitation method

It is well known that pronunciations in spontaneous connected speech tend to be much more variable than in careful, read speech, where pronunciations of words are much more likely to adhere to their citation forms. Speech elicited by word reading, picture naming, sentence reading or utterance imitation is likely to underestimate the phonetic variation and prevalence of transfer effects observed in second language learners speech in naturalistic contexts. Also, explicit ‘tests of pronunciation’ encourage a degree of conscious speech monitoring which are uncharacteristic of speech in naturalistic settings. Formal pronunciation assessment should probably reflect as closely as possible the speaking conditions under which one would wish to assess an L2 speaker’s pronunciation or intelligibility; i.e., a speaking situation where there is moderate cognitive load involved and where the speaker is more preoccupied with the linguistic formulation of the message than with monitoring their pronunciation. A *grammatical paraphrase task* was found to meet these requirements and at the same time provide control over lexical selection of items known to elicit certain transfer effects.

3.2. Grammatical paraphrase task

The grammatical paraphrase task requires subjects to transform a sentence, presented in spoken and written form into a meaning-equivalent form. Subjects typed in the paraphrase in response to an initial prompt word and when satisfied with their construction, read out the sentence that they had formed. The linguistic aspects of task were sufficiently complex to engage the subjects and to deflect their attention from the pronunciation aspects of the task. This yielded quite natural sounding, careful but unguarded speech, a sample of which is presented in Fig. 1 below:



2	A mask covered the soldiers face and mouth	vowel shortening
3	A mask covered the soldiers face and mouth.	aspiration final stop
4	A mask_ covered the soldiers face and mouth.	pause
5	A mask covered the soldiers face and mouth.	schwa epenthesis
6	A mask covered the soldiers face and mouth.	stopping of /th/
7	A mask covered the soldiers face and mouth.	vocalization of /l/
8	A mask covered the soldiers_ face and mouth.	pause
9	A mask covered the soldiers face and mouth.	stopping of /th/
10	A mask covered the soldiers face and mouth.	aspiration final stop

Figure 1. Speech sample: phonetic processes

The example in Fig. 1 from a 24 year old Vietnamese student illustrates the kind of speech elicited, typically quite densely populated with phonetic processes marking non-native accent such as:

- *Vowel shortening*: Vietnamese has no distinctive vowel length contrast but shortens vowels in closed syllables.
- *Aspiration of final stop*: In Vietnamese, syllable final stops are always unreleased and glottalized. The syllable typically carries the high rising (sac) tone. A heavily transfer-induced Vietnamese English pronunciation of *mask,* would be [m**X**t]. By contrast, this is a sophisticated, but somewhat compensatory pronunciation with marked aspirated [k] release, where liaison with the following [k] of *covered* is the native expectation.
- *Pause or break index*: Insertion of a major pause break.
- *Schwa epenthesis*: to enable pronunciation of voiced stop [d] for past tense *covere**d***. A sophisticated transfer feature.
- *Stopping of ‘th’*: Vietnamese lacks a dental fricative; a dental stop is substituted.

3.3. Annotation

Phonetic processes are segmentally labeled in terms of IPA diacritic features using the EMU speech database system (Cassidy, 1999). Phonetic processes or features are annotated with reference to normal or standard Aust. English pronunciation. That is, only those phonetic features that signal departures from expected pronunciation and are likely to contribute to the perception of ‘foreign accent’ are annotated. We have made an exception in the case of CSPs which may occur frequently in native speech, but which also show a variable incidence of occurrence across native and non-native speakers alike. Annotations are entered into a multi-tiered speech database. Segmental annotations are entered in the form of three-letter codes or unicode IPA phonetic symbols on the annotation tree. Prosodic features - pitch accents, boundary tones, and pause or break indices, are entered as separate tiers on the annotation tree, using the English ToBI framework. All annotations are time aligned to the speech signal. A complete statement of the phonological environment of any phonetic feature is retrievable from the database. A complete list of phonetic and prosodic processes which are analyzed in this paper is presented in table 1 (at the end of the paper).

The companion paper (Nguyen & Ingram, this volume) presents an analysis of the incidence of occurrence of non-native phonetic processes and CSPs in the Vietnamese English speakers relative to a base rate in a control group of native English speakers. In the remainder of this paper we discuss how the data base may be deployed in training an ASR system to identify non-native pronunciation characteristics.

4. Pronunciation modeling

Most state of the art systems for ASR, use acoustic models of phone-sized segments of speech as their building blocks. Phonetic representations of words are obtained either from a pronouncing dictionary or through synthesis by rule. Such systems typically do not handle connected speech well, where words often do not conform to their citation forms. In recent years there have been numerous attempts to address both the problem posed by the operation of CSP's and phonetic variation introduced by dialect diversity.

One obvious approach is to allow alternative pronunciation variants and to train the system to accept alternative pronunciations:

$$\textit{dance} \qquad = \qquad d \quad \begin{array}{c} a \\ \\ \text{æ} \end{array} \quad n\ s$$

But as, Saraçlar et al. (2004) point out 'the degree of deviation from the canonical pronunciation varies on a continuum. Most of the time the deviation is not large enough to be clearly identifiable at the phonemic level'. There is also the problem that the greater the phonetic latitude that is allowed in the pronunciation model of a word, the higher will be its confusability with other words. From the conventional perspective of speaker independent continuous speech recognition, accommodating a broad range of non-standard or foreign accents may well pose an intractable problem.

A certain change in perspective is needed on the problem of ASR for purposes of pronunciation diagnosis. The test dialogue ensures that the speaker's choice of linguistic content is highly constrained. (Word choice, except for function words and grammatical inflections is virtually eliminated in the syntactic paraphrase task.) The speech recognizer's task is to choose correctly from among various pronunciations of a word, rather than to recognize the identity of a word through its various pronunciations. What is of interest is not *what* the speaker has said - we can be pretty confident of what he is trying to say - but *how* he has said it. We want to be able to spot significant departures from standard pronunciation and be able to inform the speaker of such departures from the standard.

The critical problem for phonetic analysis is the selection of the pronunciation variants to be modelled, by their frequency of occurrence, homogeneity of type, and approximation to the target norm. The most frequently observed non-standard pronunciation variants need to be reflected in the model. Each pronunciation variant should capture a particular type of phonetic departure from standard pronunciation. The set of variants should reflect a range of performances from 'heavily' to 'mildly' foreign accented.

We propose a separate HMM model for each test utterance, segmented into accentual phrase groups - or stress bearing units (basically lexical words plus their associated function words). These are the likely pause insertion points in Vietnamese accented utterances. Each stress bearing unit (inflected word) will be modelled by a set of HMMs of 5 -8 states, depending on its length.

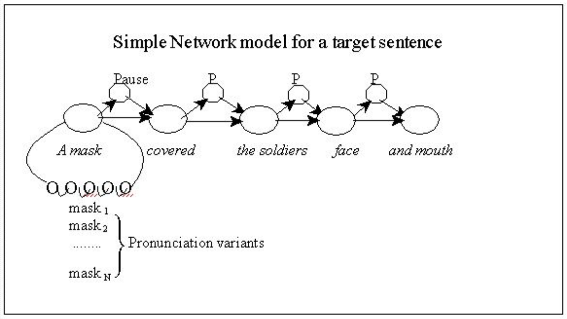


Figure 2. Basic HMM word unit model

More elaborate HMM models than this could of course be entertained. But the simple whole-word catenation model seems to be an appropriate place to begin. Success will possibly depend on how clearly categories of pronunciation variants can be identified. Statistical analysis of our annotated speech corpus, can clearly assist in this task.

However, an obvious objection to a simple concatenative HMM word model is that it fails to articulate with the acoustic state-based segmentation used in phonetic annotation of the speech data base and has no way of effectively exploiting the stochastic probabilities inherent in the hierarchy of tags which signal phonetic state changes in relation to the position of a segment in the prosodic hierarchy.

Firstly, with respect to phonetic labelling, we have previously observed that acoustic state transitions in the speech signal correspond quite well with category boundaries of a narrow phonetic transcription and that phonetic process annotation provides an explicit mapping to phonemic (segmental phonological) representations and a way of keeping track of CSP mediated segment deletions and intrusions. The SUMMIT landmark-based approach to ASR, posits an acoustic segmentation of the signal which enables one to deal explicitly with deletion and epenthesis (Seneff & Wang, 2002; Hazen, et al. 2002). Secondly, with the addition of a finite state transducer (FST) and a parsing framework developed for it (ANGIE: Senef et al. 1996), it should be possible to model the stochastic probabilities associated with states in a ‘pronunciation graph’ that models the pronunciation possibilities of a test item. The path through the network would then provide the basis for user feedback on their pronunciation of the item in question. But these are first musings only on possible approaches to the modelling of pronunciation characteristics. We would welcome comments or suggestions from those in the field of ASR with more experience than our own.

5. References

Cassidy, S. (1999) Compiling Multi-Tiered Speech Databases into the Relational Model: Experiments with the Emu System. In *Proceedings of Eurospeech '99*, Budapest, September 1999

Greenberg S. (1999) Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, **29**, 159 – 176.

Harrington M. and Ingram J. (2003) LanguageMAP^{TM} : a multi-media language assessment program, http://www.languagemap.com/

Hazen, T. J. Hetherington, I. L. Shu, H. and Livescu, K. (2002) Pronunciation modeling using a finite state transducer. *Proc. ICSLP ’02* Denver, Col. http://www.clsp.jhu.edu/pmla2002/cd/papers/hazen.pdf

Kerswill, P. E. (1987) Levels of linguistic variation in Durham. *Journal of Linguistics* **23,** 25 – 49.

Senef, R. Lau, and H. Meng, ANGIE: A new framework for speech analysis based on Morpho-phonological modeling. *Proc. ICSLP ’96*, Philadelphia, PA, vol. 1, pp. 110-113, Oct. 1996.

Seneff, R. and Wang, C. (2002) Modeling phonological rules through linguistic hierarchies. *Proc. ICSLP ’02* Denver, http://www.sls.csail.mit.edu/wangc/papers/icslp02-pmla.pdf.

Zue, V. (1983). The use of Phonetic Rules in Automatic Speech Recognition, Speech Communication 2, pp. 181 – 186, 1983.

Table 1. A list of phonetic and prosodic processes

No.	**Phonetic processes**	**Phonetic codes**
1	Vowel processes	Raised-lowered, fronted-backed,rounded-unrounded, root advanced-retracted, shortened-lengthened, centralized, reduced-strengthened, monophthongised-diphthongised, nasalized
2	Syllable structure processes	gliding of vowel, vowel epenthesis, segment deletion, Lvocalisation
3	Laryngeal processes	voiced-devoiced, prevoiced, laryngealized, breathy, creaky
4	Stop consonant processes	checked, lenis release, implosive, spirantized, initial stop aspiration final stop release
5	Fricative processes	stopping, vocalized
6	Connected speech assimilation processes	liaison, coalescence, Vowel reduction, Syllabic consonant
	Prosodic processes	Prosodic codes
1	Standard English Tones	H*, L*, L*+H, H+!H*, L+H*
2	Transfer Tones	Checked tone, sustained high tone on unstressed syllable
3	Intermediate phrases	L-, H-