SCoPE, SYLLABLE CORE AND PERIPHERY EVALUATION: AUTOMATIC SYLLABIFICATION AND APPLICATION TO FOREIGN ACCENT IDENTIFICATION

Kay Berkling

M.I.T. Lincoln
Laboratory, 244 Wood Street,
Lexington, MA 02420-9185, USA
kay@sst.ll.mit.edu

Julie Vonwiller, Chris Cleirigh

University of Sydney, Dept. of Elect. Eng.
Sydney, Australia,
julie,cleirig@speech.su.oz.au

ABSTRACT

In this paper we apply a study of the structure of the English language towards an automatic syllabification algorithm. Elements of syllable structure are defined according to both their position in the syllable and to the position of the syllable within word structure. Elements of syllable structure that only occur at morpheme boundaries or that extend for the duration of morphemes are identified as peripheral elements; those that can occur anywhere with regard to word morphology are identified as core elements. All languages potentially make a distinction between core and peripheral elements of their syllable structure, however the specific forms these structures take will vary from language to language. In addition to problems posed by differences in phoneme inventories, we expect speakers with the greatest syllable structural differences between native and foreign language to have greatest difficulty with pronunciation in the foreign language. In this paper we will analyse two accents of Australian English: Arabic whose core/periphery structure is similar to English and Vietnamese, whose structure is maximally different to English.

1. INTRODUCTION

The goal of this paper is to exploit detailed knowledge of the English syllable structure model in order to add another dimension to phoneme-based feature analysis of foreign accented speech. This application to foreign accented speech in English derives from a more general study of the syllable structure of languages. The first part of this paper is therefore devoted to the application of this study to English, followed by an analysis of foreign accents in English as a function of syllable position. Properties of accented speech are expressed in terms of phoneme substitutions, deletions or insertions as a function of syllable position. A very simple example of the importance of position is provided by German phonology. Speakers tend to devoice obstruents (stops, fricatives and affricates) at ends of words but rarely in the middle. Position independent substitution probabilities would be inaccurate for both cases. By meaningfully discriminating position of the phoneme, we can potentially improve our feature set (of phoneme substitutions) for this type of phonological variation. In this paper we will analyse two accents of Australian English: (1) Arabic whose syllable structure is relatively similar to English. (2) Vietnamese, whose syllable structure is considerably different to that of English. Section 2 will describe an automatic syllabification algorithm of a pronunciation dictionary followed by a syllable structure analysis. Section 3 will analyse the differences in pronunciation as a function of syllable position for both foreign accents.

THIS WORK WAS SUPPORTED IN PART BY TWO CONSECUTIVE POST-DOC POSITIONS AT SYDNEY UNIVERSITY AND PROF. FURUI'S LABORATORY AT TOKYO INSTITUTE OF TECHNOLOGY, AND IN PART BY THE DEPARTMENT OF THE AIR FORCE. OPINIONS, INTERPRETATIONS, CONCLUSIONS, AND RECOMMENDATIONS ARE THOSE OF THE AUTHORS AND ARE NOT NECESSARILY ENDORSED BY THE UNITED STATES AIR FORCE.

2. ENGLISH SYLLABLE STRUCTURE

Syllabification of pronunciation dictionaries is an important problem because syllable information is used for text to speech synthesis and can be an important feature in speech recognition. Most theoretical approaches to syllabification take the beginning or ending of words as their guide to the sorts of syllable structures that are allowable in a given language. In contrast, this paper takes morpheme-internal syllable structures as the basic template, and treats syllable structures specific to morpheme boundaries as exceptional, inasmuch as they carry boundary information. In order to understand the syllabification algorithm that is used in this work, we first present the model of syllable structure and the rationale that motivates it.

2.1. Syllable Constituents

A syllable usually consists of an obligatory vowel with optional surrounding consonants the exception being where a schwa-like vowel and following consonant are realised singly as a syllabic consonant. One familiar way of subdividing a syllable is into *Onset* and *Rhyme*. However, these categories alone do not indicate where the syllable

is placed within the word. We propose another additional structure of the syllable as shown in Figure 1 which distinguishes between a *Core* and *Periphery*.

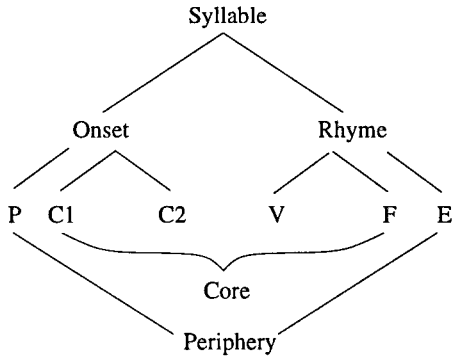

Figure 1: Constituents of a syllable as defined in this paper. (P, $C1$, $C2$, F, and E denote allowed sets of consonants. V denotes the set of vowels.)

In English, peripheral phonemes are those consonants that only occur as syllable constituents at morpheme boundaries. As such, the Periphery is a marker of morphological boundaries, and more often than not, this means word boundaries. We take the Periphery to be essentially a word-boundary phenomenon that can come to be incorporated within words historically through such processes as compounding. As an example, the word "flame" (*/fleim/*) can be broken down into the constituents as */flei/* (Core) and */m/* (Periphery), where the periphery demarcates the end of the (monomorphemic) word. Similarly, the word "lodgement" (*/lOdZm@nt/*) contains two syllables, */lOdZ/* and */m@nt/*; the first syllable has */lO/* (Core) and */dZ/* (Periphery), while the second has */m@n/* (Core) and */t/* (Periphery). Here the first Periphery */dZ/* marks the end of the first morpheme "lodge", and the second Periphery */t/* marks both the end of the second morpheme "-ment", and the end of the word "lodgement". By way of contrast, the word "freely" (*/fri:li:/*) contains two syllables, */fri:/* and */li:/*; the first syllable has */fri:/*(Core), while the second has */li:/*(Core). In this case then, although this word contains two morphemes, free and -ly, neither is demarcated by peripheral elements of syllable structure. While all languages potentially make a distinction between core and peripheral elements of their syllable structure, these structures will vary from language to language. Where English has demarcative consonants at syllable boundaries as Periphery, for tone-languages, such as Vietnamese, it is the "lexical" tone, which extends for the duration of the morpheme or word, that is analysed as the peripheral element of syllable structure. By analysing syllables in this way, we are able to identify not just differences in phoneme inventories across languages, but also differences in the ways that languages position their phonemes in syllables, and, importantly, differences in the ways that languages vary syllable structure according to the morphological location of a syllable. Comparing languages using such fine distinctions provides us with a powerful predictive tool for identifying elements of syllable structure that should prove most difficult for foreign speakers of English, and as such, a rich theoretical resource for the automated recognition of foreign accents of English.

2.2. Syllable Marking

In order to use the linguistic knowledge of syllable constituents as defined, we now want to devise an automatic method of marking syllables. Each pronunciation of a dictionary which is used by the system, will have to be split, first into syllables and then into its constituents. There are some basic rules for splitting a word into syllables. At the nucleus of any syllable is always the vowel (syllabic consonants are treated here as /@/+ consonant); long vowels and diphthongs count as a single phoneme, but occupy two syllable positions (V+F). Considering syllable structure in terms of the constituents Onset and Rhyme, the Rhyme begins with the vocalic nucleus, and anything before it in the same syllable is the Onset, a complex Onset being one containing more than one consonant. If there is only one consonant between two vowels, then that consonant is the Onset of the second syllable. If there are two consonants abutting of the same sonority, the syllable boundary falls between them, as in "threadbare." In general, if there are several consonants between vowels, then the consonant with the lowest sonority marks the start of the second syllable. The sonority hierarchy is given in Table 1 [3]. The principal exception to this is peripheral */s/*. For example, in the compound word "snakeskin" */sneikskIn/*, the word-internal proclitic */s/* that starts the second syllable falls between two consonants (*/k/*) of lower sonority Note that, on phonological criteria alone, it is not possible to determine whether peripheral */s/* is proclitic or enclitic. This can only be resolved by reference to morphological information. More generally, since our algorithm doesn't include direct knowledge of morphology (other than through knowledge of periphery), we will need to add this information if we are to match syllabification with morphology for words like "be+smirched", "be+stow", "bath+robes", and "birth+rates", which would be syllabified as */b ax s /-/ m er ch t /*, */b ax s / t ow /*, */ b ae th / r ow b z /* and */b er th /-/ r ey t s /*, respectively, by rule of sonority.

Sound	Sonority Index	Sound	Sonority Index
a	10	e,o	9
i,u	8	r	7
l	6	m,n	5
s	4	v,z,th(voiced)	3
f,th(voiceless	2	b,d,g	1
p,t,k	0.5		

Table 1: Sonority scale for phonemes.

Once the syllables are marked, we define the following three constituents as detailed in [2], where we distinguish

between Enclitic and Proclitic in the Periphery.

Proclitic: Syllable component that only occurs morpheme initially. /s/ in (*still*) or /S/ in (*shrugged*).

Core: Syllable component common to all languages types. It contains the obligatory vowel.

Enclitic: Syllable component that only occurs morpheme finally.

These three parts, thus defined, capture a certain syllable structure, where P, $C1$, $C2$, and E (Figure 1) denote allowed sets of consonants, V denotes the set of vowels, and F denotes either a consonant or vowel, the latter being the second moraic element of a long vowel or diphthong. Given a word then, which is marked at the syllable level, it is possible to automatically find the three constituents. In a complex onset (consisting of more than one consonant), the first phoneme is marked as proclitic if it is */s/* or */S/*. In the Rhyme, consonants are marked as enclitic unless they are either an */s/*, an */l/* or an "assimilating nasal" occurring immediately after a short vowel. Assimilating nasals occur in words such as pump, rant, rank, combat, bandage, languid, ranch, hinge, mince, lens, triumph, etc. The "assimilating nasal" refers to a nasal consonant whose place of articulation (labial, laminal/apical-dentalveolar/postalveolar, dorso-velar-lips, front-tongue, back-tongue), coincides with the place of articulation of the following consonant. Given these rules, we have therefore described the algorithm for marking core and periphery of syllables. The next step is then to syllabify a pronunciation dictionary so that core and periphery can be marked.

2.3. Evaluation

There is no validated reference syllabification by which to judge lexicon syllabification. So, in order to evaluate our algorithm, we want to syllabify a dictionary, which is already marked at the syllable level. The dictionary we are using for comparison has been developed at the Johns Hopkins summer school [5] and is a close variation of the high quality Pronlex lexicon, which has been automatically marked at the syllable level using Daniel Kahn's [4] Principle of English syllabification. Here, syllabification was controlled by three user-supplied lists: permitted syllable-initial consonant clusters (onsets), permitted syllable-final consonant clusters (codas), and prohibited onsets. This process is first run on native onsets and codas and then repeated for all words that failed syllabification by using corresponding lists of foreign onsets and codas while handchecking for satisfactory results. This syllabification algorithm used the generally accepted syllabification method that maximises onsets, assigning as many consonants as possible to syllable onsets while subject to the constraints of the list of permitted onsets. The dictionary contains around 71000 entries where we agreed on all but ca. 1300 syllabifications. In many cases, the phoneme */s/* was at the onset of a syllable in the dictionary while we assign */s/* to the coda (F or E) in certain compound words. Since conventional methods use beginnings of words as the way to model how syllables start, */thr/* in bathrobe, is allowed because it occurs in words such as 'throng'. English has the sequence */str/* at the beginning of words like "string", so that syllabification of "mistreat" for example is analysed as */mI/+/stri:t/*. Similarly, since English doesn't have short vowels at the end of words, in some models 'attitude' is analysed as */At/+/It/+/u:d/* rather than */A/+/tI/+/tu:d/* as in our algorithm. Such models often designate single consonants between vowels as "ambisyllabic"—ambiguous or belonging to both syllables).

Generally our syllable boundaries were correctly placed at the morphological boundaries more often than in the reference dictionary which can be explained with our indirect knowledge of morphology due to the knowledge of periphery. We take what happens at the beginnings and the ends of words to be exceptional, not the norm. We take syllable boundaries in the middle of words to be the way to model how syllables end and start generally. In addition, we differentiate between syllable transitions that occur where two morphemes meet and those that occur within a single morpheme. Though we can capture many morphologically correct syllables by this method, we need to extend our algorithm to include morphological knowledge in order to deal more effectively with prefixes and suffixes in the syllabification of words like "besmirch" */b ax s / m er ch/*.

3. FOREIGN ACCENT IDENTIFICATION

We expect speakers with greatest syllable structure differences between native and foreign language to have greatest difficulty with pronunciation in the foreign language. Similar to the example of the German accent, the behaviour of substitution of phonemes can be radically different for Core and Periphery of the syllable. We hypothesise a typology of syllable types based on Core vs. Periphery functions. At one end is English (or German) and at the other, tone languages like Vietnamese, Cantonese, Mandarin. Between these two extremes are languages without lexical tone with segmental configurations simpler than English. Syllable structures in tone languages tend to be comparatively simple in terms of phone segments, but are complicated by tones, each of which extends for the duration of a syllable or syllables expressing a grammatical unit, usually the word. The tone thus indicates the extent of the word. This difference in language typology has a strong effect on the ability to pronounce English in parts of the syllable that demarcate grammatical units. In order to study the structure of this type of foreign accent in English, we chose Vietnamese speech data. In contrast, Lebanese Arabic syllable structure has much more in common with English. We hypothesise that the pronunciation of English by Lebanese foreign speakers will be much closer to that of native speakers, and the variability less than that of a Vietnamese speaker.

3.1. DATA

The data used in this study come from the The Australian National Database of Spoken Language (ANDOSL [1]) [6]. The speech was recorded in an Anechoic chamber at the National Acoustics Laboratories of Sydney, Australia. We compare native Australian English to Vietnamese- and Lebanese-accented Australian English. The training set and test set for Australian English consist of one male speaker each. Each speaker read 200 phonetically rich and balanced sentences containing all types of phoneme combinations of Australian English pronunciation. Because the 200 sentences demanded a high degree of literacy from speakers for whom English was a non-native language, 50 sentences were chosen from the 200 and adjusted to have one member of every phoneme class in every permissible position. These were then read by the Vietnamese- and Lebanese-accented speakers. For Vietnamese, the training set and test set consist of six and three speakers respectively; the Lebanese training and test set consist of three speakers each. In order to analyze the accents, all speech was labelled by linguists with the closest Australian English phonemes achieved by the speakers. The second level of labeling consists of the transcribed words. Also available were a small dictionary covering all the words in the sentences that were uttered. This dictionary contained a single pronunciation model for each word representing the "ideal" speaker. Our syllabifier performs at 100% accuracy according to this dictionary which was syllabified by linguists.

Word	Syllable structure	actual pronunciation
1. The	D@(C)	/d/@:/
2. length	lE(C)NT(E)	/l/E/N/
3. of	O(C)v(E)	/O/b/
4. her	h@:(C)	/h/@:/
5. skirt	s(P)k@:(C)t(E)	/s/k/@:/s/
6. caused	ko:(C)zd(E)	/k/@u/s/
7. the	D@(C)	/d/@/
8. passers-by	pa:(C)s@(C)z(E)bai(C)	/p/a:/s/b/ai/
9. to	tu:(C)	/t/u:/
10. stare	s(P)te:(C)	/s/t/e:/

Table 2: Examples of English words as pronounced by a Vietnamese speaker. (E) denotes the Enclitic part, (C) the core part. Types of mistakes include: D→ d (1,7), deletion (2,8), Enclitic substitution (3,5), Enclitic devoicing (6), Enclitic simplification (6)

3.2. Aligning Utterances to Target Pronunciation

In order to study the accented speech as a function of syllable position, it is necessary to align the achieved phoneme sequence (handlabeled with English phonemes by linguists) with the target phoneme strings. An example sentence, in Table 2, "The length of her skirt caused the passers-by to stare" shows both target phonemes (in Australian English) and achieved phoneme string (as spoken by a sample Vietnamese speaker). The example shows how difficult it can be to align the two strings correctly in order to tag the syllable position of each of the actual pronunciations.

In the absence of a confusion matrix which could be obtained from training a phoneme recognizer, we use Dynamic Time Warping (DTW) in order to align the two strings with linguistic knowledge. The score to be maximized by matching achieved and target phoneme is calculated by summing up points as given in Table 3 over all shared categories over all possible phoneme pairs to be matched. Points listed in this table approximately reflect the degree of relatedness between two phonemes containing this feature. If we were to make a tree of all phoneme features, then the number reflects the depth of the tree at which is located a particular feature. For example, phonemes can be either vowels or consonants (1 point), vowels can be short or long (1.5 points), short vowels can be back or front (2 points). From this basic method, ambiguities are resolved with linguistic knowledge and points are altered by looking at the relative similarity of phonemes at different depths in the tree. So, for example, high short vowels and mid short vowels only receive 1 point, even at the same depth in the tree as back and front vowel. Matching /D/ (*loath*) to target /T/ (*bath*) results in a score: 1 (consonants) + 2 (fricatives) + 4 (laminodentals) + 1.5 (continuants) = 8.5. A perfect match to /T/ would have included 1.5 (voiceless). Matching /t/ to /T/, the score would result in 1 (consonants) + 2.5 (distal voiceless) + 1.5 (voiceless) = 5, which is smaller than 8.5; a less valuable match.

Category	Points	Category	Points
VOWELS	1	SHORT	1.5
LONG	1.5	BACK SHORT	2
CENTRAL SHORT	2	FRONT SHORT	2
BACKISH LONG	2	CENTRAL LONG	2
FRONT LONG	2	HIGH SHORT	1
LOW SHORT	1.5	MID SHORT	1
HIGH LONG	1	LOW LONG	1.5
MID LONG	1	DIPHTHONGS	1.5
RISING DIPH	3	FRONTING DIPH	0
CLOSING DIPH	3	CENTERING DIPH	2.5
INIT ROUNDING	1.5	FINAL ROUNDING	2
CONSONANTS	1	VOICELESS	1.5
VOICED	1.5	NASAL	4
LIQUID	4	APPROXIMANT	4
GLIDE	4	SONORANT	3
STOP	2.5	CONTINUANT	1.5
FRICATIVE	2	AFFRICATE	2.5
STOP FRIC	3	OBSTRUENT	1
LABIAL	2	LABIO DENTAL	4
LAMINO DENTAL	4	APICO ALVEOLAR	2
LAMINO POSTALVEOLAR	3	DORSO VELAR	4
DISTAL VOICELESS	2.5	DISTAL VOICED	2.5

Table 3: Linguistic Categories with corresponding points directly proportional to acoustic closeness (proportionate to number of common linguistic features).

The dynamic time warp returns two phoneme strings of the same length N, with each position, i, either marking a substitution, an insertion or a deletion. We thus have achieved an automatic method for marking the syllable

[1] More information on this database can be obtained at http://andosl.anu.edu.au:80/andosl/

position (Proclitic, Core, or Enclitic) within a pronunciation as inherited by the target dictionary pronunciation. While this method of alignment seems to work fine by inspection, it may be possible to improve the algorithm by acoustic analysis of closeness of phonemes within different categories.

3.3. Feature Analysis

Our goal is to look at the discrimination capability of features as a function of their position in the syllable. We want to see if position information improves the discrimination. Features used here correspond to occurrence frequencies of phoneme labels in the hand-labeled data for Vietnamese, Lebanese and Australian accented English. In order to identify discriminating features for any two classes of accented English speakers, it is essential to have a good estimate discrimination error due to a given feature. The estimate of the discriminability of two accents can be quantified for each feature based on a model of the feature distribution in the two accent classes introduced. We model each features by using a normal distribution, as shown in Figure 2, taking into account the mean occurrence frequency of a given feature, and the variation across speakers. Using this model, discriminating features can be extracted by estimating the Bayes' error due to two class-dependent distributions.

$$\text{Distance Measure} = \frac{1}{2}\exp -\frac{1}{4}\frac{(u_1[j]-u_2[j])^2}{s_1[j]^2+s_2[j]^2} \quad (1)$$

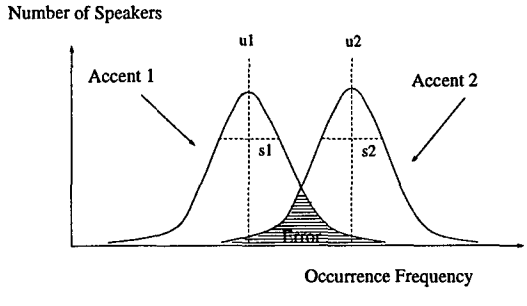

Figure 2: Normal Distribution.

For each of the features the corresponding discrimination error is estimated and thus we are able to look at the most important N features which will indicate the performance of accent discrimination based on this type of phoneme-based feature. Based on this model, we can now identify and sort the features by their classification error. Figure 3 depicts a graph of the top 40 features with respect to their corresponding estimated discrimination ability. From this graph, we can see that (1) Lebanese has less discriminating features which show less improvement when including position information. Vietnamese is a tone language and therefore, as expected, we see more improvement with this type of feature set.

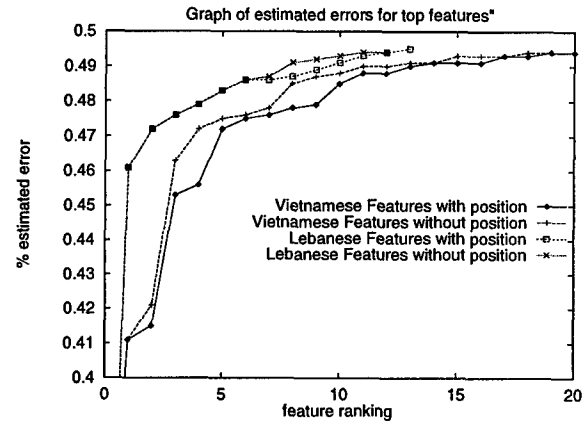

Figure 3: Top features or Lebanese vs. English and Vietnamese vs. English plotted as function of their estimated error and comparing position dependent features, with position independent Features. As expected, more improvement is seen in the Vietnamese list.

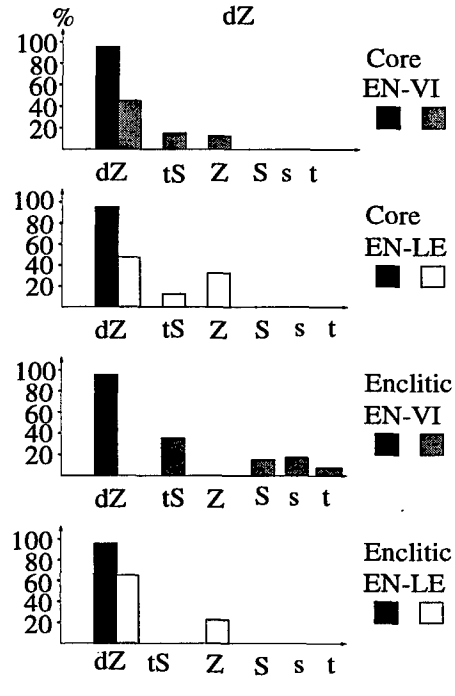

Figure 4: Comparison of language- and position-dependent substitutions for phonemes of /dZ/. Substitutions are different for Lebanese and Vietnamese and Core and Enclitic. Lebanese has less variability than Vietnamese.

3.4. Results

The total number of confusions is too large to describe here. In general, looking only at consonants, we can note the following trends:

- Confusions are different across accent groups.

- Confusions differ for *Enclitic* and *Core*.
- Lebanese speakers are more consistent in their substitutions than Vietnamese speakers. (See example for /dZ/ in Figure 4).
- Vietnamese accented speakers have a stronger accent than Lebanese accented speakers in terms of changes in voicing, manner, place and class. (See example for /dZ/ in Figure 4).
- The variability of the confusions is generally higher in the *Enclitic* than in the *Core* part of the syllable for both Vietnamese and Lebanese for /N/(*laughing*) and voiced fricatives.
- The variability of the confusions in the *Enclitic* is generally higher in Vietnamese than in Lebanese for stops, unvoiced fricatives, /T/, and /D/.
- phonemes /T/, /D/, /S/ and /z/(*zap*) are difficult for Vietnamese regardless of position.
- Voiced affricates are difficult for both accent groups.
- These trends are upheld across all speakers, however, the confusion probabilities vary.

One example, in particular, relates to the phoneme /d/ in Vietnamese. This phoneme is much more interesting for discriminability when treated as a function of position. In the Enclitic part its frequency is higher in English, but in the Core part its frequency is higher in Vietnamese. We now have the ability to study why this phenomenon takes place and why syllable position is so important. Table 4 lists some of the relevant confusions. We can see that /d/ is a substitute for /D/ (as 'th' in "the") for Vietnamese speakers—only in the Core part. In the Enclitic part of the syllable the pattern is quite different in that /D/ is simply devoiced. In addition, it can be seen that while /d/ is mostly pronounced correctly by Vietnamese speakers in the Core, /d/ is devoiced to /t/ in the Enclitic. All these effects combine to result in Vietnamese accent with a higher frequency of /d/ in the Core and a lower frequency of /d/ in the Enclitic when compared to native English.

Confusions including /d/				
Position	Target	Achieved	English	Vietnamese
Core	D	D	0.99	0.33
	D	d	0.00	0.60
Enclitic	D	D	1.00	0.15
	D	T	0.00	0.27
	D	s	0.00	0.19
	D	t	0.00	0.27
Core	d	d	0.96	0.93
Enclitic	d	d	0.99	0.48
	d	s	0.00	0.12
	d	t	0.01	0.28

Table 4: Shows importance of location information of phoneme /d/ in Vietnamese accent.

3.5. Conclusions

No statistical analysis of these trends have been made due to the small amount of data used for analysis. However, having applied this information to a larger system, we have shown in [1] that accent identification can be improved by using syllable dependent information. In this paper we have shown that the position within the syllable is important because the pronunciation patterns of accented speakers vary as a function of the phoneme's position within the syllable and that the linguistic theory is reflected in real speech data and can be systematically captured. The linguistic understanding of this theory provides a means of predicting the discrimination potential for a given accent group when using this method. Having shown the connection between linguistics, theory and real data, we have gained the ability to reason about system performance at the linguistic level. This algorithm may also serve as a powerful tool for language teaching or alternatively for speaker identification/verification as certain habits of speakers might be captured much more effectively within the syllable constituents.

4. REFERENCES

[1] K. M. Berkling, Chris Cleirigh, Julie Vonwiller, and Marc Zissman. Improving accent identification through knowledge of english syllable structure. In *Proceedings International Conference on Spoken Language Processing*, Sydney, Australia, 1998.

[2] C. Cleirigh and J. Vonwiller. Accent identification with a view to assisting recognition. In *Proceedings International Conference on Spoken Language Processing*, volume 1, pages 375–379, Yokohama, Japan, apr 1994.

[3] J. A. Goldsmith. *Autosegmental and Metrical Phonology*. Basil Blackwell, 1 edition, 1990.

[4] Daniel Kahn. *Syllable-based generalizations in English phonology*. PhD thesis, Massachusetts Insitute of Technology, 1980.

[5] M. Ostendorf, B. Byrne, M. Bacchian, M. Finke, A. Gunwardana, K. Ross, S. Roweis, E. Shirberg, D. Talkin, A. Waibel, B. Wheatley, and T. Zeppenfield. Modeling systematic variations in pronunciation via language-dependent hidden speaking mode. In *Proc. 1996 Summer Workshop on Speech Recognition*, 1996.

[6] J. Vonwiller, I. Rogers, Ch. Cleirigh, and W. Lewis. Speaker and material selection for the australian national database fo spoken languages. *Journal of Quantitative Linguistics*, 2(3):177–211, 1995.