IMPROVING ACCENT IDENTIFICATION THROUGH KNOWLEDGE OF ENGLISH SYLLABLE STRUCTURE

Kay Berkling, Marc Zissman

M.I.T., Lincoln Laboratory
244 Wood Street
Lexington, MA 02420-9108, USA
kay,maz@sst.ll.mit.edu

Julie Vonwiller, Chris Cleirigh

University of Sydney,
Dept. of Electrical Engineering,
Sydney, Australia,
julie,cleirig@speech.su.oz.au

ABSTRACT

This paper studies the structure of foreign-accented read English speech. A system for accent identification is constructed by combining linguistic theory with statistical analysis. Results demonstrate that the linguistic theory is reflected in real speech data and its application improves accent identification. The work discussed here combines and applies previous research in language identification based on phonemic features [1] with the analysis of the structure and function of the English language [2]. Working with phonemically handlabelled data in three accented speaker groups of Australian English (Vietnamese, Lebanese, and native speakers), we show that accents of foreign speakers can be predicted and manifest themselves differently as a function of their position within the syllable. When applying this knowledge, English vs. Vietnamese accent identification improves from 86% to 93% (English vs. Lebanese improves from 78% to 84%). The described algorithm is also applied to automatically aligned phonemes.

1. INTRODUCTION

The ability to approximate English phonology depends on native language similarity of **articulation** (phone inventories, syllable structure), **intonation**, and **rhythm**. In the past, research of different accent groups has focused on phone inventories and sequences, acoustic realizations, [4, 6] and intonation patterns [5, 3]. In this paper we describe how the study of the English syllable structure allows us to extend the range of useful features. In order to discriminate foreign-accented speech, we introduce a new feature dimension which includes the location of the phoneme within a syllable and apply it to discriminate between native speakers of Australian English (EN) and Vietnamese (VI) or Lebanese (LE).

The English language employs a particular syllable structure to assist in demarcating grammatical units. Because not all languages use segmental constituents for this function, some foreign speakers of English will have trouble pronouncing these demarcative syllable constituents. The goal of this study is to show that the linguistically-based theory is reflected in actual speech data and that this knowledge improves identification of accented speech. This paper is organized as follows: Section 2 outlines the linguistic theory underlying the design of the accent identification system. Section 3 describes the database. Section 4 will study the differences we find in foreign-accented speech and Section 5 applies this knowledge in an accent classification system.

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. LINGUISTIC BACKGROUND

A phrase in any language consists of words which in turn are realized by syllables. A syllable usually consists of an obligatory vowel with optional initial and final consonants. One familiar way of subdividing a syllable is into *Onset* and *Rhyme*, as shown in Figure 1. Here, P, $C1$, $C2$, F, and E denote allowed sets of consonants. V denotes the set of vowels in the *Rhyme*.

All syllables in all languages consist of *Onset* and *Rhyme* (phonetically, at least). However, these categories alone do not indicate where the syllable is placed within the word. In order to capture foreign accent in English, we want to highlight those constituents of the syllable that are most likely to prove difficult for speakers of languages in which they are not contained. We define the following three consituents as detailed in [2]:

- **Proclitic:** Syllable component that only occurs morpheme initially. /s/ (*still*) or /S/ (*shrugged*) is Proclitic when the Onset has more than one consonant preceding the vowel.
- **Core:** Syllable component common to all languages types. It contains the obligatory vowel.
- **Enclitic:** Syllable component that only occurs morpheme finally. A Rhyme consonant is Enclitic unless it is either /s/, /l/, or an assimilating nasal occurring immediately after a short vowel.

These three parts, thus defined, capture a certain syllable structure. Within that structure, the peripheral elements can be said to demarcate the boundary of grammatical units in the English language. As an example, the word "asked" (/a:/s/k/t/) can be broken down into the constituents as */a/(Core)* and */s/k/t/(Enclitic)*. The *Enclitic* here not only demarcates the end of the word but also includes the past-tense morpheme of the verb, realised by /t/, which thus carries grammatical meaning.

Only some languages have *Proclitics* and *Enclitics*. In contrast to English, tone languages use tone for the same function. Syllable structures in tone languages tend to be comparatively simple in terms of phone segments, but are complicated by the extension of a tone 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

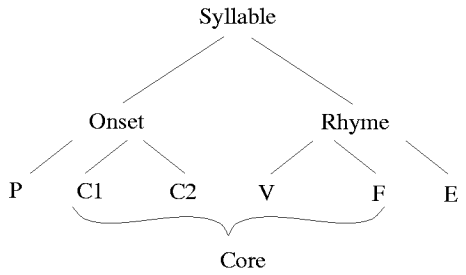

Figure 1: Constituents of a syllable as defined in this paper. P = Proclitic, E = Enclitic

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 hypothesize 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. Identifying Lebanese accents may therefore be harder at this level of analysis.

3. DATABASE

The data used in this study come from the The Australian National Database of Spoken Language (ANDOSL) [7]. The speech for this database 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 the permissible 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. The speech was labelled by linguists at the phoneme and word levels [1].

In addition, HTK was used to train a 40-phoneme recognizer on 200 utterances from each of twenty-four Australian English speakers. The accuracy of our phoneme recognizer is 41%, 43%, and 35% when evaluated on the Australian English training and test set (200 utterances from five speakers each) and the Vietnamese test set (total of 600 utterances from nine speakers) respectively. This recognizer was then used to automatically align an independent training and test set for Australian and Vietnamese accented English. Because we are now using automatically generated phoneme strings, the Australian English training and test sets are much larger than for the handlabeled utterances. The Australian English training and test set include five and six speakers respectively, with 200 utterances each. The Vietnamese training and test set are the same as for the experiment using aligned and handlabeled data.

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

Category	Pts.	Category	Pts.
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 1: Linguistic Categories with corresponding points directly proportional to acoutic closeness (proportionate to number of common linguistic features).

4. FEATURE ANALYSIS

Before building a system for accent classification, we want to study the structure of manifested accent. To do this, we match a target pronunciation as given by the dictionary to the achieved string of phonemes for each utterance. Normally, a confusion matrix obtained from training a phoneme recognizer is used for this purpose. Since no recognizer was trained, we use linguistic knowledge to obtain a matching score, which is maximized during the dynamic time warping algorithm. A matching score between achieved and target phoneme is calculated by summing up points as given in Table 1 over all shared categories. 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 (voicless) = 5, which is smaller than 8.5; a less valuable match.

Such a dynamic time warp returns two phoneme strings of the same length N, with each position, i, either matching a phoneme, marking an insertion or a deletion. We thus have a means of looking at the confusions between target and achieved phonemes as a function of the syllable position (Proclitic, Core, or Enclitic), dictated by the target, and the language. Looking only at consonants, we note the following trends (Figure 2 shows some typical examples.

1. Confusions are substantially different across accent groups.
2. Confusions differ substantially for *Enclitic* and *Core*.
3. Lebanese speakers are much more consistent in their substitutions than Vietnamese speakers.
4. Vietnamese accented speakers have a much stronger accent than Lebanese accented speakers in terms of changes in voicing, manner, place and class.

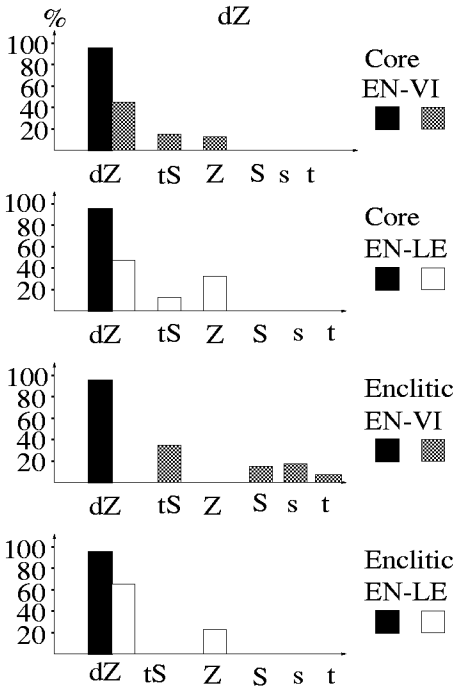

Figure 2: Comparison of language- and position-dependent substitutions for phonemes of /dZ/. Subistitutions are differnent for Lebanese and Vietnamese and Core and Enclitic. Lebanese has less variability than Vietnamese.

5. The variablity 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.
6. The variablity of the confusions in the *Enclitic* is generally higher in Vietnamese than in Lebanese for stops, unvoiced fricatives, /T/, and /D/.
7. phonemes /T/, /D/, /S/ and /z/(*zap*) are difficult for Vietnamese regardless of position.
8. Voiced affricates are difficult for both accent groups.
9. These trends are upheld across all speakers, however the confusion probabilities vary.

5. ACCENT IDENTIFICATION SYSTEM

We now build a simple accent-identification baseline system as shown in the block diagram of Figure 3. For each accent (native, Vietnamese, and Lebanese) denoted by α, a confusion matrix P_α is computed relating the probability of a target phoneme given an achieved phoneme. A given achieved phoneme sequence A is classified by calculating the probability of a match with the target sequence T as given by Equation 1, where N corresponds to the length of the match. The classified accent $\hat{\alpha}$ corresponds to the accent of the confusion matrix which yields the highest score.

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}} \prod_{i=0}^{N} P_\alpha(T_i|A_i) \quad (1)$$

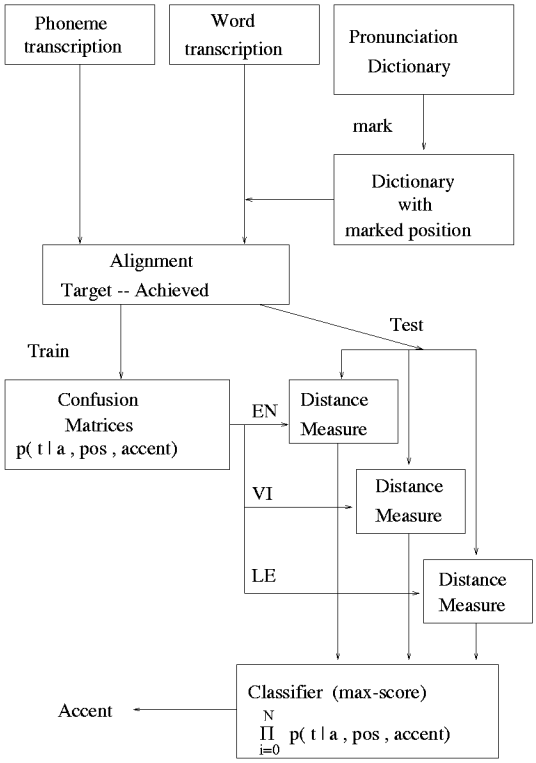

Figure 3: Block diagram of accent identification system.

In order to improve the accent identification system, we now incorporate the insight gained from the linguistic knowledge and observation of the data. Confusion matrices $\gamma_\alpha^{\phi t}$ are calculated for each language, differing from P_α in that they are calculated separately for each position $\phi_t \epsilon$ (Proclitic, Core, Enclitic) of target phoneme t. The accent is now classified as given by Equation 2.

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}} \prod_{i=0}^{N} \gamma_\alpha^{\phi_{T_i}}(T_i|A_i) \quad (2)$$

Figure 4 plots the comparative results for the test sets of Vietnamese and Lebanese vs. native speakers as a function of the number of phonemes processed[2]. Accent classification based on various levels of position information (Core/Proclitic and Enclitic/Proclitic, Core and Enclitic/none) are compared. Table 2 compares results using Eq. 1 and Eq. 2 for $N = 40$. Using position dependent information (Eq. 2), consistently improves performance: English vs. Vietnamese improves from an overall 86% to 93% correct classification. English vs. Lebanese improves from 78% to 84% correct classification[2]. The plot shows that while both Core and Periphery information are important in acoustic matching of the achieved phoneme string to the target, most of the speaker independent information seems to be contained in the Core. As predicted, Lebanese accent identification is more difficult with this method than Vietnamese identification.

In order to study how well our theory might generalize from handlabeled to automatically aligned phonemes, we align a training and test set for Australian and Vietnamese accented English as

[2]Three way accent identification improves from 69% in the test set to 77% when using Eq. 2 instead of Eq. 1

defined in Section 3. Each of the automatically aligned phoneme strings was then analyzed in the same manner as the handlabeled strings, using knowledge of the target non-time aligned word transcriptions. Even though there are obviously some improvements to be made to the recognizer, Table 2 indicates that foreign accent identification for Vietnamese vs. Australian English can be improved by using position information. Results are evaluated after processing 40 phonemes in each of the strings. When using position information performance improves from 84% to 88% for the training set and from 84% to 89% on the test set. Table 2 gives detailed results for both accent groups.

Handlabeled				
Eq.2 (Eq.1)	Training Set		Test Set	
input-output	EN	VI	EN	VI
EN	100 (100)	0 (0)	98 (96)	2 (4)
VI	3 (12)	97 (88)	13 (25)	87 (75)
Eq.2 (Eq.1)	Training Set		Test Set	
input-output	EN	LE	EN	LE
EN	100 (99)	0 (1)	90 (88)	10 (12)
LE	10 (13)	90 (87)	20 (28)	80 (72)
Automatically Aligned				
Eq.2 (Eq.1)	Training Set		Test Set	
input-output	EN	VI	EN	VI
EN	99 (97)	31 (39)	98 (97)	40 (55)
VI	1 (3)	69 (61)	2 (3)	60 (45)

Table 2: % correct accent identification after processing $N = 40$ phonemes. Results using Eq. 2, are compared to the baseline system (in parenthesis), using Eq. 1.

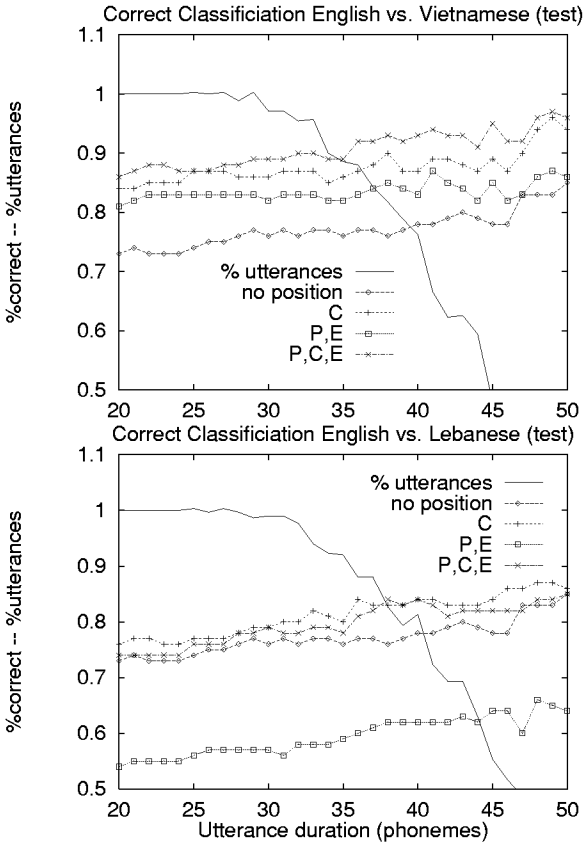

Figure 4: (a) Handlabeled, English vs. Vietnamese (b) Handlabeled, English vs. Lebanese. % correct classification using different combinations of information of C (Core), P (Proclitic), and E (Enclitic) or disregarding it. Also indicated is the % of test utterances of length N.

6. DISCUSSION AND FUTURE WORK

Our aim was to show that the position within the syllable is important because the pronunciations of speakers vary as a function of the phoneme's position within the syllable. This theory is not only useful for identification of foreign-accented speech but improves a match between achieved and target pronunciation speech (not only for accented but also for native speech). Results indicate that this algorithm can be very useful in a speech recognition system, which includes word recognition. After improving automatic recognition, we would like to extend the theory so that we can apply it without knowledge of word transcription. Finally, we plan to evaluate whether this algorithm is useful for speaker identification.

7. ACKNOWLEDGEMENTS

We want to thank Dr. Furui, who made it possible to continue this work in Japan. His helpful discussions added valuable insights.

8. REFERENCES

[1] K. M. Berkling. *Automatic Language Identification with Sequences of Language-Independent Phoneme Clusters*. PhD thesis, Oregon Graduate Institute, October 1996.

[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.H.L. Hansen and L.M. Arslan. Foreign accent classification using source generator based prosodic freatures. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 836–839, Detroit, may 1995.

[4] K. Kumpf and R.W. King. Foreign speaker accent classification using phoneme-dependent accent discrimination models and comparisons with human perception benchmarks. In *Proceedings Eurospeech*, volume 4, pages 2323–2326, Rhodos, Greece, Sep 1997.

[5] H. Mixdorff. Foreign accent in intonation patterns - a contrastive study applying a quantitative model of the f0 contour. In *Proceedings International Conference on Spoken Language Processing 96*, volume 2, pages 1469–1472, Philadelphia, Oct 1996.

[6] C. Teixeira, I. Trancoso, and A. Serralheiro. Recognition of non-native accents. In *Proceedings Eurospeech*, volume 4, pages 2375–2379, Rhodos, Greece, Sep 1997.

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