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.

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/ (gill) 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
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. 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 consist of 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.

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

<table>
<thead>
<tr>
<th>Category</th>
<th>Pts.</th>
<th>Category</th>
<th>Pts.</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOWELS</td>
<td>1</td>
<td>SHORT</td>
<td>1.5</td>
</tr>
<tr>
<td>LONG</td>
<td>1.5</td>
<td>BACK SHORT</td>
<td>2</td>
</tr>
<tr>
<td>CENTRAL SHORT</td>
<td>2</td>
<td>FRONT SHORT</td>
<td>2</td>
</tr>
<tr>
<td>BACKISH LONG</td>
<td>2</td>
<td>CENTRAL LONG</td>
<td>2</td>
</tr>
<tr>
<td>FRONT LONG</td>
<td>2</td>
<td>HIGH SHORT</td>
<td>1</td>
</tr>
<tr>
<td>LOW SHORT</td>
<td>1.5</td>
<td>MID SHORT</td>
<td>1</td>
</tr>
<tr>
<td>HIGH LONG</td>
<td>1</td>
<td>LOW LONG</td>
<td>1.5</td>
</tr>
<tr>
<td>MID LONG</td>
<td>1</td>
<td>DIPHTHONGS</td>
<td>1.5</td>
</tr>
<tr>
<td>RISING DIPH</td>
<td>3</td>
<td>FRONTING DIPH</td>
<td>0</td>
</tr>
<tr>
<td>CLOSING DIPH</td>
<td>3</td>
<td>CENTERING DIPH</td>
<td>2.5</td>
</tr>
<tr>
<td>INIT ROUNDED</td>
<td>1.5</td>
<td>FINAL ROUNDED</td>
<td>2</td>
</tr>
<tr>
<td>CONSONANTS</td>
<td>1</td>
<td>VOICELESS</td>
<td>1.5</td>
</tr>
<tr>
<td>VOICED</td>
<td>1.5</td>
<td>NASAL</td>
<td>4</td>
</tr>
<tr>
<td>LIQUID</td>
<td>4</td>
<td>APPROXIMANT</td>
<td>4</td>
</tr>
<tr>
<td>GLIDE</td>
<td>4</td>
<td>SONORANT</td>
<td>3</td>
</tr>
<tr>
<td>STOP</td>
<td>2.5</td>
<td>CONTINUANT</td>
<td>1.5</td>
</tr>
<tr>
<td>FRICATIVE</td>
<td>2</td>
<td>AFFRICATE</td>
<td>2.5</td>
</tr>
<tr>
<td>STOP FRIC</td>
<td>3</td>
<td>OBSTRUENT</td>
<td>1</td>
</tr>
<tr>
<td>LABIAL</td>
<td>2</td>
<td>LABIO DENTAL</td>
<td>4</td>
</tr>
<tr>
<td>LAMINO DENTAL</td>
<td>4</td>
<td>APICO ALVEOLAR</td>
<td>2</td>
</tr>
<tr>
<td>LAMINO POSTALVEOLAR</td>
<td>3</td>
<td>DORSO VEOLAR</td>
<td>2</td>
</tr>
<tr>
<td>DISTAL VOICELESS</td>
<td>2.5</td>
<td>DISTAL VOICELESS</td>
<td>2.5</td>
</tr>
</tbody>
</table>

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 /l/ (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 /l/ 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.

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.
5. The variability of the confusions is generally higher in the 
Enclitic than in the Core part of the syllable for both Viet-
namese and Lebanese for /N/(laughing) and voiced frica-
tives.
6. The variability 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 Viet-
namese 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 \( P_\alpha \) 
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.

\[
\hat{\alpha} = \arg \max_\alpha \prod_{i=0}^N P_\alpha(T_i|A_i) 
\]

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 \( \gamma^\phi_{t,\alpha} \) are calculated 
for each language, differencing from \( P_\alpha \) in that they are calculated sep-
arately for each position \( \phi \) (Proclitic, Core, Enclitic) of target 
phoneme \( t \). The accent is now classified as given by Equation 2.

\[
\hat{\alpha} = \arg \max_\alpha \prod_{i=0}^N \gamma^\phi_{t,\alpha}(T_i|A_i) \]

Figure 4 plots the comparative results for the test sets of Viet-
namese and Lebanese vs. native speakers as a function of the 
number of phonemes processed. Accent classification based on 
various levels of position information (Core/Proclitic and Enclitic/ 
Proclitic, Core and Enclitic/none) are compared. Table 2 com-
pares results using Eq. 1 and Eq. 2 for \( N = 40 \). Using position 
dependent information (Eq. 2), consistently improves perfor-
mane: English vs. Vietnamese improves from an overall 86% to 
93% correct classification. English vs. Lebanese improves from 
78% to 84% correct classification. 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 diffi-
cult 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 train-
ing 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.

<table>
<thead>
<tr>
<th></th>
<th>Training Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Handlabeled</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EN</td>
<td>100 (100)</td>
<td>98 (96)</td>
</tr>
<tr>
<td>VI</td>
<td>3 (12)</td>
<td>13 (25)</td>
</tr>
<tr>
<td><strong>Automatically Aligned</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EN</td>
<td>99 (97)</td>
<td>98 (97)</td>
</tr>
<tr>
<td>VI</td>
<td>1 (3)</td>
<td>2 (3)</td>
</tr>
</tbody>
</table>

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.

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


