ELSEVIER

Speech Communication 35 (2001) 125–138

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

SCoPE, syllable core and periphery evaluation: Automatic syllabification and foreign accent identification ☆

Kay Berkling *,[1]

Speech Technology Group, Department of Electrical Engineering, Sydney University, Sydney, Australia

Abstract

In this paper we apply a study of the structure of the English language towards an automatic syllabification algorithm and consequently an automatic foreign accent identification system. Any word consists of syllables which can in turn be divided into its constituents. Elements within the syllable structure are defined according to both their position within the syllable and the position of the syllable within the 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 (a detailed analysis of comparative phoneme inventories across the languages treated here is outside the scope of this paper), 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 analyze two accents of Australian English: Arabic whose core/periphery structure is similar to English and Vietnamese, whose structure is maximally different to English.

Zusammenfassung

In dieser Veröffentlichung wenden wir eine Studie über die Englische Silbenstruktur an, um eine automatische Silbentrennung zu erhalten und darauf aufbauend, ein System für das automatische Erkennen von Ausländischen Akzenten. Dieser Algorithmus wird auf handmarkierte Daten und automatisch markierte Daten angewendet. Elemente in der Sylbe werden definiert sowohl durch Ihre Position in der Silbe als auch im Wort. Solche Elemente, die an der Morphem-grenze liegen oder über die ganze länge des Morphemes gültig sind, werden als peripherale Elemente bezeichnet; solche, die an jeglicher Stelle stehen können, gehören zum Kern der Silbe. Alle Sprachen besitzen eine Art Kern und Peripherie der Silbe, aber spezielle Formen dieser Strukturen werden von Sprache zu Sprache verschieden sein. Zu den Problemen, die durch das unterschiedliche Phonem-inventar zustande kommen, stellt die unterschiedliche Silbenstruktur eine zusätzliche Anforderung an die Aussprache für Sprecher, deren Muttersprache sich maximal in den Silbenstruktur von dem Englischen underscheidet. In dieser Veröffentlichung werden wir zwei ausländische Akzente in Australischem Englisch vergleichen: Sprecher deren Muttersprache Arabisch ist (mit ähnlicher Silbenstruktur), und Sprecher deren Muttersprache Vietnamesisch ist (eine Sprache deren Silbenstruktur dem Englischen maximal unähnlich istl.

☆ 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.

* Present address: Buckhauser Str. 3, 8048 Zurich, Switzerland. Tel.: +1-400-5650.

E-mail addresses: kay@berkling.com, kay.berkling@swisslife.ch (K. Berkling).

[1] The author is presently employed by Swiss Life in Zurich, Switzerland, working in the e-Business Competence Centre.

0167-6393/01/$ - see front matter
PII: S0167-6393(00)00100-X

Résumé

Dans le présent article, nous utilisons les résultats d'une étude de la langue anglaise et l'appliquons dans un algorithme de syllabification. Notre système permet aussi l'identification automatique d'accents étrangers et peut etre utilisé avec des données générées de façon manuelle ou automatique. Les éléments de la structure syllabique sont definis selon leurs positions à l'intérieur des structures des syllables et des mots. Les éléments de la structure syllabique qui apparaissent uniquement à la bordure d'un morphème sont identifiés comme éléments périphériques; ceux qui se présente à n'importe quel endroit de la morphologie du mot sont identifiés comme éléments centraux. Toutes les langues font potentiellement une distinction entre les éléments centraux et péripheriques de leur structure syllabique. Cependant, les formes que prennent ces structures syllabiques varient d'une langue à l'autre. En plus des problémes posés par les différences entre les inventaires de phonèmes, nous nous attendons à ce que les personnes avec de grandes différences structurelles entre la langue maternelle et étrangere sont celles qui ont les plus grandes difficultés dans la prononciation de la langue étrangère. Dans cet article, nous analysons deux accents étrangères de l'anglais australien: la première est basé sur l'arabe, dont la structure centrale et périphérique est similaire à l'anglais, et la seconde est fondé sur le vietnamien, dont la structure diffère au plus au degré de celle de l'anglais.

Keywords: Foreign accent identification; Syllabification; Linguistic knowledge

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 about different accent groups has focused on phone inventories and sequences, acoustic realizations (Kumpf and King, 1997; Teixeira et al., 1997), and intonation patterns (Mixdorff, 1996) and (Hansen and Arslan, 1995). In this paper, we describe how the study of the English syllable structure allows us to extend this 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 database that we used for this study is introduced in Section 2.

The goal of the algorithm presented here is to exploit detailed knowledge of the English syllable structure model. Properties of accented speech are expressed in terms of phoneme substitutions, deletions or insertions as a function of syllable position. A simple example of the importance of position is demonstrated by a typical German speaker of English. Such speakers tend to devoice 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 phoneme position, we can potentially improve our feature set (of phoneme substitutions) for this type of phonological variation.

The algorithm's application to foreign accented speech in English derives from a more general study of the syllable structure of languages. Some time is devoted to the application of this study (Cleirigh, 1998) to English in Section 3.1, followed by an implementation of an automatic syllabification algorithm, which is also able to mark syllable constituents in Section 3.2, and an evaluation of the algorithm on a large standard database in Section 3.3. The algorithm is applied to foreign accent identification by aligning achieved pronunciations to target pronunciations in Section 4.1, analyzing the types of features in Section 4.2, and using the algorithm to build an automatic foreign accent identification system in Section 4.3.

2. Database

The data used in this study come from the The Australian National Database of Spoken Language (ANDOSL) [2] (Vonwiller et al., 1995). The

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

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 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. In order to analyze the accents, all speech was labeled by linguists with the closest Australian English phonemes achieved by the speakers. The second level of labeling consists of the transcribed words. Also available is 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.

In order to complete an additional experiment described in Section 4.3, involving only Australian English and Vietnamese, HTK was used to train a 40-phoneme recognizer on 200 utterances from each of 24 Australian English speakers. The accuracy of the resulting 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. For the experiments including the automatically labeled dataset, the Australian English training and test sets include five and six speakers respectively, with 200 utterances each. The Vietnamese training and test sets consist of the same speakers as for the experiment using hand labeled data.

3. The English syllable structure

Syllabification of pronunciation dictionaries is an important application because syllable information is used for text to speech synthesis and can be an important feature in speech recognition as well. 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 the syllable structure and the rationale that motivates it.

3.1. Linguistic background

A syllable usually consists of an obligatory vowel with optional surrounding consonants, the exception being where a schwa-like vowel and following consonant are realized 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 Fig. 1 that distinguishes between a *Core* and a

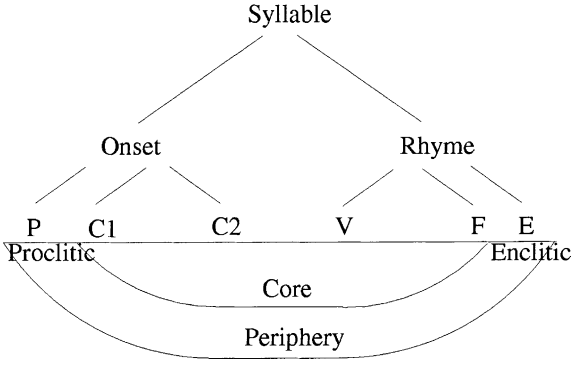

Fig. 1. Constituents of a syllable as defined in this paper. P, C1, C2, F, denote allowed sets of consonants and E denotes certain allowed sequences of consonants, V denotes the set of vowels. (A detailed description of these sets and sequences is beyond the scope of this paper but can be found in (Cleirigh, 1998).)

Periphery. The two parts of the periphery are called proclitic (placed before the core) and enclitic (placed after the core).

In Section 3.2 we show that *Core* and *Periphery* can be marked automatically, but first we would like to describe the underlying linguistic theory. 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 with 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 analyzed as the peripheral element of syllable structure. By analyzing 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 within 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.

3.2. Automatic syllable marking

In order to use the linguistic knowledge of syllable constituents as defined, we now want to devise an automatic method for 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.

3.2.1. The syllable

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 (Goldsmith, 1990). Sonority is equated with acoustic energy in order to establish this sonority index on a scale from less audible sounds like voiceless plosives and fricatives, to most audible sounds like vowels. 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

Table 1
Sonority scale for phonemes

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		

proclitic or enclitic. This can only be resolved by reference to morphological information. More generally, since our algorithm does not 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.

3.2.2. The syllable-constituents

Once the syllables are marked, we define the following three constituents as detailed in (Cleirigh and Vonwiller, 1994), 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 language 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, F and E (see Fig. 1) denote allowed sets of consonants and V denotes the set of vowels. 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 /s/, /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.

3.3. 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 that we used for our database was syllabified by linguists and our algorithm matched this syllabification with 100% accuracy. However, this dictionary is very small and we wanted to demonstrate how well this algorithm works on a larger and standard dictionary. For this purpose, we selected another dictionary, which has been developed at the Johns Hopkins summer school (Ostendorf et al., 1996) and is a close variation of the high quality pronlex lexicon, which has been automatically marked at the syllable level using Daniel Kahn's Principle of English syllabification (Kahn, 1980). 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 hand checking for satisfactory results. This syllabification algorithm used the generally accepted syllabification method that maximizes 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 71 000 entries where we agreed on all but around 1300 syllabifications. In many cases, the phoneme 's' or 'th' was at the onset of a syllable in the dictionary while we assign /s/ or 'th' 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/ as in bathrobe, is allowed within a syllable 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 analyzed as /mI/+/stri:t/. Similarly, since English does not have short vowels at the end of words, in conventional models "attitude" is analyzed 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. The way to model how syllables end and start in our algorithm is based on syllable boundaries in the middle of words. By differentiating, in addition, between syllable transitions that occur at morphemes boundaries and non-morpheme boundaries we are further able to define the Periphery. Though we can capture many morphologically correct syllables with this method (by marking syllables guided by the knowledge of Core and Periphery), 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/. This particular phenomenon has already been identified as a potential problem in the previous section.

4. 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 behavior of substitution of phonemes can be radically different for Core and Periphery of the syllable. We hypothesize a typology of syllable types based on Core versus 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 hypothesize that the pronunciation of English by Lebanese foreign speakers will be much closer to that of native speakers, and the variability (in manner and place of articulation approximating native pronunciation) across and within speakers less than that of a Vietnamese speaker.

4.1. 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 (hand labeled 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

Table 2
Examples of English words as pronounced by a Vietnamese speaker[a,b]

No.	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:/

[a] (E) denotes the Enclitic part, (C) the core part.
[b] Types of mistakes include: D→ d (1,7), deletion (2,8), Enclitic substitution (3,5), Enclitic devoicing (6), Enclitic simplification (6).

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 phoneme-based Dynamic Time Warping (DTW) in order to align the two strings using 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

Table 3
Linguistic categories with corresponding points, directly proportional to depth in tree based on linguistic similarity (i.e. number of common linguistic features)

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	Diphthong	1.5
Rising diphthong	3	Fronting diphthong	0
Closing diphthong	3	Centering diphthong	2.5
Initial 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 fricative	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

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. A simple example consists of matching an achieved /D/, (as in *loath*) to a target /T/, (as in *bath*) resulting 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.

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 position (Proclitic, Core or Enclitic) within a pronunciation as inherited by the matched 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.

4.2. 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. A feature in this context corresponds to the 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 of the 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 feature by using a normal distribution, as shown in Fig. 2, taking into account the mean occurrence frequency of a given feature, and the variation across and within speakers (simple histograms have shown that this approximation is reasonable despite the small number of speakers). Using this model, discriminating features can be extracted by estimating the Bayes' error due to two class-dependent distributions.

$$\text{Distance measure} = \frac{1}{2}\exp\left[-\frac{1}{4}\frac{(u1[j]-u2[j])^2}{s1[j]^2+s2[j]^2}\right]. \tag{1}$$

For each of the features the corresponding discrimination error is estimated and thus we are able to look at the most important N features that 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. Fig. 3

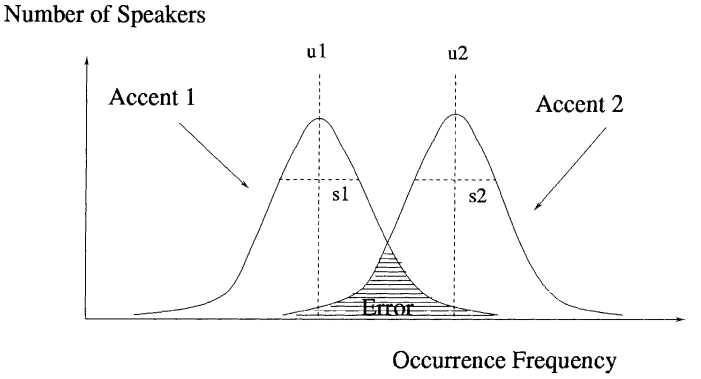

Fig. 2. Error caused by two overlapping normal distributions.

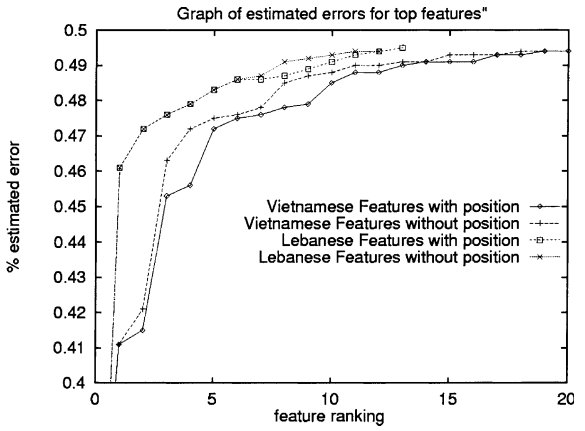

Fig. 3. Top features or Lebanese versus English and Vietnamese versus English plotted as function of their estimated error and comparing position-dependent features, with position-independent features. As expected, more improvement is seen in the Vietnamese list.

depicts a graph of the top features with respect to their corresponding estimated discrimination error (less than chance, which is 0.5). From this graph, we can see that Lebanese has less discriminating features (13, depicted in plot, that are less than 0.5) and shows little decrease in estimated error when including position information. In contrast, Vietnamese is a tone language and therefore, as expected, shows a better set of features (20 features with error <0.5) and more decrease in estimated error when including position information.

The total number of confusions is too large to describe here. In general, looking only at consonants, we can note the following trends (see examples in Table 4):

- Confusions are different across accent groups.
- Confusions differ between *Periphery* and *Core*.
- Lebanese speakers are more consistent in their substitutions than Vietnamese speakers.
- Vietnamese accented speakers have a stronger accent than Lebanese accented speakers in terms of changes in voicing, manner, place and class (see Fig. 4).
- The variability of the confusions is generally higher in the *Periphery* than in the *Core* part of the syllable for both Vietnamese and Lebanese for /N/ (as in *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.

One example, in particular, relates to the phoneme /d/ in Vietnamese. This phoneme is much more interesting for discriminating Vietnamese accents from native Australian English when considered as a function of position.

In particular, Table 4 exemplifies some of the relevant confusions. We can see here that /d/ is a substitute for /D/ (as 'th' in "the") for Vietnamese speakers – but only in the Core part. In the Enclitic part of the syllable however, 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 Periphery. 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. Therefore, making distinction of position, the occurrence frequency of /d/ becomes a feature that may well have been lost without position information. Table 4 also shows a particularly strong difference for the pronunciation of nasals in all three accent groups as a function of syllable position.

No statistical analysis of these trends have been made due to the small amount of data used for analysis. However, Section 4.3 will show that accent identification can be improved by using syllable dependent information even for larger datasets.

4.3. Automatic foreign accent identification

We now build a simple accent-identification baseline system as shown in the block diagram of Fig. 5. 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 (information that

Table 4
Except of most important confusion probabilities for three accent groups showing the importance of discriminating between Core and Periphery during parsing[a]

Position	Target	Achieved	English	Vietnamese	Lebanese
Affricates					
Core	dZ	Z	0.00	0.13	0.33
	dZ	dZ	0.97	0.46	0.48
	dZ	g	0.00	0.05	0.04
	dZ	tS	0.00	0.13	0.11
Enclitic	dZ	S	0.00	0.17	0.04
	dZ	Z	0.05	0.02	0.21
	dZ	d	0.00	0.06	0.04
	dZ	dZ	0.95	0.03	0.67
	dZ	s	0.00	0.19	0.04
	dZ	t	0.00	0.08	0.00
	dZ	tS	0.00	0.36	0.00
Core	tS	S	0.00	0.00	0.08
	tS	t	0.00	0.09	0.00
	tS	tS	1.00	0.84	0.92
Enclitic	tS	S	0.03	0.05	0.03
	tS	s	0.00	0.14	0.00
	tS	tS	0.97	0.70	0.97
Fricatives					
Core	D	D	0.99	0.33	0.91
	D	d	0.00	0.60	0.03
Enclitic	D	D	1.00	0.15	0.67
	D	T	0.00	0.27	0.22
	D	s	0.00	0.19	0.00
	D	t	0.00	0.27	0.00
	D	z	0.00	0.00	0.11
Core	T	T	1.00	0.56	0.87
	T	t	0.00	0.34	0.06
Enclitic	T	D	0.30	0.03	0.27
	T	T	0.66	0.39	0.62
	T	s	0.02	0.07	0.05
	T	t	0.00	0.42	0.05
Nasals					
Core	N	N	0.94	0.87	0.93
	N	n	0.06	0.13	0.07
Enclitic	N	N	0.98	0.52	0.79
	N	n	0.00	0.42	0.21
Sibilants					
Core	S	S	1.00	0.74	0.92
	S	s	0.00	0.25	0.00
	S	tS	0.00	0.01	0.06
Enclitic	S	S	1.00	0.55	1.00
	S	s	0.00	0.33	0.00
	S	tS	0.00	0.10	0.00

Table 4 (Continued)

Position	Target	Achieved	English	Vietnamese	Lebanese
Core	z	s	0.00	0.42	0.08
	z	z	0.97	0.52	0.89
Enclitic	z	s	0.01	0.81	0.23
	z	z	0.98	0.10	0.75
Core	s	s	1.00	0.98	0.97
Enclitic	s	s	0.99	0.91	0.93
	s	z	0.01	0.00	0.07

[a] Probabilities are based on the training set: six Vietnamese speakers, one English speaker and three Lebanese speakers.

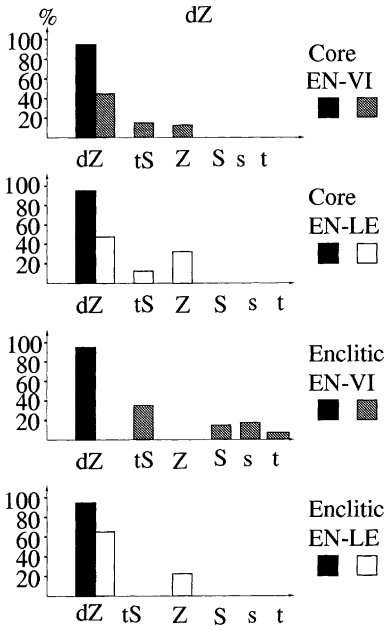

Fig. 4. Comparison of language- and position-dependent substitutions for phonemes /dZ/ in English (EN). Substitutions are different for Lebanese (LE) and Vietnamese (VI) and Core and Enclitic. Lebanese (LE) has less variability than Vietnamese (VI).

is obtained when applying the DTW-algorithm). A given achieved phoneme sequence A is classified by calculating the probability of a match with the target sequence T as given by Eq. (2), where N corresponds to the length of the DTW-match. The classified accent $\hat{\alpha}$ corresponds to the accent of the confusion matrix that yields the highest score.

$$\hat{\alpha} = \arg\max_{\alpha} \prod_{i=0}^{N} P_{\alpha}(T_i|A_i). \tag{2}$$

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 Eq. (3).

$$\hat{\alpha} = \arg\max_{\alpha} \prod_{i=0}^{N} \gamma_{\alpha}^{\phi_{T_i}}(T_i|A_i). \tag{3}$$

Fig. 6 plots the comparative results for the test sets of Vietnamese and Lebanese versus native speakers as a function of the number of phonemes processed. [3] Accent classification based on various levels of position information (1. Core (C), 2. Proclitic and Enclitic (P,E), 3. Proclitic, Core and Enclitic (P,C,E), 4. no position information) are compared. Tables 5 and 6 compares results using Eqs. (2) and (3) for $N = 40$. Using position dependent information (Eq. (3)), consistently improves performance: English versus Vietnamese improves from an overall 86% to 93% correct classification. English versus Lebanese improves from 78% to 84% correct classification. [3] The plot shows that while both Core and Periphery information are important in acoustic matching of the achieved phoneme string to the target string, most of the speaker-independent information seems to be contained in the Core. As predicted, Lebanese accent identification is more difficult with this

[3] Three-way accent identification improves from 69% in the test set to 77% when using Eq. (3) instead of Eq. (2).

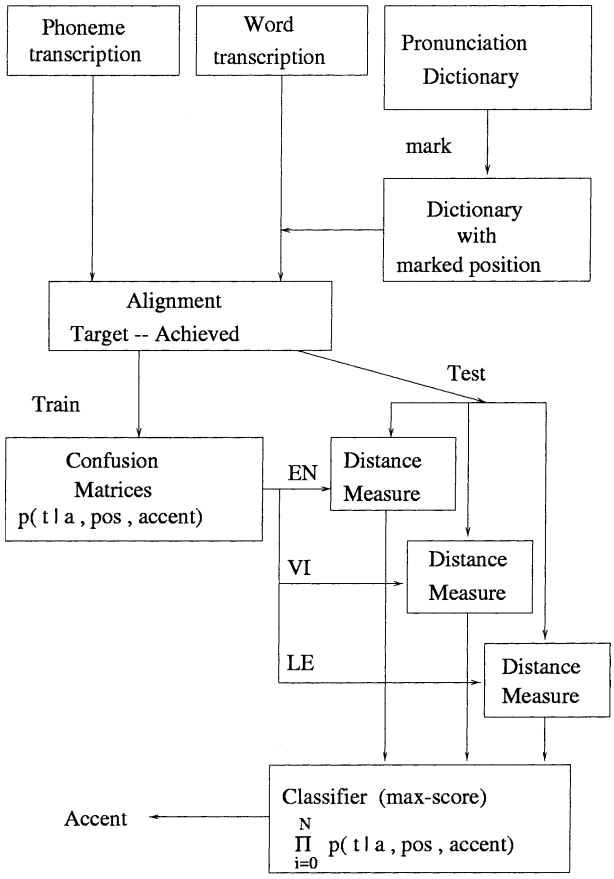

Fig. 5. Block diagram of accent identification system.

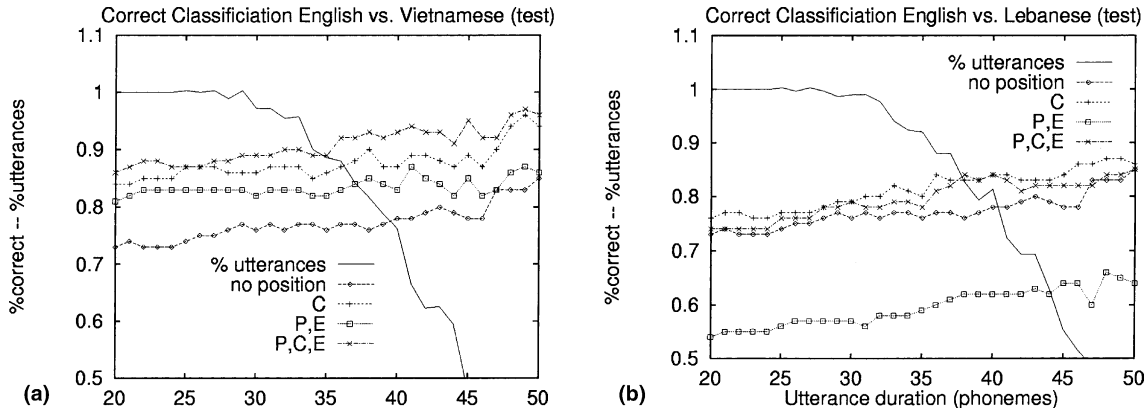

Fig. 6. (a) Hand labeled, English versus Vietnamese. (b) Hand labeled, English versus Lebanese. % correct classification using different combinations of information of C (Core), P (Proclitic) and E (Enclitic) or disregarding it. Also indicated is the % of test utterances of length N.

Table 5
Shows importance of location information of phoneme /d/ in Vietnamese accent

Position	Target	Achieved	English	Vietnamese
Confusions including /d/				
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 6
% correct accent identification after processing $N = 40$ phonemes[a]

Eq. (3) (Eq. (2))	Training set		Test set	
Hand labeled				
Input–output	EN	VI	EN	VI
EN	100 (100)	0 (0)	98 (96)	2 (4)
VI	3 (12)	97 (88)	13 (25)	87 (75)
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				
Input–output	EN	VI	EN	VI
EN	99 (97)	1 (3)	98 (97)	2 (3)
VI	31 (39)	69 (61)	40 (55)	60 (45)

[a] Results using Eq. (3), are compared to the baseline system (in parenthesis), using Eq. (2).

method than Vietnamese identification. It seems that features in the periphery are consistent across Vietnamese speakers, while Lebanese speakers are not consistent in their pronunciation patterns in the periphery. This may be the reason that, for Lebanese, using no position results in better performance than using only periphery information.

In order to study how well our theory might generalize from hand labeled to automatically aligned phonemes, we align a training and test set for Australian and Vietnamese accented English as defined in Section 2. Each of the automatically labeled phoneme strings was then analyzed in the same manner as the hand labeled 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 6 indicates that foreign accent identification for Vietnamese versus Australian English becomes possible 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 6 shows results for both accent groups. It is important to note here that the automatic labeling of phonemes can be considered a worst-case scenario. Ideally, this algorithm should be applied, when we have a well working system that has produced a phoneme string based on a hypothesis of what was said. In this case, the string was produced with some constraints which would

considerably improve the phoneme alignment. The hypothesis of what was said would be used to mark the positions of each of the phonemes in the achieved phoneme string. What this experiment shows, is that between a best-case (hand-labled) and a worst-case scenario (unconstrained automatic labeling) this algorithm is useful to a varying degree in capturing accent dependent information.

5. Discussion and future work

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. It can also be seen that the difference in syllable structure of the native language of a speaker compared to the non-native language has a direct influence on the degree of foreign accent. The examples discussed in this paper show the contrast in accent for Vietnamese (very different from English syllable structure) and Lebanese (much closer to English in syllable structure) clearly.

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. It would be desirable to implement this algorithm within a large vocabulary speech recognition system. More details about the algorithm, analysis and data are available upon request from the author (see also (Berkling et al., 1998)).

Acknowledgements

This work is the result of two consecutive postdoc positions at the University of Syndey, Australia and the Tokyo Institute of Technology, Japan.

References

Berkling, K., Cleirigh, C., Vonwiller, J., Zissman, M., 1998. Improving accent identification through knowledge of english syllable structure. In: Internat. Conf. Spoken Language Processing, Vol. 2, Sydney, Australia.

Cleirigh, C., 1998. A selectionist model of the gensis of phonic texture: systemic phonology and universal Darwinism. PhD thesis, Department of Linguistics, Sydney University.

Cleirigh, C., Vonwiller, J., 1994. Accent identification with a view to assisting recognition. In: Internat. Conf. Spoken Language Processing, Vol. 1, Yokohama, Japan.

Goldsmith, J., 1990. Autosegmental and Metrical Phonology, first ed., Blackwell, Basil.

Hansen, J., Arslan, L., 1995. Foreign accent classification using source generator based prosodic freatures. In: IEEE Internat. Conf. Acoustics, Speech, and Signal Processing, Vol. 1, Detroit, USA.

Kahn, D., 1980. Syllable-based generalizations in English phonology. PhD thesis, Massachussetts Institute of Technology.

Kumpf, K., King, R., 1997. Foreign speaker accent classification using phoneme-dependent accent discrimination models and comparisons with human perception benchmarks. In: Eurospeech, Vol. 4, Rhodos,Greece.

Mixdorff, H., 1996. Foreign accent in intonation patterns – a contrastive study applying a quantitative model of the f0 contour. In: Internat. Conf. Spoken Language Processing, Vol. 2, Philadelphia, USA.

Ostendorf, M. Byrne, B., Bacchian, M., Finke, M., Gunwardana, A., Ross, K., Roweis, S., Shirberg, E., Talkin, D., Waibel, A., Wheatley, B., Zeppenfield, T., 1996. Modeling systematic variations in pronunciation via languagedependent hidden speaking mode. In: Summer Workshop on Speech Recognition, Johns Hopkins University Baltmore, MD.

Teixeira, C., Trancoso, I., Serralheiro, A., 1997. Recognition of non-native accents. In: Eurospeech, Vol. 4, Rhodos, Greece.

Vonwiller, J., Rogers, I., Cleirigh, C., Lewis, L., 1995. Speaker and material selection for the Australian national database fo spoken languages. J. Quantitative Linguistics 2 (3), 177–211.