A Study of Temporal Features and Frequency Characteristics in American English Foreign Accent

Article *in* The Journal of the Acoustical Society of America · May 1999

DOI: 10.1121/1.419608 · Source: CiteSeer

2 authors, including:

Levent M. Arslan
Bogazici University
95 PUBLICATIONS **1,384** CITATIONS

SEE PROFILE

A Study of Temporal Features and Frequency Characteristics in American English Foreign Accent[§]

Levent M. Arslan and John H.L. Hansen
Robust Speech Processing Laboratory, Department of Electrical Engineering, Box 90291, Duke University, Durham, North Carolina 27708-0291

Technical Report RSPL-96-10

Abstract

In this paper, a detailed acoustic feature study of foreign accent using temporal features, intonation patterns, and frequency characteristics in American English is performed. Using a database which consists of words uttered in isolation, temporal features such as voice onset time, word-final stop closure duration, and characteristics of duration are investigated. Accent differences for native-produced versus Mandarin, German, and Turkish accented English utterances are analyzed. Of the dimensions considered the most important is found to be word-final stop closure duration. Mandarin accented English utterances show significant differences in terms of this feature when compared to native speaker utterances. In addition, the intonation characteristics across a set of foreign accents in American English is investigated. It is shown that Mandarin speaker utterances possess a larger negative continuative intonation slope than native speaker utterances, and German speaker utterances had a more positive intonation slope when compared to native speaker utterances. Finally, a detailed frequency analysis of foreign accented speech is conducted. It is shown that the mid-frequency range (1500-2500 Hz) is the most sensitive frequency band to non-native speaker pronunciation variations. Based on this knowledge a new frequency scale for the calculation of cepstrum coefficients is formulated which is shown to outperform the Mel-scale in terms of its ability to classify accent automatically among four accent classes.

Mail All Correspondence To:
Prof. John H.L. Hansen
Duke University
Robust Speech Processing Laboratory
Department of Electrical Engineering, Box 90291
Durham, North Carolina 27708-0291 U.S.A.

internet email: jhlh@ee.duke.edu Phone: 919-660-5256 FAX: 919-660-5293
http://www.ee.duke.edu/Research/Speech

---

*PACS Code 43.72Lw Language Acquisition

§Submitted May 1, 1996 to *The Journal of Acoustical Society of America.* Revised August 1, 1996. Revised again October 22, 1996.

1 Introduction

Foreign accent can be defined as the change in pronunciation patterns of a non-native speaker due to his or her first language background. Speakers of a second language often exhibit varying degrees of foreign accent traits based on a number of factors such as the age of second language learning and the length of residence in the second language speaking country (Asher and Garcia 1969; Leather, 1983; Flege, 1988; Flege and Fletcher 1992b). An understanding of the causes and acoustic properties of foreign accent can be quite useful in several areas such as speech synthesis, speech coding, and speech recognition. Defining the acoustic norms of American English in terms of pronunciation patterns may lead to more natural sounding speech output from text-to-speech systems. In addition, for some applications it may be desirable to produce speech with a specific regional or foreign accent. On the other hand, improved speech recognition performance can be achieved by incorporating accent information as a means of adapting to speaker differences. In a recent study, Arslan and Hansen (1996) showed that incorporating accent information into an isolated word speech recognition system can lead to substantial improvement in recognizer performance. In a separate study, Brousseau and Fox (1992) showed that improvement in the continuous speech recognition rate can be achieved by retraining on European French as opposed to Canadian French, or British English as opposed to American English.

There has been considerable research directed at understanding the causes and acoustic characteristics of foreign accent in English. In the text by Chreist (1964), a brief overview of sound problems in foreign accent is presented, where the issue of accent is regarded as a speech pathology problem. An alternative was considered in a study by Wells (1982) where the dialects in British English were examined from a linguistics point of view. From an historical perspective, these studies can be said to be more general in their treatment of accent. Recently, there have been more focused studies detailing the acoustic characteristics of foreign accent. For example, in a number of studies (Flege *et al.* 1980,1984a,1984b,1987,1988,1992a,1995; Port and Mitleb 1983; Munro 1993; Crowther and Mann 1992; Bohn and Flege 1992; Arslan and Hansen 1996) specific language accents were investigated in terms of their acoustic characteristics. It was shown that the second formant (F_2) is statistically the most significant resonance frequency in discriminating French accent from American accent for the French syllables /tu/ ('tous') and /ty/ ('tu'). In another study, Arslan and Hansen (1996) compared computer algorithm performance with that of human listener performance in the detection and classification of foreign accent based on isolated words. It was shown that a hidden Markov model (HMM) based computer algorithm could both detect and classify accent better than the average human listener for isolated word based acoustic features derived from utterances of native and non-native speakers.

In this paper, we consider a study of the temporal features, intonation patterns, and frequency characteristics of accented speech based on native-produced, Turkish, German, and Mandarin accents. In Section 2, we describe the database that was established at Duke University for analysis of foreign accent. In Section 3, an assessment of each temporal feature's ability to discriminate accent is made based on statistics generated from the accent database. The investigated features include voice onset time, word-final stop closure duration, average voicing, and average word duration. In Section 4, the differences between intonation patterns of native and non-native speaker utterances are analyzed. In Section 5, we present an analysis to determine which frequency bands are more sensitive to foreign accent. In addition, the validity of using the Mel-scale in accent classification is questioned, and a more appropriate scale for accent classification is proposed. Finally, a discussion of the results and conclusions are given in Section 6.

2 Accent Database

Based on an extensive literature review of foreign accent problem in American English, a test vocabulary was selected which contain a rich collection of phoneme class to phoneme class transitions. Particular attention was paid to select the set of words that were identified to be more problematic for non-native speaker production (Chreist, 1964). The chosen vocabulary consists of twenty isolated words, and four test sentences. These words and phrases are listed in Table 1. The data corpus was collected using a mixture of two environmental conditions including a head-mounted microphone in a quiet office environment and an on-line telephone interface (43 speakers used microphone input, 68 speakers used telephone input). The speakers were from the general Duke University community. All speech was sampled at 8 kHz and each vocabulary entry was repeated 5 times. Practice was not permitted before recording began. Available speech includes native-produced American English, and English under the following accents: German, Mandarin, Turkish, French, Persian, Spanish, Italian, Hindi, Rumanian, Japanese, Greek, and others. For the studies conducted here, the focus was on American English speech from 48 male[1] speakers between the ages of twenty and forty across the following four accents: native-produced, Turkish, Mandarin and German. In the evaluations, the microphone speech was bandpass filtered between 100 Hz and 3800 Hz in order to simulate the same telephone channel response, and thereby provide consistency in the database.

[1] In terms of frequency analysis it was necessary to separate male and female speakers. Unfortunately, there was not enough female speakers in the database in order to perform a similar acoustic analysis.

FOREIGN ACCENT DATABASE				
WORDS				
aluminum	catch	line	student	thirty
bird	change	look	target	three
boy	communication	root	teeth	white
bringing	hear	south	there	would
PHRASES				
This is my mother				
He took my book				
How old are you?				
Where are you going?				

Table 1: List of words and phrases that are included in the foreign accent database

3 Temporal Features

Studies have suggested that duration is an important suprasegmental feature in perception of foreign accent. A speaker often exhibits accent through hesitations, pauses, and the amount of time spent in producing or forming strings of different phonemes or phoneme classes. A number of studies have considered the analysis of temporal features in accented speech. For example, studies by Caramazza *et al.* (1973), Flege (1980,1984b), and Port and Mitleb (1983) showed that voice onset time is an important parameter in detecting the presence of French accent. Crowther and Mann (1992) investigated native language factors affecting use of vocalic cues to final consonant voicing. They found that Japanese and Mandarin speakers of English show less difference in F_1 offset frequencies in their tokens of *pod* compared to their tokens of *pot*. In another study, Byrd (1984) investigated the speaking rate differences among 8 regional dialects of American English in the TIMIT database[2]. Byrd found significant differences between the "Army Brat"[3] group and "South", and "Army Brat" and "South Midland" dialects. Part of the difference resulted from the frequency of pauses. Speakers from South Midland and South paused more often than expected, while speakers from the North Midland, West, and the "Army Brat" paused less often than expected given a random distribution.

In this study, we investigate voice onset time and word-final stop closure duration across native-produced, Turkish, Mandarin, and German accented English. Two of the language accents studied

[2]Detailed information about the TIMIT database is available through the Linguistics Data Consortium (LDC) (URL address: *http://www.cis.upenn.edu/~ldc/home.html*). Details can be found in Fisher *et al.* (1986).

[3]The American English group "Army Brat" refers to a person who has moved frequently across the U.S., and therefore may possess less of a regional dialect. This term is derived from U.S. military personnel who are normally moved frequently across the U.S.

(Turkish, Mandarin) do not possess voiced stops at the word-final position. In addition, Mandarin does not have unvoiced stops at the word-final position. These facts led us to choose word-final stop closure duration as one variable in our study. Voice onset time has been investigated by many researchers (Caramazza *et al.*, 1973; Flege, 1980; Flege and Hillenbrand, 1984b) for the study of foreign accent, and therefore was chosen as the second variable to analyze. In addition, average voicing duration and average word duration are investigated which are found to be affected to a large extent by accent related factors based on time-frequency analysis of the accent database. The study here is by no means intended to cover all aspects of foreign accent in English. We rather attempted to analyze a subset of acoustic features that are found to be significant in our analysis of the language accents considered here.

3.1 Word-Final Stop Closure Duration

Word-final stop closure duration has been investigated in several recent research studies. Port and Mitleb (1983), for example, showed that there are significant differences in consonant closure durations between word-final lax and tense stops produced by Arab and American speakers of English. In a separate study, Flege, *et al.* (1992a), reported significantly longer closure durations in final stops produced by Mandarin and Spanish speakers when compared to native English speakers.

An extensive analysis of word-final stop closure duration was conducted using selected entries from the accent data corpus. This included the following words: *would, bird, target, look, root, white.* In our analysis of time-frequency responses of native and non-native speaker utterances, we observed that the closure duration prior to the release of a stop consonant at the end of a word (word-final stop closure duration) is in general longer for non-native speakers than for native speakers. In Figure 1, spectrograms of the English word "would" from 8 different speakers are shown. The four spectrograms in Figure 1a belong to American speakers, and the four spectrograms in Figure 1b belong to Mandarin speakers. The closure before the release of the stop consonant /D/[4] is significantly longer (+55 ms on average) in duration for all Mandarin speakers. This result complies with a previous study by Flege, *et al.* (1992a) which indicates that Mandarin speakers produce the stop /D/ in bVd and sVd words with 30-40 ms longer than American speakers. The small difference in results may be due to different speaker population characteristics. Descriptive statistics of word-final stop closure durations are obtained across six words under the four accents (native-produced, Mandarin, German, and Turkish). In this analysis, a total of 12 male speakers from each accent group and 3 tokens of each word from each speaker

[4] In this study, uppercase ARPABET notation is used to describe phonemes for American English. See Deller, Proakis, and Hansen (1993) page 118 for a summary.

were used in estimating the statistics[5]. In order to assess the significance of differences among the four accent classes in terms of word-final stop closure duration, a one-way analysis of variance was performed. The results are summarized in Table 2. The word-final stop closure duration for the word *would* was found to be highly dependent on speaker accent. However, it should be noted here that the largest contribution to statistical significance of differences across the four language accents comes from voiced stops produced by Mandarin speakers. Therefore, pairwise statistical significance tests were performed where native English closure durations were compared to Turkish, German, and Mandarin closure durations. The results of these tests are also summarized in Table 2 with † symbol indicating ($p<0.05$), and †† indicating ($p<0.01$). Mandarin speakers had significantly longer stop closure durations for both word-final /T/ and /D/ sounds when compared to American speakers. However, for the /K/ sound in *look* the differences did not reach statistical significance. Both Turkish and German speakers had significantly longer closure durations for the voiced stop /D/ when compared to native speakers. However, for unvoiced stops such as /T/ and /K/ the closure durations for Turkish and German speakers were not significantly different from American speakers. These results verify the presence of a direct influence from the first language background of the non-native speaker to his/her accent, since the Mandarin language does not allow voiced or unvoiced word-final stops, whereas Turkish does not allow voiced stops in word-final position.

ANALYSIS OF VARIANCE RESULTS FOR WORD-FINAL STOP CLOSURE DURATION (ms)					
WORD	English	Mandarin	Turkish	German	F(3,140)
would	64.5 (13.7)	119.0 (34.1) ††	85.5 (24.0) ††	81.7 (31.1) ††	27.8 **
bird	64.1 (12.2)	104.0 (35.5) ††	94.9 (30.3) ††	76.2 (17.8) ††	18.7 **
target	89.8 (33.1)	118.3 (28.8) ††	85.8 (15.7)	80.5 (25.1)	19.2 **
look	124.4 (37.9)	138.0 (33.4)	111.5 (28.2)	108.2 (21.5) †	7.3 **
root	104.7 (33.7)	138.1 (28.0) ††	99.0 (16.9)	99.3 (30.7)	19.1 **
white	102.6 (34.0)	128.1 (30.7) ††	102.6 (14.5)	111.4 (29.0)	7.5 **

Table 2: Statistical analysis of word-final stop closure duration across accent groups. The mean and standard deviation (in parentheses) of word-final stop closure duration for each accent class across various words are listed. Null hypothesis represents the case where the averages of word-final stop closure durations are equal across four accents (native-produced, Turkish, German, Mandarin). (*: $p<0.05$, **: $p<0.01$). Pairwise ANOVA results comparing native-produced versus non-native accents are summarized as: †:$p<0.05$, ††:$p<0.01$

.

Among the stop consonants, non-native speakers consistently used longer closure duration for the

[5]Three out of five tokens of each speaker was selected for statistical analysis which are identified to have the most audible/visible release bursts for human operator labeling.

/D/ sound in *would* and *bird*. For the accents considered, Mandarin speakers were found to employ consistently longer closure duration for the investigated stop consonants.

In order to evaluate the discriminative power of word-final stop closure duration in terms of accent classification, the mean[6] and standard deviation are used to generate Gaussian probability density functions (PDF) for each word under each accent. Next, a maximum likelihood classifier is developed which employs the PDFs to make a decision as to which accent PDF results in the closest match for an input set of data samples. Finally, the true accent classes and accent classes obtained after employing the maximum likelihood classifier are compared to calculate the average accent classification rate based on the word-final stop closure duration. The closed set accent classification rate obtained following this approach is shown in Figure 2. The average classification rate using only word-final stop closure duration is 44.9%, which is significantly higher than the chance rate (25% for four accents). Accent classification performance over several words with final stops are also summarized. Consistent classification results were obtained for both voiced and unvoiced stops. Here, it should be emphasized that the classification rate is based on closed test data, and it should only be interpreted as a relative measure of confidence on the feature. This number will make more sense after it is compared to the classification rates obtained from voice onset time and average voicing duration features in the following sections. Based on this average classification rate and the statistical significance test results, it can be concluded that word-final stop closure duration is a useful discriminator of accented speech.

3.2 Voice Onset Time

In this section, we consider voice onset time as a potential accent discriminator. Voice onset time (VOT) represents the interval between the release burst of a leading stop consonant and the onset of voicing for a following vowel.[7] Caramazza and Yeni-Komshian (1974) found that French talkers produced the /T/ with VOT values that were significantly longer than the VOT measured in French words produced by French monolinguals. However, in an experiment conducted by Flege (1984a), the /T/ edited from French speaker utterances of the /tu/ and /ti/ syllables were 30 ms and 15 ms shorter than that of native English speakers on the average. This result was also consistent with other studies on foreign accent (Caramazza *et al.*, 1973; Flege, 1980; Flege and Hillenbrand, 1984b).

In this study, VOT values were investigated across four accents for a set of words which include a stop consonant in the initial position (*target, teeth, catch, communication*). In general, Turkish

[6] In order to better characterize the accented speech data, outliers in the data were removed in the estimation of the mean.

[7] VOT is normally measured from the stop release to the instant of voicing. As such, VOT is positive for unvoiced stops, and negative for voiced stops. Here, only unvoiced stops were considered for accent classification.

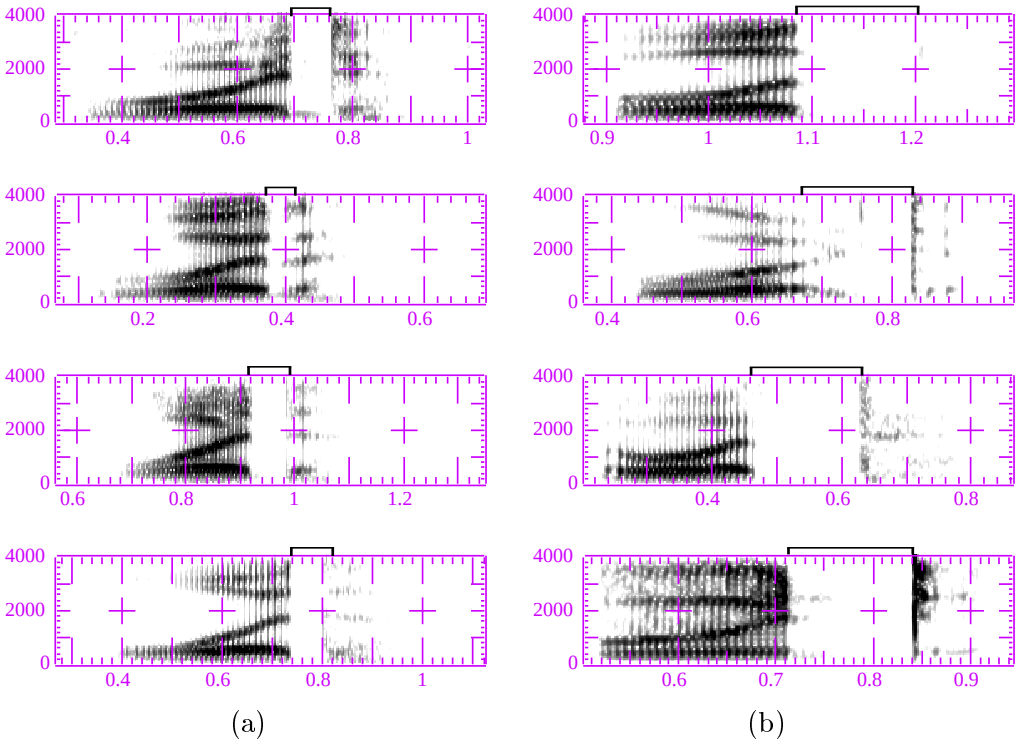

Figure 1: Illustration of the change of stop closure durations (i.e., stop closure durations are indicated by solid bars above spectrograms) for the word "would" due to Mandarin accent. (a) Spectrograms of 4 native speaker utterances, (b) Spectrograms of 4 Mandarin speaker utterances

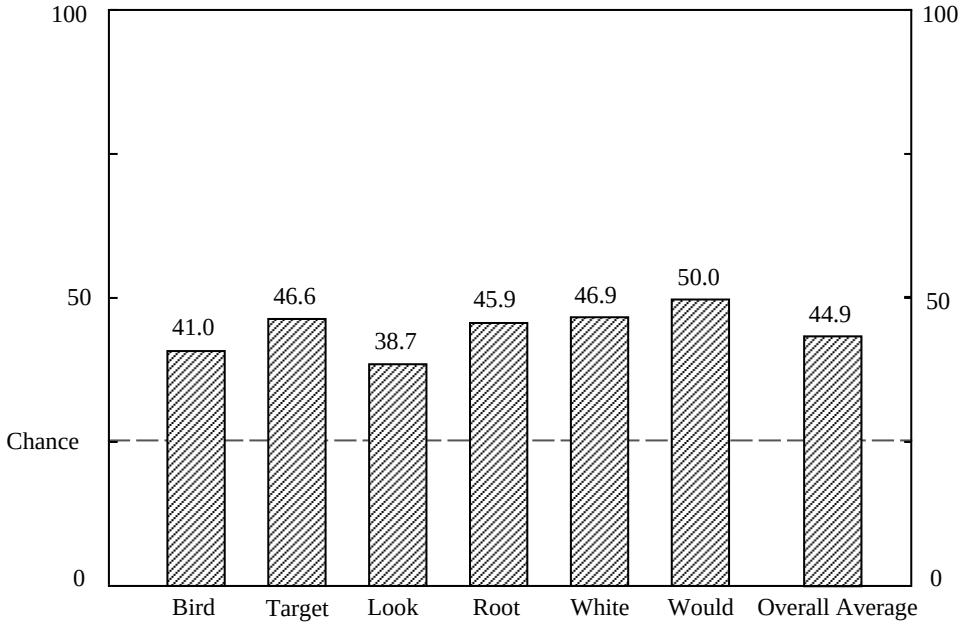

Figure 2: Closed set accent classification rates among 4 accent classes (native-produced, Mandarin, Turkish, German) based on the maximum likelihood estimate of word-final stop closure durations.

speakers demonstrated shorter voice onset times, whereas Mandarin and German speakers had longer voice onset times when compared to American speakers. However, as shown in Table 3, the statistics of VOT are not significantly different enough to discriminate reliably among the four accents. For the words *teeth* and *catch*, the average VOT differences across accent classes did not reach the minimum significance level ($p>0.05$). Pairwise statistical significance tests (also shown in Table 3) verified that, in general, voice onset time is not a significant discriminator among the accent classes considered in this study. The VOT value for the unvoiced /T/ in *target* was found to be significantly shorter for Turkish speakers (57 ms) when compared with American speakers (69 ms). In addition, the VOT value for the /K/ sound in *communication* was found to be significantly shorter for Mandarin speakers (57 ms) when compared to American speakers (65 ms). As a result, when the maximum likelihood classifier was employed for accent classification based on voice onset times for these words, an average classification rate of 32.0% was achieved in the closed set. The detailed statistical results across the four words are shown in Figure 3. Although this feature did not prove to be a strong indicator among the accent classes considered here, it may be more useful as a secondary feature in accent classification or potentially useful for other accent classes not considered here.

ANALYSIS OF VARIANCE RESULTS FOR VOICE ONSET TIME (ms)					
WORD	English	Mandarin	Turkish	German	F(3,140)
target	69.2 (20.1)	72.4 (22.1)	57.4 (16.3) ††	73.0 (28.5)	4.1 *
teeth	76.1 (18.4)	79.6 (23.9)	72.7 (23.1)	78.6 (22.5)	0.7
catch	70.5 (18.0)	73.1 (16.4)	69.9 (13.3)	80.6 (28.8)	2.3
communication	65.4 (19.1)	57.1 (14.6) †	63.3 (16.2)	69.3 (15.2)	3.7 *

Table 3: Statistical analysis of voice onset time across accent groups. Null hypothesis represents the case where the averages of voice onset times are equal across four accents (native-produced, Turkish, German, Mandarin). (*: p<0.05, **: p<0.01). Pairwise ANOVA results comparing native-produced versus non-native accents are summarized as: †:p<0.05, ††:p<0.01.

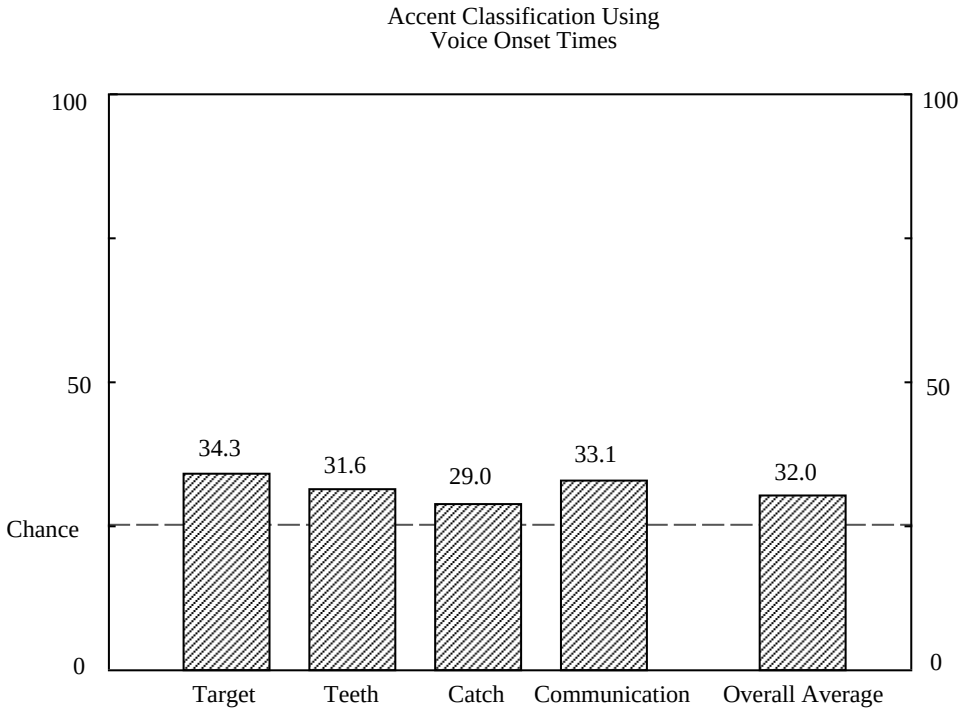

Figure 3: Closed set accent classification rates among the 4 accents based on a maximum likelihood classifier using voice onset time.

3.3 Average Voicing Duration

The third temporal feature investigated for accent discrimination is average voicing duration. There have been a number of studies on vowel duration as a perceptual cue to the voicing feature in word-final stops produced by native speakers of English (Raphael 1972; Repp, 1978; Wardrip-Fruin, 1982; Hogan and Rozsypal, 1980; Port and Dalby, 1982). Voicing duration is measured from the first positive peak in the periodic portion of voiced speech until where the voicing ends (i.e., the speech becomes unvoiced). The following eight words from the vocabulary set were selected for analysis: *bird, catch, change, hear, look, south, boy, teeth.* The statistics obtained here were based on the same speakers and accent classes as in the previous sections. Table 4 summarizes the statistical analysis of average voicing duration. Excluding the word *change*, all words considered showed statistically significant differences among accent classes in terms of their average voicing duration. Accent classification rates across the eight words after employing the maximum likelihood classifier are shown in Figure 4. Again we see consistent performance, with an average classification rate of 37.0% which is clearly above the chance probability. In general, non-native speakers spent longer times in voiced speech sections. In particular, average voicing duration for Mandarin speakers was consistently longer than for native speakers.

ANALYSIS OF VARIANCE RESULTS FOR AVERAGE VOICING DURATION (ms)					
WORD	English	Mandarin	Turkish	German	F(3,140)
bird	233.5 (48.8)	283.7 (53.5) ††	234.1 (46.2)	245.4 (43.8)	8.8 *
catch	137.9 (25.5)	179.3 (40.0) ††	148.6 (22.4)	150.5 (27.1) †	13.0 *
change	306.2 (47.8)	325.3 (65.7)	303.1 (42.4)	316.3 (51.2)	1.4
hear	273.8 (60.0)	325.7 (55.0) ††	259.4 (35.3)	325.3 (66.8) ††	14.2 *
look	194.8 (41.6)	220.0 (45.3) †	197.0 (43.0)	221.5 (45.2) ††	4.1 *
south	224.0 (39.2)	258.3 (48.1) ††	229.6 (31.4)	255.0 (37.4) ††	7.1 *
boy	322.8 (67.9)	383.6 (73.0) ††	318.7 (60.4)	346.6 (74.8)	6.7 *
teeth	163.2 (23.1)	216.6 (64.6) ††	205.0 (44.9) ††	195.4 (42.6) ††	9.0 *

Table 4: Statistical analysis of average voicing duration across accent groups. Null hypothesis represents the case where the averages of average voicing duration are equal across four accents (native-produced, Turkish, German, Mandarin). (*: $p<0.05$, **: $p<0.01$). Pairwise ANOVA results comparing native-produced versus non-native accents are summarized as: †:$p<0.05$, ††:$p<0.01$.

3.4 Average Word Duration

Next, the average word duration is analyzed among the four accents as an indicator of the speaking rate. Using the same speaker and accent set, the average word duration and its standard deviation for

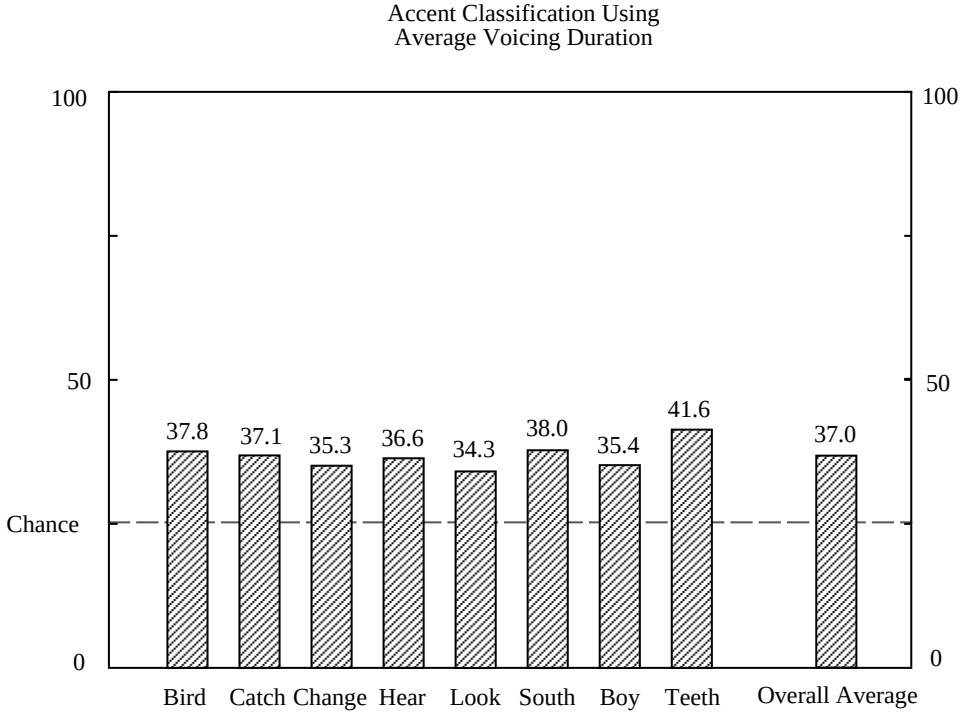

Figure 4: Closed set accent classification rates among the 4 accents based on the maximum likelihood estimate of average voicing duration.

the 20 words available in the database were obtained and tabulated in Table 5. Only the differences in average word duration for *bringing*, *change*, and *would* did not reach statistical significance out of the 20 words considered. Mandarin speakers had significantly slower phoneme production rates when average word duration was chosen as the criterion. Average word duration for Mandarin speakers was 813 ms, whereas it was 741 ms for native speakers. In contrast with Mandarin speakers, Turkish speakers were found to be using shorter duration (725 ms) for the production of word utterances when compared to native speakers. One possible explanation for this might be that Turkish speakers generally substitute shorter steady-state vowels for diphthongs in American English.

4 Intonation

When a native speaker listens to foreign accented speech, it is quite common to perceive a change in the intonation patterns. The role of intonation in English and other languages has been studied extensively (Bolinger 1958; Chomsky and Halle 1991; Leon and Martin 1980; Pilch 1970). In general, each language shows different intonation patterns depending on syntax, semantics, and phonemic structure. An experiment by Grover *et al.* (1987) verified that French, English and German speakers differ in the slopes (fundamental frequency divided by time) of their continuative intonation. In Figure 5, the

STATISTICS RELATED TO SPEAKING RATE DIFFERENCES AMONG ACCENTS (in ms)					
WORD	Native-Produced $\mu(\sigma)$	Mandarin $\mu(\sigma)$	Turkish $\mu(\sigma)$	German $\mu(\sigma)$	F (3,143)
aluminum	794(97)	894(123)	782(114)	867(102)	20.4 *
bird	671(123)	744(194)	659(112)	676(99)	4.9 *
boy	648(131)	727(120)	623(94)	662(101)	10.6 *
bringing	753(156)	783(153)	756(117)	771(132)	2.1
catch	742(111)	794(119)	691(88)	763(75)	4.9 *
change	828(106)	842(130)	763(97)	834(71)	2.3
communi.	1047(118)	1133(155)	1092(144)	1134(94)	5.2 *
hear	685(88)	773(124)	671(100)	754(84)	11.4 *
line	755(143)	783(105)	692(121)	776(96)	2.9 *
look	665(96)	702(139)	629(95)	714(93)	3.9 *
root	680(129)	798(176)	721(112)	795(94)	12.4 *
south	786(138)	814(140)	746(101)	814(87)	4.3 *
student	895(150)	994(151)	913(115)	967(123)	4.7 *
target	706(121)	874(158)	746(121)	790(101)	16.5 *
teeth	719(124)	768(162)	691(110)	717(82)	2.9 *
there	690(101)	776(172)	672(118)	732(112)	6.3 *
thirty	668(126)	800(163)	679(97)	734(105)	20.0 *
three	727(168)	762(156)	634(113)	723(89)	10.6 *
white	671(142)	787(141)	682(101)	729(92)	10.6 *
would	679(115)	704(165)	640(117)	680(83)	2.2
AVG.(ms)	741	813	725	782	

Table 5: Average word durations and standard deviations (in parentheses) across four accents for isolated words in the database. (*: statistically significant ($p<0.05$))

average intonation slopes among the four accents are shown. The slopes are calculated based on the following delta parameter computation:

$$\Delta f_0(l) = [\sum_{k=-K}^{K} k f_0(l-k)] \cdot G, \qquad 1 \leq n \leq P, \tag{1}$$

where l is the frame index, and G is a gain term chosen to make the variances of $f_0(l)$ and $\Delta f_0(l)$ equal.[8] Here, $\pm K$ represents a 75 ms delta pitch analysis window. It should be noted that the slope calculation is based on words produced in isolation in the accent database from the previous set of 48 male speakers. The results discussed here should not be extended to continuous speech, since other suprasegmental issues must be included. In fact, a study using continuous speech utterances could better represent intonation characteristics in a language accent. However, in this study the primary focus was on experiments using isolated speech utterances, and the results and conclusions of this study should be regarded in this context.

Based on the average intonation slopes and their variation across speakers, three important conclusions can be drawn:

- German speakers consistently exhibited more positive continuative intonation slopes than American speakers
- Mandarin speakers consistently exhibited more negative continuative intonation slopes than American speakers, which was due primarily to the sharp fall in pitch contour at the end of utterances.
- Turkish speakers showed much less variation in their intonation contours when compared to other accents.

The above results agree with previous findings that German speakers exhibit more positive intonation slope than American speakers, and that pitch information can be useful in determining the accent or language class of a speaker (Grover *et al.*, 1987). It should be noted that other factors also affect pitch characteristics such as a speaker's emotional state. In general, the pitch variance across words is quite large as would be expected for such a time-varying feature, which makes use of the maximum likelihood classifier very difficult. Therefore, although pairwise accent classification (i.e., German versus English or Mandarin versus English) may benefit from pitch information, when a larger set of accent classes are employed it is not expected to provide significant accent discrimination ability.

[8] Obtaining similar variances between actual parameters and delta parameters are important in HMM training

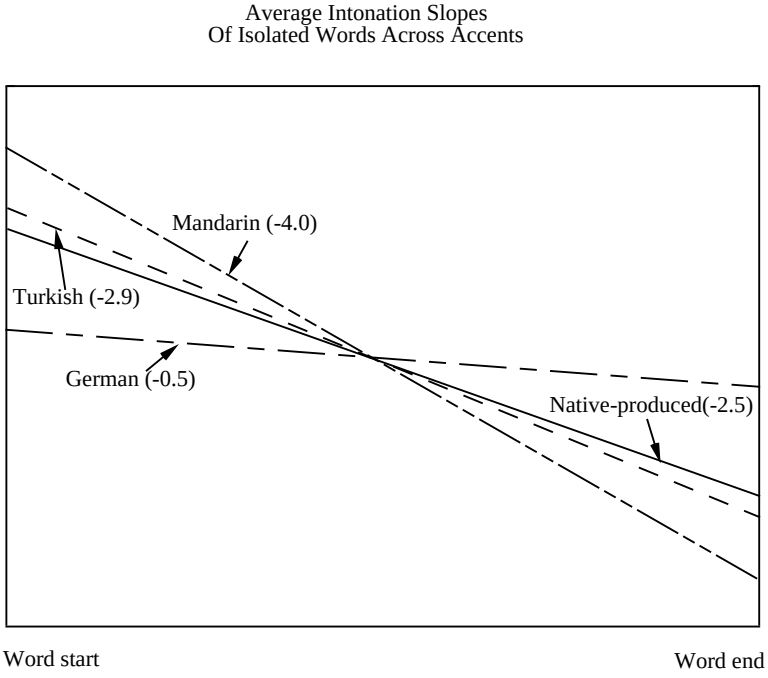

Figure 5: Average intonation slopes across the four accents based on the 20-word accent database.

5 Frequency Characteristics

In this next section, the focus for accent feature analysis shifts to frequency domain characteristics. Features that represent frequency characteristics of speech are commonly used in speech recognition systems. Although there has recently been significant improvement in speech recognition algorithm development, automatic speech recognition performance still lags far behind that achieved by humans. Therefore, in order to formulate better speech recognition algorithms, it may be beneficial to first consider aspects of the human auditory perception mechanism. Studies on psychoacoustic analysis of the human auditory perception mechanism have shown that the human ear responds differently to each acoustic tone based on their relative frequencies. Empirical evidence suggests that the human ear is more sensitive to low frequency signals. After extensive experimental analysis, the Mel-scale was formulated for the sampling of the frequency axis based on perceptual criteria (Koenig, 1949). Speech features derived using the Mel-scale have also resulted in superior speech recognition performance when compared to parameters obtained from a linear scale (Davis and Mermelstein, 1980).

In this section, our main argument is that the problem of accent classification is different than that experienced in speech recognition. Therefore care must be taken when applying standard speech recognition parameterization techniques to the problem of accent classification (the same could also be

said for speaker verification and language identification). It could be argued that a non-native speaker will focus his attention on speaking as close to an idealized native speech goal as possible based on "perception" of his own speech. As such, attempts would first be made to correct perceptually the most significant differences in pronunciation when compared to the native speaker pronunciation (e.g., what is typically done when students listen to teaching tapes of a new language). Therefore a parameter set which is based on perceptual criteria may not be the optimal feature set for the problem of accent classification. In light of this argument, a series of experiments were conducted in order to assess the statistical significance of various resonant frequencies and frequency bands for both speech recognition and accent classification. The following sections will discuss the experimental set-up followed by their results.

5.1 Formant Frequencies

In a previous study by Flege (1984a), the French syllables /tu/ ('tous') and /ty/ ('tu') produced in three speaking tasks by native speakers of American English and French talkers living in the United States were examined. Acoustic analysis revealed that the American talkers produced the /U/ sound with significantly higher F_2[9] and F_3 values than the French talkers. Fant (1970) performed a series of experiments in order to investigate the influence of the place of tongue constriction and the constriction area on formant frequencies. In Figure 6, the measurements of formant frequencies based on an electrical line analog (LEA[10]) of the vocal tract model are plotted. In the graph, the horizontal axis corresponds to the axial coordinate of the tongue constriction center. Each of the three curves represent formant frequencies corresponding to the area of constriction values ranging between 0.16 and 8.0 cm^2. Based on these curves it can be stated that a very small change in the place of the tongue constriction center or the cross sectional area at tongue constriction center can lead to large shifts in F_2 and F_3, whereas the remaining formants follow a more gradual change. Large shifts in the frequency location of F_1 can only be observed when the overall shape of the vocal tract is changing (e.g., an increasing vocal tract area as in /AA/ versus a decreasing vocal tract area in /IY/). In general, non-native speakers do not produce the same tongue movements as native speakers, since they automatically produce different sounds based on learned tongue movements of their native language. Therefore F_2 and F_3 play a bigger role in the discrimination of foreign accent.

In our analysis of frequency across the accent database, F_2 and F_3 contours of native speaker utterances were observed to be significantly different from that of non-native speaker utterances. When

[9] In this discussion, the first, second, third, and fourth formants will be represented as F_1, F_2, F_3, and F_4 respectively.

[10] For a more complete description of LEA speech synthesizer see Fant (1970) page 100.

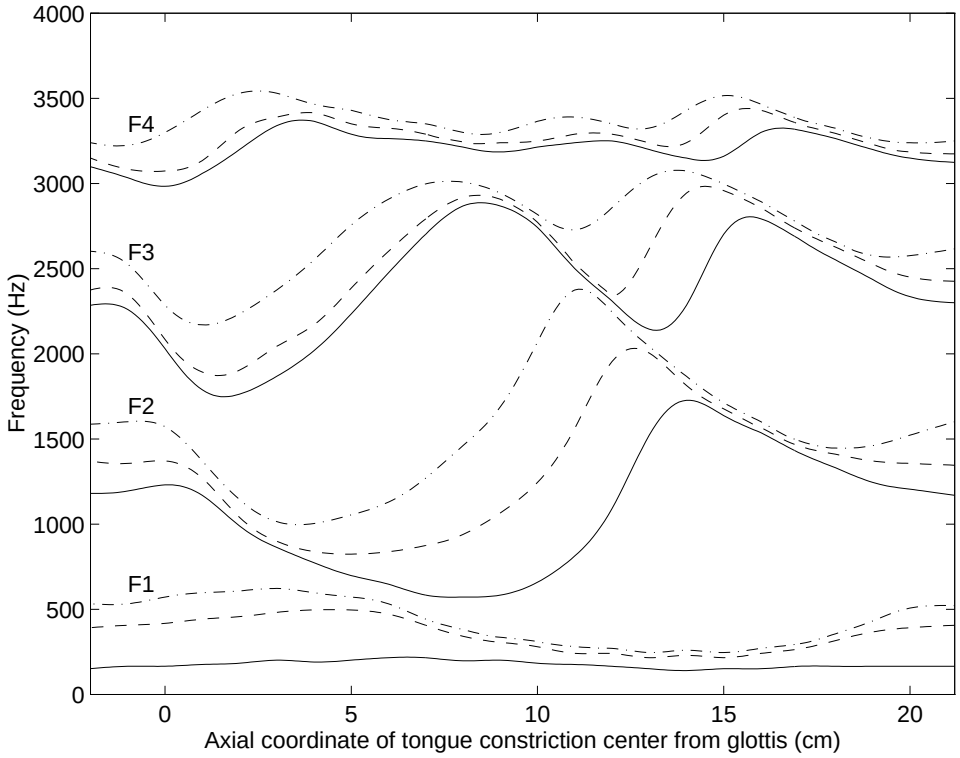

Figure 6: The influence of the place of tongue constriction and constriction area on the formant frequencies. Each of the three curves for each formant frequency represent the contours corresponding to different cross sectional areas at the place of constriction (-. : $A = 8.0cm^2$ - - : $A = 2.0cm^2$ — : $A = 0.16cm^2$). Adopted from Fant (1970).

time-frequency analysis of all the words was conducted for the four accents, we observed that the relative position of F_3 with respect to F_2 for the /ER/ sound was consistently different between native and non-native speakers. In Figure 7, a comparison between the spectrograms of native and non-native speakers for the /ER/ sound in *bird* is illustrated. For American speakers F_3 collapses into F_2 for the /ER/ sound (Figure 7a) which suggests early oral cavity closure resulting from the tip of the tongue touching the hard palate and sliding back. However, for some non-native speakers the tongue does not touch the hard palate until the very last moment in the production of the /ER/ sound, which causes some degree of separation between these two formant frequencies (Figure 7b).

A series of experiments were also performed in order to assess the relative significance of formant frequencies in the discrimination of accent. First, voiced sections of each word in the database were extracted. Next, the first four formant frequencies are estimated for each time frame. The formant frequency tracks were visually inspected and corrected after using a standard computer algorithm. For each formant frequency the derivative is estimated based on the delta parameter computation from Equation 1. A hidden Markov model (HMM) is obtained for each word in the database for each accent using one formant with its derivative at a time (e.g., a HMM is formed based on F_1 and delta F_1 parameters of the word *thirty* from the Turkish training speaker set). It is important to note here that when generating the models all the data that came from speakers sharing the same first language background is used without using subjective judgment of the level of accent exhibited by each utterance. Arslan and Hansen (1996b) attempted to address this issue by proposing a selective training algorithm to automatically detect the unreliable tokens in the data and reduce their weights in the model generation phase. Although improvements have been reported this issue still poses a big problem in automatic accent model generation. The HMM based accent recognizer using formant structure was evaluated. Open test accent classification results for the first four formant frequencies are shown in Figure 8a. Using the HMM set trained from American speakers, we evaluated speech recognition performance based on the 20-word vocabulary using a new (i.e., open) set of American speakers. In this case the open test speech recognition performance for each formant is shown in Figure 8b. Here, speech recognition and accent classification evaluation is conducted for each formant. When accent classification and speech recognition performance are compared, F_2 is found to be the most significant resonant frequency contributing to correct classification for both problems. However, F_1 which is known to be important in speech recognition (and demonstrated here) was not found to be as useful in accent classification. This result also supports our previous argument regarding the accented speech production mechanism, which states that previously developed perceptual based

critical frequency scales may not play such a significant role in differentiating between accent classes.

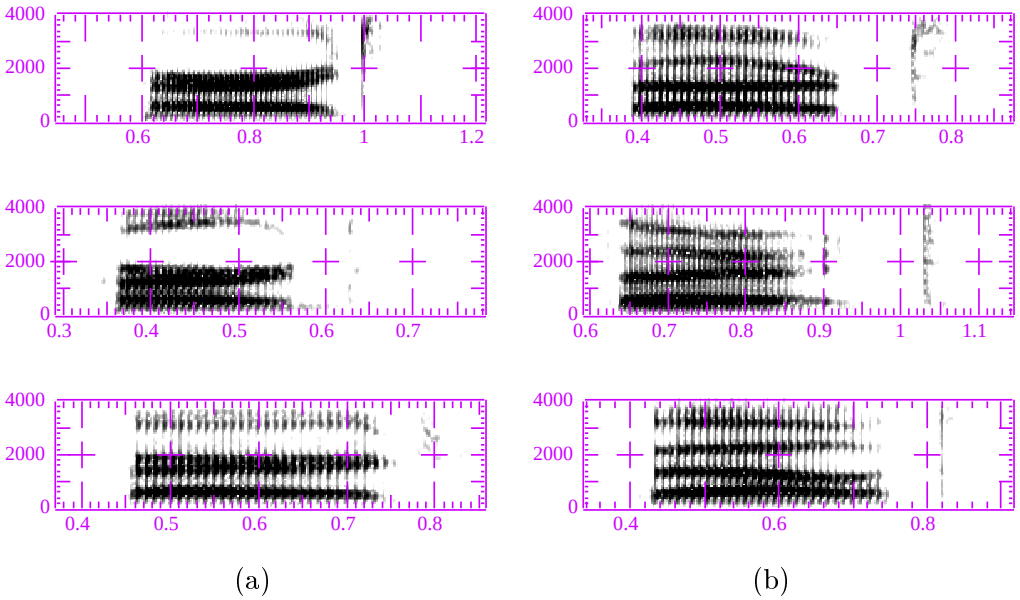

Figure 7: Illustration of the influence of accent on F_2-F_3 separation in /ER/ sound in *bird*: (a) three native speakers (b) three non-native speakers (from top to bottom, Turkish, Mandarin, German).

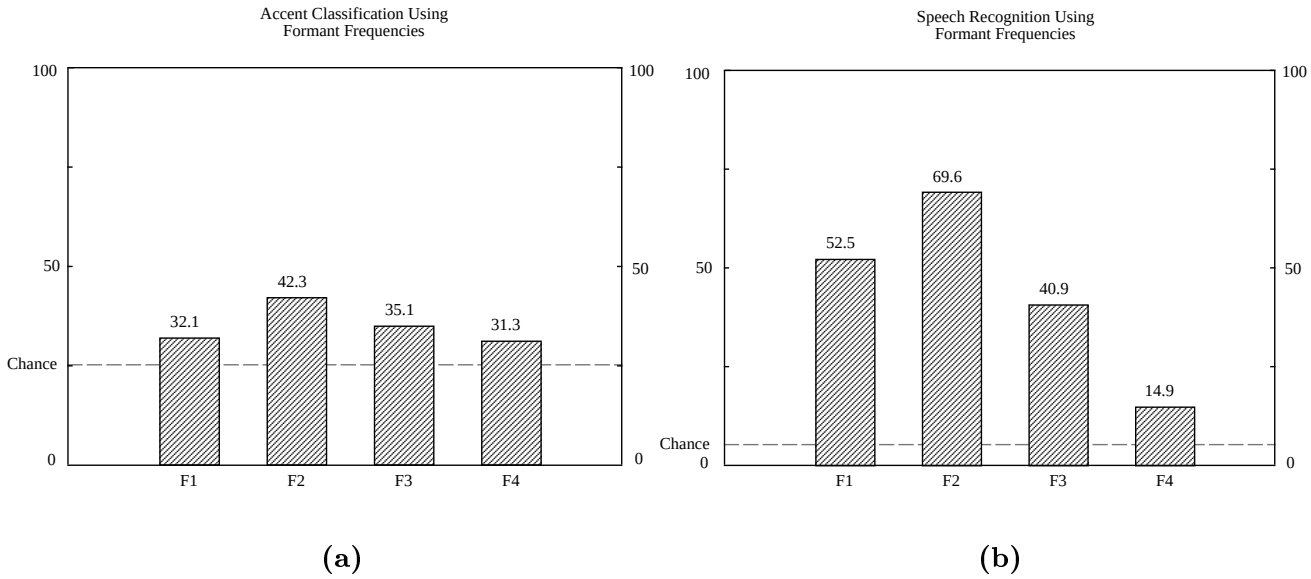

Figure 8: The influence of formant frequencies on the performance of (a) accent classification and (b) speech recognition.

5.2 Filter Banks

In automatic speech recognition systems, it is often difficult to work with formant frequencies as a standard parameter set because of the uncertainty involved with the estimation of the resonances. Moreover, the formant frequencies are not accurate representations of unvoiced sounds such as /F/

and /P/. Generally, a parameter set that represents the entire frequency band (e.g., filter bank coefficients or Mel-frequency cepstrum coefficients) is used in practice. In order to investigate the accent discrimination ability of various frequency bands, a series of experiments were performed. The frequency axis (0-4 kHz) was divided into 16 uniformly spaced frequency bands, as shown in Figure 9. Energy in each frequency band is weighted with triangular windows. Next, the output of each filter bank is used as a single parameter in generating an HMM for each word across the four accent classes. Using a single filter bank output as the input parameter, isolated word HMMs for native-produced, Turkish, Mandarin, and German accented English were generated via the Forward-Backward training algorithm. The HMM topology was a left-to-right structure with no state skips allowed. The number of states for each word was between 7 and 21 and was set proportional to the duration of each word. In the training phase, 11 male speakers from each accent group were used as the closed set and 1 male speaker from each accent group was set aside for open speaker testing. In order to use all speakers in the open test evaluations, a round robin training scenario was employed (i.e., the training simulations were repeated 12 times to test all 48 speakers under open test conditions).

In Figure 10, plot (a) shows accent classification performance across the 16 frequency bands. In order to compare accent classification performance to speech recognition performance across frequency bands, a second experiment was performed. Using only HMMs trained with native-produced English utterances obtained in the previous experiment, open set American speaker utterances were tested to establish speech recognition performance on the 20-word vocabulary. The speech recognition performance as a function of frequency is shown in Figure 10(b). From the graphs, it can be concluded that the impact of high frequencies on both speech recognition and accent classification performance is reduced. However, mid-range frequencies (1500-2500 Hz) contribute to accent classification performance to a greater extent than for speech recognition, whereas low frequencies improve speech recognition performance more than for accent classification. These results are consistent with those obtained with individual formant frequencies, since the F_2-F_3 range which was shown to be significant in accent discrimination roughly corresponds to the 1500-2500 Hz frequency range, and F_1 which was shown not to be as significant in accent discrimination corresponds to lower frequencies in Figure 10a.

The Mel-scale which is approximately linear below 1 kHz and logarithmic above (Koenig, 1949), fits suitably with that of speech recognition performance across frequency bands. However, it is not the most appropriate scale to use for accent classification. Therefore, a new frequency axis scale was formulated for accent classification which is shown in Figure 11. Since a larger number of filter banks are concentrated in the mid-range frequencies, the output coefficients are able to emphasize accent

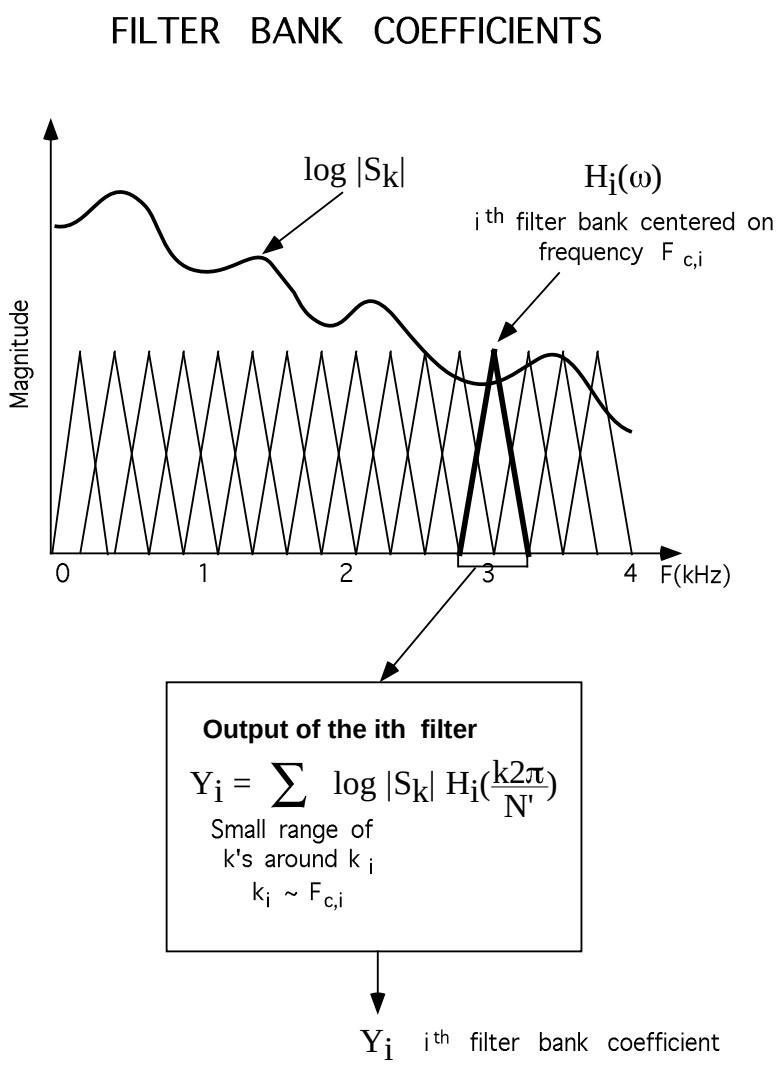

Figure 9: The extraction of filter bank coefficients.

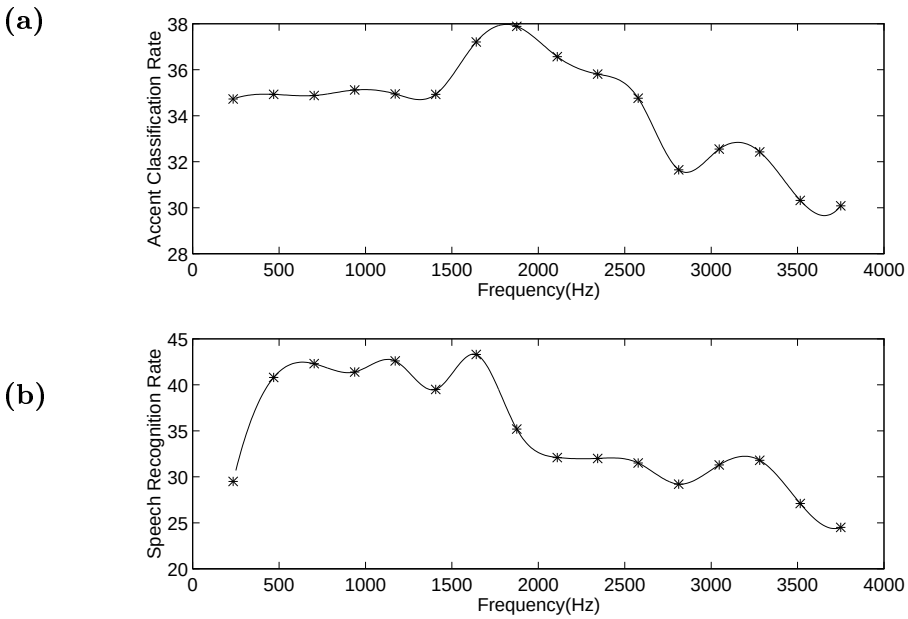

Figure 10: Comparison of accent classification versus speech recognition performance based on the energy in each frequency band.

sensitive features better. The 16 center frequencies of the filter bank which range between 0-4 kHz are also given in Figure 11. While this new frequency scale will prove to be useful for accent, it could be argued that the frequency based results in Figure 10 could be biased due to the chosen vocabulary or accent classes under consideration. Although this is a valid argument, the results presented here do show statistically significant performance improvement. Future studies may consider larger databases or a wider range of accent classes. Unfortunately, at the present time, no extensive database is available (such as the OGI multi-language or TIMIT available through LDC) in order to perform more extensive analysis across a wider range of accents. However, the major goal in this section on frequency is to show that accent classification is a different problem than speech recognition. Since we evaluated both accent classification and speech recognition performance on the same vocabulary set, respective rates based on each frequency band provide sufficient evidence to suggest that accent sensitive frequencies are somewhat different than perceptually motivated frequencies for recognition. The reason for the relatively poor accent classification rates at low frequencies can be explained by our previous hypothesis. Since the human auditory system is highly sensitive to low frequencies (Zwicker, 1990), non-native speakers concentrate on correcting the low-frequency characteristics in their speech. On the other hand, the frequency response in mid-range frequencies (1500-2500 Hz) differ from that of native speech, because the non-native speaker's auditory system is not as perceptually sensitive to changes in this frequency range. These frequencies (F_2-F_3 range) also represent detailed tongue movement which most

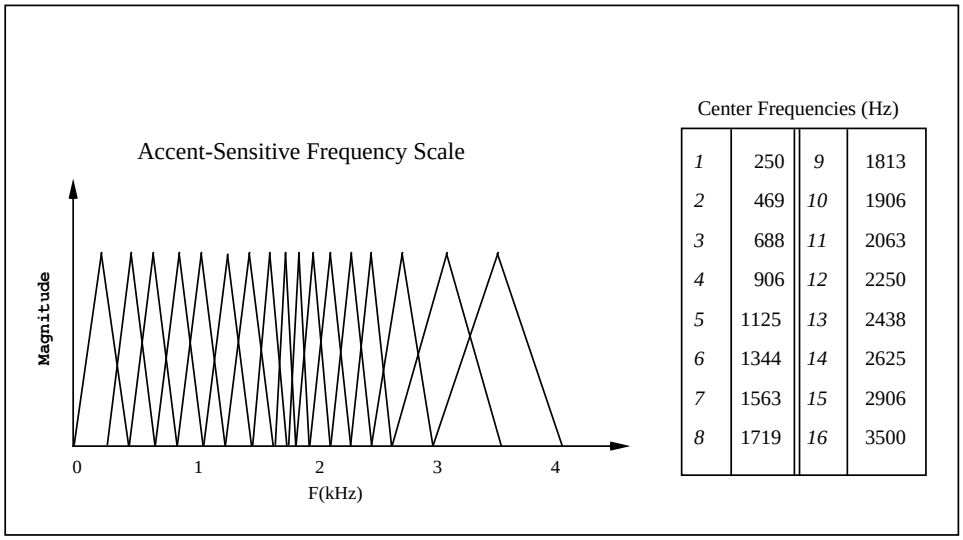

Figure 11: A new sampling scheme for the filter banks which is more sensitive to accent characteristics.

non-native speakers have difficulty in adopting.

In order to compare the performance of the accent-sensitive frequency sampling scheme to that of the Mel-scale and linear scales, an accent classification experiment was performed on the accent database. The following three sets of cepstrum coefficients were extracted from the isolated word utterances across the four accents (native-produced, German, Turkish, Mandarin): i) cepstrum coefficients derived from uniform sampling of the filter banks, ii) Mel-frequency cepstrum coefficients, and iii) cepstrum coefficients derived from accent-sensitive sampling of the filter banks. Using each parameterization approach, separate isolated word HMMs were generated using the Forward-Backward training algorithm. In order to reduce spectral bias, the long-term cepstral mean removal method is applied to each parameter set. The average accent classification rates for these parameter sets are shown in Table 6. The parameter set derived from the accent-sensitive scale resulted in the highest performance. It is also observed that the Mel-scale performed better than the linear scale for accent classification. When delta parameters as calculated using Equation 1 are added to the feature set, an increase in accent classification rate is obtained across all three frequency scales, while the same ordering of performance among the three parameter sets is retained. The improved results after addition of delta parameters are also shown in Table 6. The same frequency scale has been tested for a related task, language classification, and significant improvement is achieved on OGI multi-language database (Arslan, 1996).

COMPARISON OF DIFFERENT FREQUENCY SCALES IN TERMS OF ACCENT CLASSIFICATION PERFORMANCE			
	Linear Scale	Mel-Scale	Accent-sensitive
Accent Classification %	55.4	57.1	58.3
With Delta	60.0	60.7	61.9

Table 6: Comparison of the linear scale, Mel-scale, and accent-sensitive scales in terms of their accent classification performance among native-produced, Mandarin, Turkish, and German accented English.

6 Discussion and Conclusion

In this study, a detailed acoustic feature analysis of foreign accent in American English has been considered using temporal features, intonation patterns, and frequency characteristics. First, it was shown that word-final stop closure duration is a significant indicator of accent. This was especially true for Mandarin speakers who could be identified reliably by their usage of long closures before the release of stop consonants at the end of word utterances. Voice onset time, on the other hand, was not found to be as useful in the discrimination of accented utterances for the set of accent classes considered. In general, durational parameters at the segmental and word level were found to be significant features in the study of foreign accent. The slope of the intonation contour differed among the four accent classes studied, namely native-produced, Mandarin, Turkish, and German. In general, Mandarin speaker utterances had a more negative intonation slope than native speaker utterances, and German speaker utterances had a more positive continuation slope than native speaker utterances.

In terms of frequency analysis of foreign accent a number of significant results were obtained. The motivation for the study of frequency analysis was due to our belief that a parameter set designed based on perceptual criteria for speech recognition would not be the best parameter set for the problem of accent classification. This is because non-native speakers concentrate their efforts highly on following native speaker pronunciation patterns, and in order to accomplish this they rely on feedback from their auditory system. The auditory perception mechanism, in turn, is highly sensitive to changes in low frequencies (perceptually significant frequencies). Therefore, speakers attempt to correct the low-frequency component of their utterances. With minor changes in their tongue movements, they can accomplish this correction to a large extent, but major tongue movements which are a learned habit through years of their first language experience are not as easy to modify. Therefore, the frequencies in the F_2-F_3 range should be the most sensitive frequencies to assess accent characteristics. In order to test this hypothesis, formant frequencies were first analyzed in terms of their ability in discrimi-

nating accented speech from native speech. In general, the second formant was found to be the most significant indicator of foreign accent characteristics. An additional experiment employing filter banks to identify accent sensitive frequency bands was also conducted, and similar results to that found in formant frequency analysis were obtained. The frequencies in the 1500-2500 Hz range were shown to be the most important frequencies based on accent classification performance. Finally, a new frequency sampling scheme was proposed and evaluated in the calculation of cepstrum coefficients in place of the commonly used Mel-scale. Consistent improvement over the Mel-scale was obtained through the use of the proposed accent-sensitive frequency scale with or without the inclusion of delta coefficients in the parameter set. In conclusion, this study has shown that a variety of both temporal and frequency domain characteristics are modified when accent is present in American English, and that these issues can be used to formulate effective accent classification algorithms.

References

Asher, J., and Garcia, G. (**1969**). "The optimal age to learn a foreign language," *Modern Language Journal,* **38**:334–341.

Arslan, L.M. **(1996)**. *Foreign Accent Classification in American English.* PhD thesis, Duke University.

Arslan, L.M., and Hansen, J.H.L. **(1996)**. "Language Accent Classification in American English". *Speech Communication.* **18**(4):353–367.

Arslan, L.M., and Hansen, J.H.L. **(1996b)**. "Improved HMM Training and Scoring Strategies". *IEEE Proc. ICASSP*, vol. 2, pp. 589-592, Atlanta, USA, May 1996.

Bohn, O., and Flege, J.E. **(1992)**. "The Production of New and Similar Vowels by Adult German Learners of English". *Studies in Second Language Acquisition*, **14**(2):131–158.

Bolinger, D. **(1958)**. "A Theory of pitch accent in English". *Word*, **14**:109–149.

Brousseau, J., and Fox, S.A. **(1992)**. "Dialect-dependent Speech recognizers for Canadian and European French". *Proc. Inter. Conf. on Spoken Language Processing*, pp.1003-1006.

Byrd, D. **(1994)**. "Relations of sex and dialect to reduction". *Speech Communication*, **15**:39–54.

Caramazza, A., Yeni-Komshian, G., Zurif, E., and Carbone, E. **(1973)**. "The acquisition of a new phonological contrast: The case of stop consonants in French-English bilinguals ". *J. Acoust. Soc. Am.*, **54**:421–428.

Caramazza, A., and Yeni-Komshian, G. **(1974)**. "Voice onset time in two French dialects". *J. Phonet.*, **2**:239–245.

Chomsky, N., and Halle, M. **(1991)**. *The Sound Pattern of English.* MIT Press, Cambridge, MA.

Chreist, F. **(1964)**. *Foreign Accent.* Prentice-Hall, Englewood Cliffs, N.J.

Crowther, C.S., and Mann, V. **(1992)**. "Native language factors affecting use of vocalic cues to final consonant voicing in English". *J. Acoust. Soc. Am.*, **92**(2):711–723.

Davis, S., and Mermelstein, P. **(1980)**. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". *IEEE Trans. on Acoustics, Speech, and Signal Processing*, **ASSP-28**(4):357–366.

Deller, J., Proakis, J., and Hansen, J.H.L. **(1993)**. *Discrete Time Processing of Speech Signals*,

Macmillan Series for Prentice-Hall, New York.

Fisher, W., Doddington, G., and Goudie-Marshall, K. **(1986)**. "The DARPA speech recognition research database: Specifications and status". In *Proceedings of the DARPA Speech Recognition Workshop*, pp. 93–99.

Flege, J.E. **(1980)**. "Phonetic approximation in second language acquisition". *Language Learning*, **30**:117–134.

Flege, J.E. **(1984a)**. "The detection of French accent by American listeners". *J. Acoust. Soc. Am.*, **76**:692–707.

Flege, J.E., and Hillenbrand, J. **(1984b)**. "Limits on pronunciation accuracy in adult foreign language speech production". *J. Acoust. Soc. Am.*, **76**(3):708–721.

Flege, J.E., and Eefting, W. **(1987)**. "Cross-language switching in stop consonant perception and production by Dutch speakers of English". *Speech Communication*, **6**(3):185–202.

Flege, J.E. **(1988)**. "Factors affecting degree of perceived foreign accent in English sentences". *J. Acoust. Soc. Am.*, **84**:70–79.

Flege, J.E., Munro, M., and Skelton, L. **(1992a)**. "Production of the word-final English /t/-/d/ contrast by native speakers of English, Mandarin, and Spanish". *J. Acoust. Soc. Am.*, **92**(1):128–143.

Flege, J.E., and Fletcher, K.L. **(1992b)**. "Talker and listener effects on degree of perceived foreign accent". *J. Acoust. Soc. Am.*, **91**:370–389.

Flege, J.E., Munro, M., and MacKay, I. **(1995)**. "Effects of age of second-language learning on the production of English consonants". *Speech Communication*, **16**(1):1–26.

Grover, C., Jamieson, D., and Dobrovolsky, M. **(1987)**. "Intonation in English, French and German: perception and production". *Language and Speech*, **30**(3):277–295.

Harmegnies, B., Delplancq, V., Esling, J., and Bruyninckx, M. **(1994)**. "Effects of Deliberate Changes of Global Quality on the Vocal Signals in English and French". *Revue de Phonetique Appliquee*, **111**:139–153.

Hogan, J., and Rozsypal, A. **(1980)**. "Evaluation of vowel duration as a cue for the voicing distinction in the following word-final consonant". *J. Acoust. Soc. Am.*, **67**:1764–1771.

Leather, J. **(1983)**. "Second language pronunciation and language teaching". *Language Teaching Abstracts*, **16**(3):198–223.

Léon, P., and Martin, P. **(1980)**. *The Melody of a Language*, chapter Des Accents, 177–186. University Park Press, Baltimore.

Munro, M. J. **(1993)**. "Productions of English Vowels by Native Speakers of Arabic: Acoustic Measurements and Accentedness Ratings". *Language and Speech*, **36**(1):39–66.

Pilch, H. **(1970)**. "The elementary intonation contour of English: A phonemic analysis". *Phonetica*, **22**:82–111.

Port, R., and Dalby, J. **(1982)**. "Consonant/vowel ratio as a cue for voicing in English". *Percep. Psychophys.*, **32**:141–152.

Port, R., and Mitleb, F. M. **(1983)**. "Segmental Features and Implementation in Acquisition of English by Arabic Speakers". *Journal of Phonetics*, **11**(3):219–229.

Raphael, L. **(1972)**. "Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English". *J. Acoust. Soc. Am.*, **51**:1296–1303.

Repp, B. **(1978)**. "Perceptual integration and differentiation cues for intervocalic stop consonants". *Percep. Psychophys.*, **24**:471–485.

Wardrip-Fruin, C. **(1982)**. "On the status of temporal cues to phonetic categories: Preceding vowel duration as a cue to voicing in final consonants". *J. Acoust. Soc. Am.*, **71**:187–195.

Wells, J. **(1982)**. *Accents of English*. Cambridge University Press, England.

Zwicker, E., and Fastl, H. **(1990)**. *Psychoacoustics: Facts and Models*. Springer-Verlag, Berlin.