L. Arslan, J.H.L. Hansen, "Language Accent Classification in American English," Speech Communications, vol. 18(4), pp. 353-367, July 1996.

Language Accent Classification in American English

Levent Arslan and **John H.L. Hansen**

Robust Speech Processing Laboratory
Center for Spoken Language Research
Erik Jonsson School of Engineering & Computer Science Department of Electrical Engineering
The University of Texas at Dallas
P.O. Box 830688, EC33
Richardson, TX 75083-0688
972 – 883 – 2910 (Phone) 972 – 883 - 2710 (Fax)
John.Hansen@utdallas.edu (email)

Speech Communication, vol. 18(4), pp. 353-367, July 1996.

ELSEVIER

Speech Communication 18 (1996) 353–367

SPEECH COMMUNICATION

Language accent classification in American English [1]

Levent M. Arslan, John H.L. Hansen *

Robust Speech Processing Laboratory, Department of Electrical Engineering, Box 90291, Duke University, Durham, NC 27708-0291, USA

Received 28 August 1995; revised 26 February 1996

Abstract

It is well known that speaker variability caused by accent is one factor that degrades performance of speech recognition algorithms. If knowledge of speaker accent can be estimated accurately, then a modified set of recognition models which addresses speaker accent could be employed to increase recognition accuracy. In this study, the problem of language accent classification in American English is considered. A database of foreign language accent is established that consists of words and phrases that are known to be sensitive to accent. Next, isolated word and phoneme based accent classification algorithms are developed. The feature set under consideration includes Mel-cepstrum coefficients and energy, and their first order differences. It is shown that as test utterance length increases, higher classification accuracy is achieved. Isolated word strings of 7–8 words uttered by the speaker results in an accent classification rate of 93% among four different language accents. A subjective listening test is also conducted in order to compare human performance with computer algorithm performance in accent discrimination. The results show that computer based accent classification consistently achieves superior performance over human listener responses for classification. It is shown, however, that some listeners are able to match algorithm performance for accent detection. Finally, an experimental study is performed to investigate the influence of foreign accent on speech recognition algorithms. It is shown that training separate models for each accent rather than using a single model for each word can improve recognition accuracy dramatically.

Zusammenfassung

Es ist bekannt, daß die durch den Akzent verursachte Sprechervariabilität einer der Faktoren ist, die die Leistungsfähigkeit von Spracherkennungsalgorithmen vermindern. Wenn der Akzent eines Sprechers genau bestimmt werden kann, kann eine veränderte Sammlung von Erkennungmodellen, die den Sprecherakzent berücksichtigt, eingesetzt werden, um die Erkennungsrate zu erhöhen. In dieser Studie wird das Problem der Sprachakzentklassifikation im amerikanischen Englisch behandelt. Es wird eine Datenbank von fremdsprachlichen Akzenten erstellt, die aus Wörtern und Sätzen besteht, von denen bekannt ist, daß sie auf verschiedene Akzente empfindlich reagieren. Danach werden einzelwort- und phonembasierte Algorithmen zur Akzentklassifikation entwickelt. Die dabei betrachtete Menge von Sprachfaktoren umfaßt Mel-Cepstrum-Koeffizienten und Energie sowie deren erste Ableitungen. Es wird gezeigt, daß mit steigender Äußerungslänge die

* Corresponding author. E-mail: jhlh@ee.duke.edu; http://www.ee.duke.edu/Research/Speech .

[1] Audiofiles available. See http://www.elsevier.nl/locate/specom .

PII S0167-6393(96)00024-6

Klassifikationsgenauigkeit steigt. Die Verwendung von Einzelwortketten von 7 bis 8 Wörtern eines Sprechers erzielt eine akzentklassifikationsrate von 93% unter vier verschiedenen Sprachakzenten. Es wird auch ein subjektiver Hörtest durchgeführt, um die menschliche mit der algorithmischen Leistungsfähigkeit bei der Akzentunterscheidung zu vergleichen. Die Ergebnisse zeigen, daß die computerbasierte Akzentklassifikation in konsistenter Weise eine höhere Leistungsfähigkeit aufweist als die menschlichen Klassifikationsantworten. Es wird jedoch gezeigt, daß einige Hörer die gleichen Leistungen bei der Akzentererkennung erreichen wie die algorithmische Erkennung. Zuletzt wird eine experimentelle Studie durchgeführt, um den Einfluß von fremdsprachlichem Akzent auf Sprachererkennungsalgorithmen zu untersuchen. Es wird gezeigt, daß das Trainieren von getrennten Modellen für jeden Akzent gegenüber der Verwendung eines einzelnen Modells für jedes Wort die Erkennungsgenauigkeit dramatisch erhöhen kann.

Résumé

Il est bien connu que la variabilité inter-locuteur liée à l'accent est un facteur important de la dégradation des performances des systèmes de reconnaissance. Si l'on peut faire une estimation précise de l'accent d'un locuteur, alors un ensemble de modèles de reconnaissance modifiés pour prendre en compte cet accent peut être utilisé pour améliorer les scores de reconnaissance. Dans cette étude, on traite de la question de l'identification de l'accent en Anglais Américain. Une base de données d'accents étrangers a été établie: elle comporte des mots et des syntagmes connus pour être sensibles à l'accent. Des algorithmes de classification d'accent ont ensuite été développés, basés sur des mots isolés ou sur des phonèmes. L'ensemble des traits considérés comprend les coefficients cesptraux en Mels et l'énergie, ainsi que leurs dérivées du premier ordre. On montre que la précision de la classification augmente avec la longeur de la phrase test. Des séquences de 7 à 8 mots isolés donnent lieu à un taux de classification correcte d'accent de 93%, dans un ensemble de quatre type d'accents différents. Un test d'écoute subjectif a également été mené pour comparer les performances humaines et automatiques sur cette tâche. Les résultats montrent que la classification automatique donne, de façon cohérente, des performances supérieures à celles des réponses humaines pour cette tâche de classification. Toutefois, on montre que certains auditeurs sont capables d'égaler les performances des algorithmes pour la détection d'accent. Enfin, une étude expérimentale a été menée pour étudier l'influence de l'accent étranger sur la reconnaissance. On montre que la précision de la reconnaissance peut être améliorée de façon notable en faisant un apprentissage séparé des modèles pour chaque accent plutôt qu'en utilisant un modèle unique pour chaque mot.

Keywords: Accent analysis; Accent classification; Hidden Markov models; Robust speech recognition; Listener accent assessment

1. Introduction

Foreign accent can be defined as the patterns of pronunciation features which characterize an individual's speech as belonging to a particular language group. In general, individuals who speak languages other than their own identify themselves as non-native speakers by the voluntary and/or involuntary appearance of certain pronunciation patterns. If a speaker acquires a second language at an early age, his ability to minimize these accent factors improves. In addition, most speakers of a second language improve their ability to converse with a reduction of accent traits as they gain more experience (i.e., the length of time a person speaks the second language) (Asher and Garcia, 1969).

1.1. Accent classification versus language identification

Language identification in telecommunications has received much attention recently (House and Neuberg, 1977; Zissman, 1993, 1995; Muthusamy et al., 1994; Berkling and Barnard, 1994a,b; Hazen and Zue, 1994). A related problem which has not been explored in detail is foreign accent classification. In this section we will discuss the similarities and differences between the two problems.

Every individual develops a characteristic speaking style at an early age which will depend heavily on his language environment (i.e., the native language spoken). When learning a second language, the speaker will carry traits of this style into the new

language. Therefore, many features of his native language will persist in his speech. As an example, for unfamiliar phonemes in the second language, the speaker will substitute more commonly used phonemes available in his native language. These phonemes can be relayers of both accent and language. For language identification, much success can be gained by considering phoneme concentration, phoneme positioning, and of course word and sentence content. Accent classification assumes that speakers are intentionally speaking the same language. The level of accent exhibited in second language pronunciation will depend on a number of speaker related factors such as (i) the age at which a speaker learns the second language, (ii) the nationality of the speaker's language instructor, and (iii) the amount of interactive contact the speaker has with native talkers of the second language. In this respect, accent classification is a more challenging problem than language identification since there are no clear boundaries among accent classes. For example, a person may be judged as having a slight German accent, or a heavy German accent. However, the language identification decision is always binary (i.e., the speaker either produces speech using the German language or not).

Some distinct language dependent prosodic features may not be good relayers of accent, since a person is normally taught to deemphasize those features which are perceptually more audible when learning a second language. However, there may be other clues present in the speech signal for detection of accent such as hesitation, pause duration between words, and others. Hesitation between words is typically due to the fact that early second language learners normally study vocabulary lists, and normally will not have as much experience in speaking the given word in context.

1.2. Why classify accent?

It is well known that accent is one of the most important factors next to gender that influence speaker independent recognition algorithms (Hansen and Arslan, 1995; Gupta and Mermelstein, 1982; Rabiner and Wilpon, 1977). Currently, most recognition algorithms are gender dependent, with no accent information utilized. In order to motivate the problem of accent classification, we wish to determine the influence of accent on recognition performance. An experiment was conducted using a 20-word isolated speech database. A hidden Markov model (HMM) recognizer was trained using five tokens of each word, from 11 speakers of American English (in Section 2.1, a more complete discussion of the accent database is presented). The recognizer was tested using 12 separate native speakers of American English, 12 Turkish, 12 German and 12 Chinese speakers of English. The open set recognition rate for American speakers was 99.7%, whereas it was 92.5% for Turkish speakers, 88.7% for Chinese speakers, and 95.3% for German speakers. While a more detailed discussion of the experiment is given in Section 4.4, the results here clearly show that the presence of accent impacts overall recognition performance.

This experiment therefore demonstrates the clear need for the formulation of speech recognition algorithms which are less sensitive to accent. To accomplish this, we propose to develop an accent classification algorithm which in turn, can extract the necessary information to direct an accent dependent recognition system. Such a scheme could improve automatic speech recognition in aircraft cockpits or communication scenarios involving multi-national speakers (e.g., multi-national United Nations exercises).

1.3. Overview and background

Second language learning requires a speaker to develop a modified set of patterns for intonation, stress and rhythm. In addition, an alternate collection of phonemes may also be necessary for proper speech language production. The role of intonation in foreign accent has been studied extensively. An experiment verified that French, English and German speakers differ in the slopes (fundamental frequency divided by time) of their intonation contours (Grover et al., 1987). Flege (1984) performed experiments to test the detectability of native listeners for French accented speakers. Detection rates between 63% and 95% were obtained. It was also shown that the listener's detection performance was not influenced by whether the speech was read in isolation, or produced in a spontaneous story. Another interesting result was that speech of even 30 ms (roughly, the

release burst of the phoneme /T/) was enough for some listeners to detect the accent.

Accent is also a challenging research problem in speech recognition. Gupta and Mermelstein (1982) showed that the recognition rate for French accented speakers of English was lower than that for native English speakers. Studies have also been conducted which attempt to normalize the influence of regional accent prior to speech recognition. In one study, Barry et al. (1989) proposed a two stage procedure for British English accent normalization. In stage one, an accent classification procedure selects one of four gross regional English accents on the basis of vowel quality differences within four calibration sentences. In stage two, an adjustment procedure shifts the regional reference vowel space onto the speaker's vowel space as calculated from the accent identification data.

In a recent study, Waibel investigated the use of a variety of prosodic features in speech recognition (Waibel, 1988). He showed that prosodic and phonetic information are complementary, and that improved speech recognition performance could be attained by combining the two sources of information. In another study, Ljolje and Fallside (1987) used hidden Markov models to represent prosodic features such as fundamental frequency, fundamental frequency time derivative, energy, and smoothed energy of isolated words. An interesting observation of their experiment was that the poorest error rate accounted for non-native speakers of English. This result suggests that prosodic structure may be useful in accent discrimination.

In this paper, we investigate the ability of a proposed HMM algorithm to classify accent accurately. During algorithm formulation, particular attention is placed on accent sensitive acoustic speech characteristics, with an effort to show quantitatively the differences among accents.

The outline of this paper is as follows. In Section 2, the accent database is described, and sound problems relating to foreign accent are discussed. Particular examples are given which illustrate phoneme substitution for American English by non-native speakers. The accent classification system is formulated in Section 3. Evaluations are conducted in Section 4 to establish accent classification performance for three test scenarios based on isolated versus continuous speech, and partial versus full model search. A set of subjective listening tests are also conducted in order to compare computer algorithm performance. Finally, accent information is used to improve the robustness of a speech recognition system. A discussion of the results and conclusions are presented in Section 5.

2. Speech and foreign accent

A number of studies have been directed at formulating a better understanding of the causes of foreign accent in English (Flege, 1988; Piper and Cansin, 1988; Tahta and Wood, 1981). Flege (1988) investigated factors affecting the degree of perceived foreign accent in English sentences. From these studies it was determined that each person develops a speaking style up to the age of 12, which consists of phoneme production, articulation, tongue movement and other physiological phenomena related to the vocal tract. Non-native speakers preserve this speaking style when learning a second language, and therefore substitute phonemes from their native language when they encounter a new phoneme in the second language. It has been shown that there exists a critical time period for a speaker learning a second language in another country. During this critical period, there is rapid pronunciation improvement. However, after this period is over, the learning curve levels off. Therefore, Flege (1988) showed that pronunciation scores for non-native speakers living in the United States from one to five years were not appreciably different. In the remainder of this section, a sequence of observations regarding accent and phonemic content will be considered. The intent here is to illustrate that accent affects prosodic structure as well as aspects of phoneme content (i.e., phoneme substitution, addition, etc.).

Chreist (1964) investigated the sounds that emphasize non-native speaker accent during English speech production. For example the /AE/ sound [2] as in *cat* is not available in most other languages.

[2] In this study, uppercase ARPABET notation is used to describe phonemes for American English. See (Deller et al., 1993) for a summary.

Norwegian and Chinese are two of the few languages in addition to English that use the /AE/ sound. Speakers whose native language is Arabic, substitute the /AA/ for the /AE/ phoneme consistently in such words as: *add*, *and*, *bat*, *dad*. The voiced flap /DH/ as in *there*, and the unvoiced flap /TH/ as in *three* present another difficulty for most non-native English speakers. Turkish, Polish, Hungarian, Spanish and Indian speakers substitute the /D/ and /T/ for /DH/and /TH/ consistently, whereas the Arabic and Chinese speakers substitute the /S/ and /Z/ for the same phonemes. For speakers whose native language contains only pure vowels, glides from one vowel target to another are not permitted, thus the diphthong presents an enigma. For these speakers, the rules for their language prevent their vocal systems from producing the necessary shifts in posture. For example in Japanese, there are no diphthongs. When two vowels appear consecutively in the same word, Japanese speakers cannot produce the correct articulatory movement from one target to the other. For example, in the word *eat*, the tendency is to pronounce *eat* as *it*. The words *boy*, *how*, *line* are shortened from their native diphthong character to a single pure vowel sound. Portuguese, Spanish, Italian, German and Norwegian have sounds similar to the diphthongs of English. However German, Italian and Norwegian use different diphthongs than those in English. This imposes additional difficulties on the speaker, since those diphthongs present in his native language are normally substituted for the proper diphthong.

In American English, there are twelve principal vowels. Phoneticians often recognize a thirteenth vowel called the "schwa" vowel. It is sometimes called a "degenerate vowel" since other vowels gravitate towards this neutral vowel when articulated hastily in the course of continuous speech. Since it is substituted so freely, the non-native speaker finds the schwa to be one of the most difficult sounds of American English. The schwa appears in the initial, medial and final position of many word units in American English. Many secondary stressed syllables have neutralized vowels which approach the "schwa" position. The four characteristics of defective articulation are evident for this sound among non-native English speakers; the schwa sound is heard as an addition, distortion, omission or a substitution when foreign accent is present. Among the second language learners, sound substitution is the most common problem for the schwa. In American English, for words that begin with "a" or "un" such *as above* and *unknown*, the tendency of the non-native speaker is to substitute the /AA/ sound for the leading vowels. The medial substitution in words such as *uniform*, *disappear*, *disappoint*, *disability* is another source of problem. Other words which include the schwa sound are *laboratory*, *president* and *communication*. The word initial and word final additions of the schwa are also prevalent among non-native speakers. For example, Spanish speakers add a schwa sound to the beginning of words that begin with the /S/ sound (e.g., *school*, *spend*, *space*). On the other hand, when attempting to produce what the second language hears as the final sound in such words as *bag* or *pod*, the speaker adds a voiced release to the end. The prominence of these initial and final additions becomes a vivid indicator of foreign accent.

Among the back vowels, the rounded, lax sound /UH/ contrasts with the front, unrounded, lax vowel /AE/ just discussed. In such words as *took*, *look*, *full*, *pull*, *should*, *would*, the sound represents three combinations of written or printed letters.

The majority of European languages have no /UH/ phoneme, and among those who have difficulty in acquiring the sound in American English are the speakers of French, Italian, Spanish, Portuguese, Greek, Swedish, Danish, Russian, Polish and Czech. Such expressions as "He took my book", and "The moon could put him in a good mood", demonstrate alternate substitutions.

Most non-native speakers have problems in producing the sound /NX/ as in: *sing*, *bringing*. The absence of the sound in French, Russian, Navajo, Italian and Polish often creates real difficulty in its production by such speakers. The fact that several of these languages have a similar sound in their speech causes additional confusion in production when native speakers of those languages attempt to produce the /NX/ in *sing*, *bring*, *sprinkle*.

The variety of /R/ sounds produces problems not only for native American children and adults, but also for students of English as a second language. Establishing the American English /R/ in the speech of the second language learner involves learning

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

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?				

both a consonant form and a vowel form. For example, Chinese speakers substitute /L/ for /R/ in words such as *fear*, *car*, *dart*. The vowelized /ER/ appears in words such as *bird*, *heard*, *turn*, though it creates less overall trouble. For the non-native speaker accustomed to pure vowels, the glide /R/ offers double difficulty. The speaker must first produce a new sound, and then must learn to glide into another position in order to use the sound syllabically.

In this section, we have considered a number of examples of how the placement of phoneme content signifies the presence of accent in speech production. It should be noted that a listener's judgement of speaker accent is based on perception, and therefore each listener may selectively choose those accent relayers modified by the speaker to determine accent. With this knowledge, it is suggested that an effective accent classification algorithm could be proposed which capitalizes on accent sensitive word or phoneme sequences.

2.1. Accent database

Based on the extensive literature review of language education of American English as a second

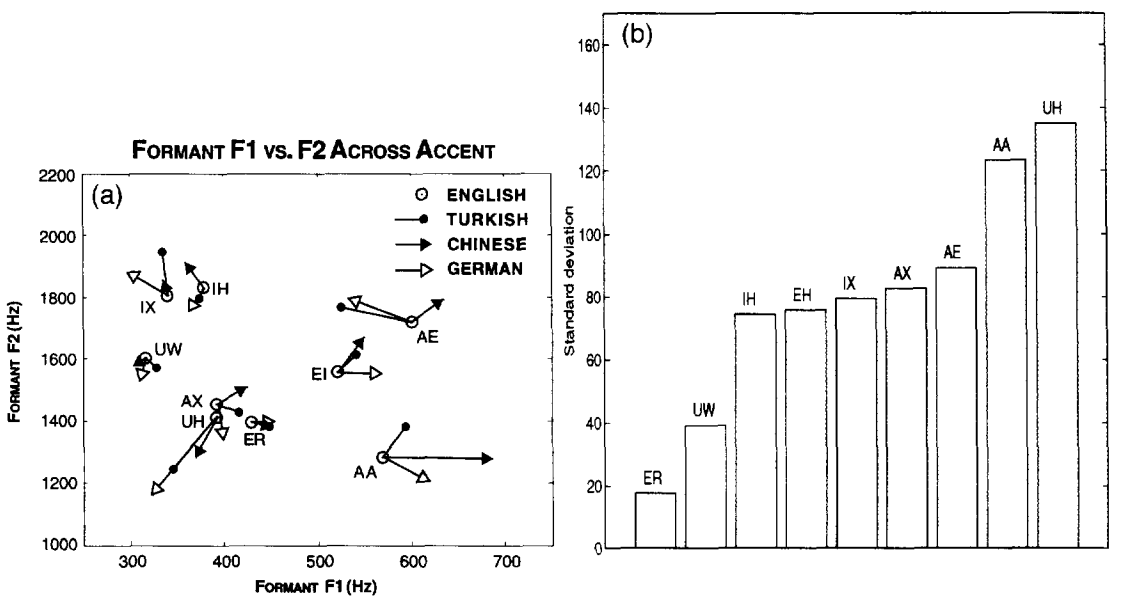

Fig. 1. The first formant versus second formant (a) scatter plot and (b) variance for available phonemes in the vocabulary.

language, a test vocabulary was selected using a set of twenty isolated words, and four test sentences. These words and phrases are listed in Table 1. A portion of the data corpus was collected using a head-mounted microphone in a quiet office environment, and the remaining portion was collected through an online 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. Available speech includes neutral American English, and English under the following accents: German, Chinese, Turkish, French, Persian, Spanish, Italian, Hindi, Romanian, Japanese, Greek and others. For the studies conducted here, the focus was on American English speech from 48 male speakers across the following four accents: neutral, Turkish, Chinese and German.

2.2. Analysis of foreign accented speech

In order to illustrate the sound variation among different accents, vowel codebooks were obtained for first (F_1) and second (F_2) formants for four accents (neutral, Chinese, Turkish, German) by averaging the formant frequencies across all speakers in each accent group. The codebooks contain only those vowels available in the accent database. Fig. 1(a) illustrates an F_1 versus F_2 scatter plot of the four accent codebooks. A quantitative measure of accent difference is shown in Fig. 1(b) based on the standard deviation of the distance from the centroids of each vowel among four accents. From this figure, it can be seen that the English and Chinese /AE/ sounds are well separated from German and Turkish /AE/ sounds, which confirms the fact that no /AE/ phoneme exists in Turkish and German, and that these speakers tend to substitute the /EH/ sound in its place. The /UH/ sound is found to be the most sensitive phoneme for the four accents considered. Measurable differences in both F_1 and F_2 for all four accents were noted. Finally, only the first formant F_1 changed for English versus Chinese, and primarily F_2 for Turkish versus German in the /AA/ vowel. These results clearly demonstrate that vowel centroids can be useful in accent assessment.

In order to investigate the influence of accent on

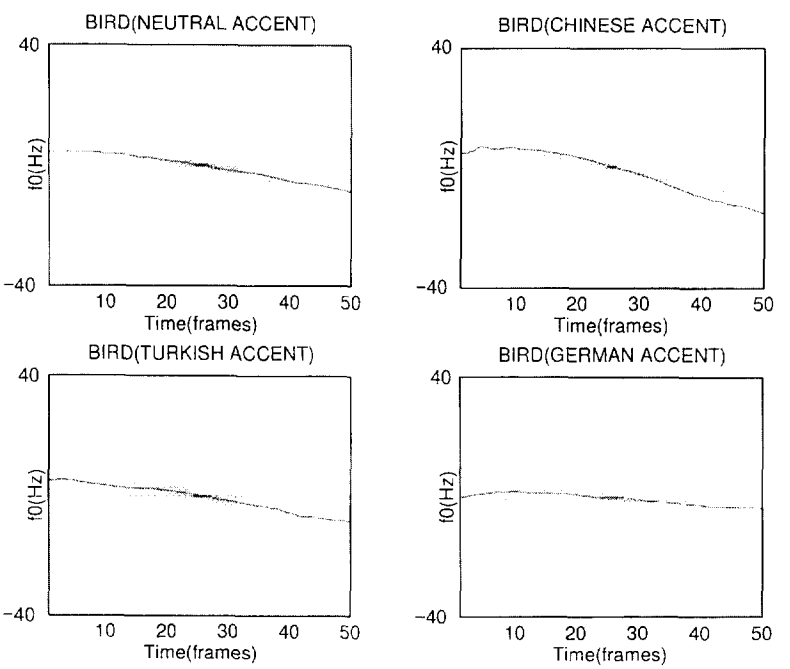

Fig. 2. The gray-scale histograms of the normalized pitch (fundamental frequency) contours for four different accents for the word *bird*.

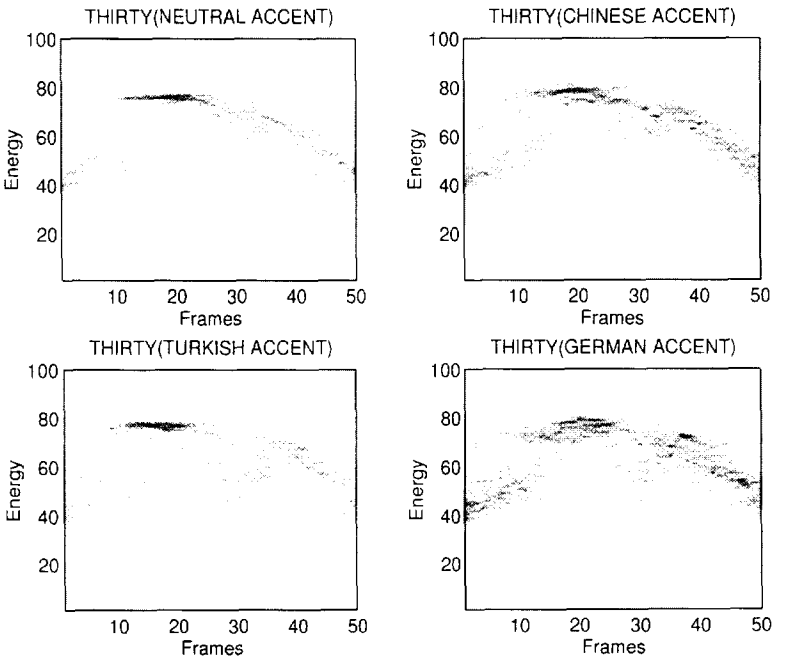

Fig. 3. The gray-scale histograms of the energy (in dB) contours for four different accents for the word *thirty*.

prosodic features, normalized pitch and energy (in dB) contours for various words were generated using accented speech data. In Fig. 2, four gray-scale plots illustrate pitch histogram contours for the word *bird* spoken under four different accents using 5 tokens of 12 speakers (60 tokens) from each accent group. The histograms were generated after all waveforms were time-aligned. Since all words were uttered in isolation, it is not possible to capture how pitch contours vary in spontaneous speech. However, given the number of speakers and tokens, statistically significant trends can be drawn. The solid lines in the figure represent the median values at each time frame. Although the histograms can be distinguished from each other visually, there is significant inter-speaker variability caused by factors other than accent, such as speaker-dependent traits, emotion, environment, etc. In spite of this, it is clear that overall differences exist across the four accents. In Fig. 3, gray-scale histograms of the energy contours for the word *thirty* under the same accents are shown. The contours are better defined in these plots as compared to pitch contours. The energy level drop in the middle of each contour indicates the pause duration between the /ER/ and /DX/-/IY/ sounds while pronouncing the word *thirty*. It can be concluded

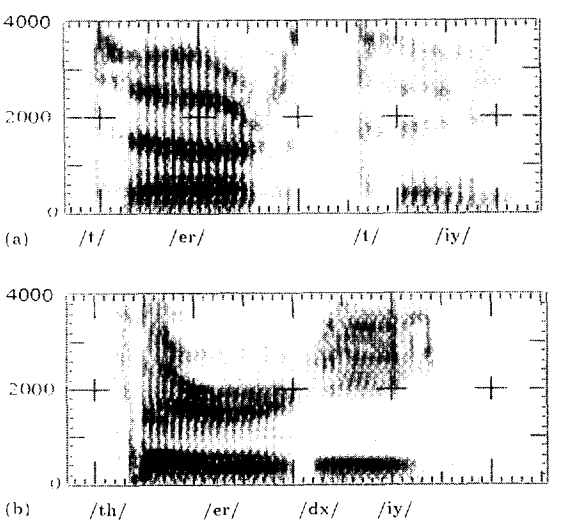

Fig. 4. The spectrograms of the word "thirty" spoken by (a) a Turkish speaker, (b) an American speaker. These audiofiles can be retrieved at http://www.elsevier.nl/locate/specom .

that the American speakers insert a short pause between these syllables, while Turkish speakers tend to insert a pause of much longer duration.

Next, spectrograms of the word "thirty" spoken by a Turkish speaker and an American speaker are compared in Fig. 4. The neutral accent phoneme sequence is /TH/-/ER/-/DX/-/IY/ for this word. For the Turkish speaker in Fig. 4(a), there is a well defined substitution of /T/ for /TH/, where the /T/ sound has greater energy in the high frequencies. After the /ER/ phoneme, the Turkish speaker chooses to pause for about 70 ms, whereas the American speaker pauses for only 10 ms. Next, the Turkish speaker substitutes the /T/ for /DX/ sound. These spectrograms help illustrate the drastic changes in pronunciation patterns of words when spoken by non-native speakers.

3. Classification system

A flow-diagram for the proposed accent classification system is shown in Fig. 5. Input continuous speech is first sampled at 8 kHz. Next, acoustic feature extraction is performed on sampled data on a frame-by-frame basis, and pre-emphasized with the filter $1 - 0.95z^{-1}$. Hamming window analysis is applied to frames that are 25 ms long with a 15 ms overlap. Next, energy and 8th order Mel-cepstrum coefficients (Deller et al., 1993) are computed. The

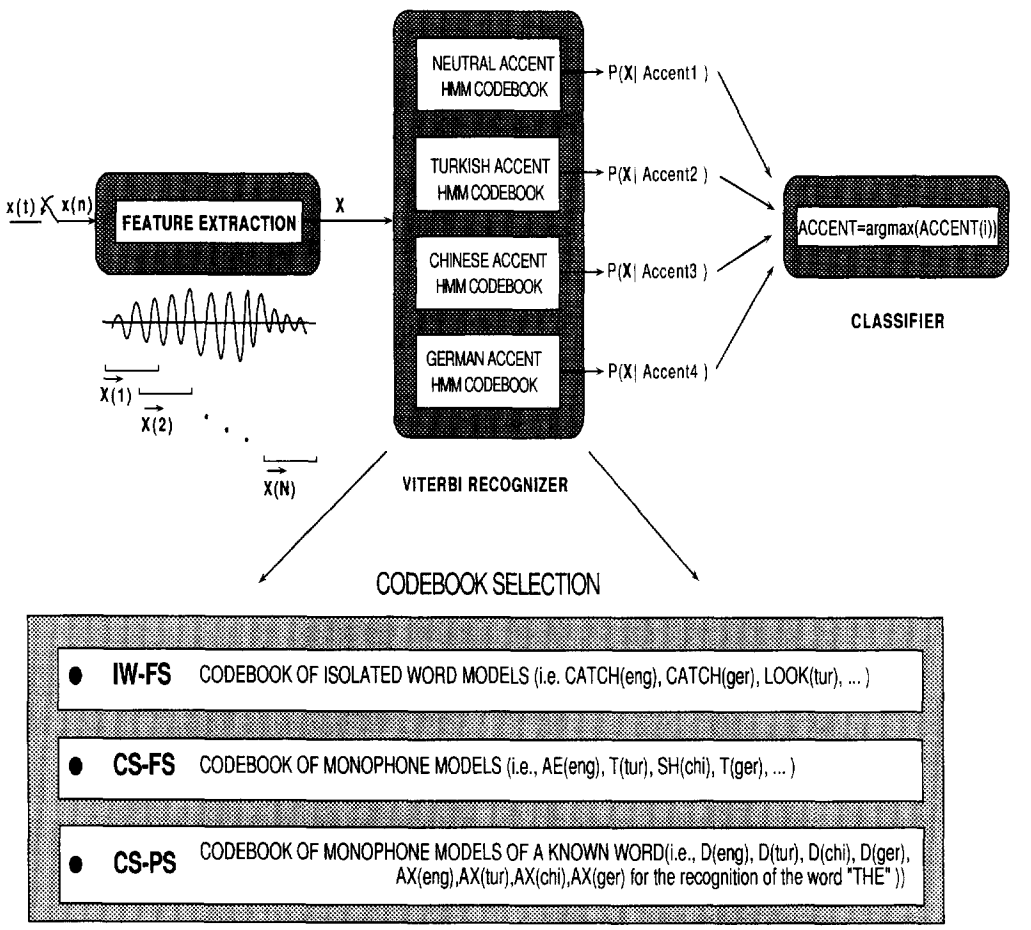

Fig. 5. Framework for the accent classification algorithm.

final acoustic feature set (i.e., $\vec{X}_i$, $i = 1,\ldots,N$) includes the Mel-cepstrum coefficients and energy along with their first order differences (i.e., delta Mel-cepstral, and delta energy).

In this study, three scenarios are considered in the formulation of the accent classification algorithm. All three scenarios employ a left-to-right hidden Markov model (HMM) topology with no state skips allowed. The differences among the scenarios are based on the speech units employed in their grammars and the amount of a priori knowledge utilized. For the first scenario (IW-FS, i.e., Isolated Word – Full Search), isolated word HMM recognizers are trained for each accent consisting of the 20 words in the vocabulary set. The number of states for each word is set proportional to the average duration of the word. The number of states in the IW-FS based HMMs ranged from 7 to 21. In the classification testing phase, the input utterance is submitted to each word model, and the accent associated with the most likely model is selected as the accent of the speaker.

The second scenario (CS-FS, i.e., Continuous Speech – Full Search) uses monophone models, and is therefore vocabulary independent. Codebooks of monophone HMMs are created for each accent type in the training phase. Each monophone HMM uses 5 states, including the non-emitting initial and final states. The HMM codebooks generated after training are used in the classification phase to obtain a score of the likelihood of each accent codebook for a given test utterance. The scoring procedure uses a Viterbi decoder to estimate an average probability that the given observation vector is produced by each accent (i.e. $P(\boldsymbol{X} \mid \text{Accent}_i)$, $i = 1,\ldots,M$). Finally, the accent resulting in the maximum probability is selected as the test speaker's accent (i.e., $\text{argmax}(P(\boldsymbol{X} \mid \text{Accent}_i), i = 1,\ldots,M)$).

In the third scenario (CS-PS, i.e., Continuous Speech – Partial Search), the same monophone models generated for the CS-FS are used, but the monophone text sequence of the test utterance is assumed to be known a priori (i.e., we assume that the speaker is required to produce a particular input text sequence). Therefore only those phonemes which are present in the utterance are searched in the Viterbi recognizer.

For algorithm evaluation, speech from the head-mounted microphone portion of the accent database is used. The training set consisted of 5 tokens of 20 isolated words from 4 American, 4 Turkish, 4 Chinese and 4 German speakers. Twelve speakers (5 American, 4 Turkish, 3 Chinese) were set aside in order to evaluate the performance of the classifier under open speaker test conditions.

Table 2
A summary of the evaluation of the proposed accent classification algorithms. The test speaker set includes 5 American, 4 Turkish and 3 Chinese speakers

Comparison of 3 different approaches for accent identification	
System	Open test accent ID rate
IW-FS	74.5
CS-FS	61.3
CS-PS	68.3

4. Evaluations

4.1. Comparison of IW-FS, CS-PS and CS-FS

The algorithms described above were tested using the following 5 words (catch, communication, target, thirty, bringing) using open test speakers (i.e., (5 words) × (5 tokens) × (12 open test speakers) = 300 trials). In Table 2, the average accent classification rates for the 5 word vocabulary are summarized for each scenario described above. Not surprisingly, the CS-PS algorithm configuration performed better than CS-FS configuration (68.3% versus 61.3% classification rate), since phoneme substitution is one of the most important accent relayers, and it becomes transparent under the CS-FS scheme. The reason for this is that each time the system scores a phoneme, every monophone model in the codebook is considered (i.e., the word "the" is allowed to be scored with a monophone sequence /D/-/AX/whereas the true search sequence should have been/DH/-/AX/). Therefore, the phoneme substitution of /T/ for /TH/ in the word "three" for most language accents is not considered as an outlier in the CS-FS system. The IW-FS algorithm configuration performed better (74.5%) than the other two methods. This was also expected, since by using whole word models, the system is better able to track the resulting articulatory movements present in accented speech.

4.2. Dependence of IW-FS performance on utterance context and length

In Fig. 6, the accent classification rates for each of the 20 words are summarized. After comparing performance across the vocabulary set, it can be concluded that some words are better identifiers of accent than others for the four accents considered. For example, the word *target* resulted in the highest accent classification rate (90%). In general, higher classification rates were achieved for words of longer duration (e.g., *aluminum*, *communication*, *bringing*). In order to evaluate the importance of utterance length for accent classification, words were randomly selected (1 at a time, 2 at a time, etc.), from the open test speaker recordings. These random word lists with varying word count length were evaluated by the IW-FS accent classifier. The log-probability was estimated for each word belonging to each accent. Next, the log-probabilities of the words for that list and for each accent were summed to make an overall decision concerning accent. In these tests, again only open test speakers are considered with decisions made among the models for neutral, Turkish, German and Chinese accents. The results are shown in Fig. 7. The graph in Fig. 7(i) shows the

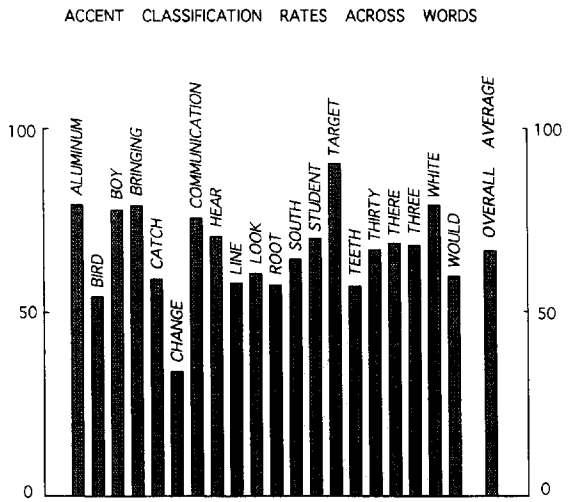

Fig. 6. The accent classification (IW-FS) rates among the 4 language accents for the 20 words in the vocabulary. All 12 speakers are open-test speakers. Audiofiles corresponding to some of these words (3 words×4 speakers) can be retrieved at http://www.elsevier.nl/locate.specom .

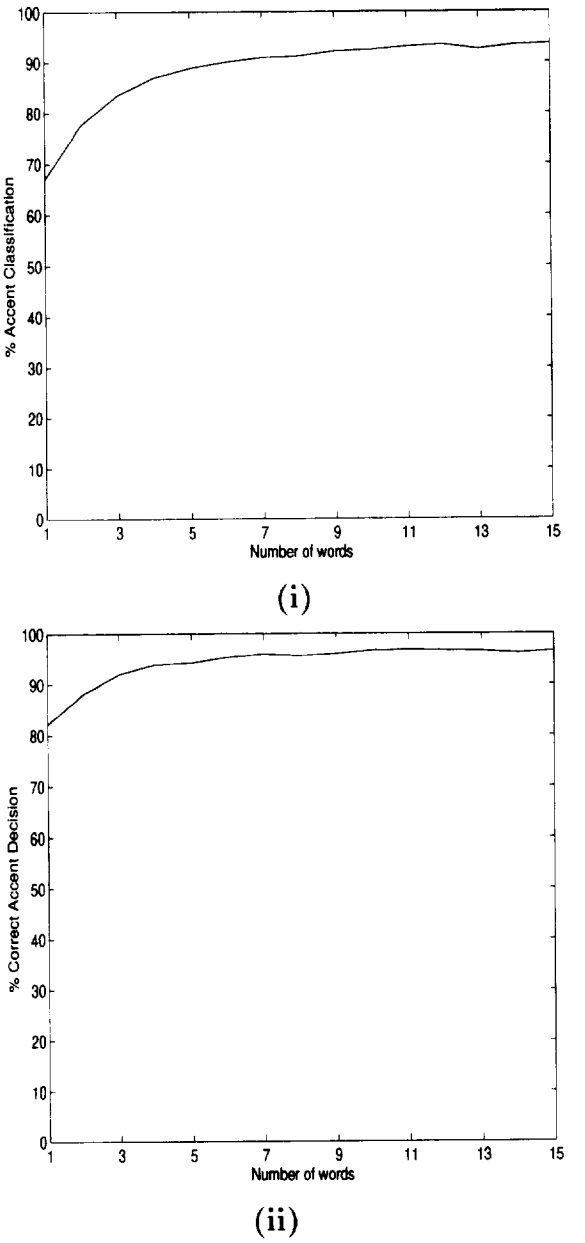

Fig. 7. The effect of speech duration on (i) accent classification and (ii) accent detection rates for 12 open-test speakers among 4 different language accents.

improvement of accent classification rate as a function of the number of words uttered by the speakers. An accent classification rate of 93% is achieved after 7–8 words, with little improvement resulting as ob-

servation test word count increases. The problem of accent detection is considered in these evaluations as well. This involves a binary decision as to whether the person is speaking with a foreign accent, or has neutral accent. Therefore, classifying a Turkish speaker as having a Chinese accent is not considered an error when the goal is accent detection. In Fig. 7(ii), the graph shows the accent detection rate as a function of the number of words uttered. Again after 7–8 words, the curve levels off, this time achieving a 96% accuracy. The level of misclassification after 7–8 words should not be judged as pure system error, since the degree of accent depends strongly on the speaker. Issues such as (i) whether the speaker was taught English by a native speaker, (ii) the age when English learning began, (iii) at what age the speaker moved to the United States, and (iv) how long he has resided in the United States are all important factors that affect the level of foreign accent that a person exhibits. After a more careful analysis of the classification errors, it was not surprising that a number of the accent detection errors were committed on speech data from a speaker who started learning English at the age of 12 from an American teacher. He also arrived in the United States earlier than all other non-native speakers.

4.3. Comparison of computer and human performance in accent discrimination

In order to compare human performance in foreign accent discrimination to that of the computer algorithm (IW-FS), we performed a listening test experiment on 12 native and 9 non-native speakers of English. The listeners were volunteers among students and faculty at Duke University. First, sample words of different speakers from each accent (neutral, Turkish, Chinese, German) were played so that each listener acquired a fixed amount of training for the types of accented speech prior to the experiment. When questioned, most listeners indicated they already had some familiarity with the accent types to be considered. Each listener was presented with 48 words, which were selected randomly from open test speaker data. After listening to each word, the listener was asked to specify whether the utterance possessed any foreign accent or not. When the listener decided that an accent was present, he was

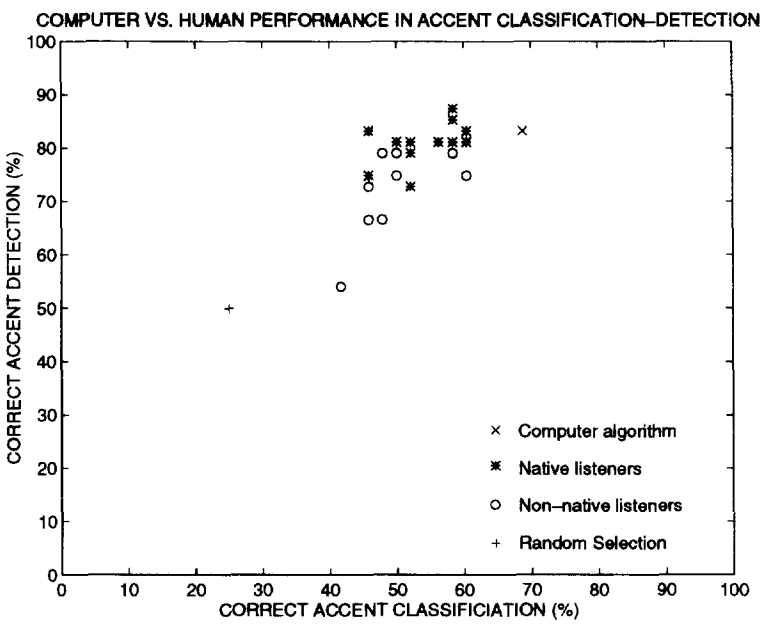

Fig. 8. A scatter plot of accent classification versus accent detection rates for all test listeners (* native listeners, O non-native listeners), the proposed accent computer algorithm (×), and a statistically random selection (+: a random selection of 1 in 4 accents for classification, a 0.5 detection probability of accented versus non-accented speech).

further asked to classify the accent as one of Turkish, Chinese or German accents. During the test session, the listener was allowed to listen to the same word more than once before making his decision. The individual results are shown in Table 3. The average classification rate of the listeners was 52.3%, and the average detection rate was 77.2%.

Next, the computer algorithm (IW-FS) was tested using the same word set presented to the listener group. A scatter plot of each listener test result combined with computer algorithm and the random chance rate is shown in Fig. 8. The computer algorithm was able to classify accent with an 68.8% accuracy, while the best classification rate achieved by listeners was 60.4%. However, for accent detection, native listener performance was in general comparable to that of the computer algorithm (83.3%), with two listeners outperforming the algorithm by a slight margin.

In Fig. 9, a comparison of the average classification and detection rates are given. The results suggest that listeners who are native speakers of English

Table 3
Results obtained from the subjective listening test on foreign accent classification and detection

Listener test results			
Listener no.	Native tongue	Classification rate	Detection rate
1	English	60.4	81.2
2	English	58.3	81.2
3	English	52.1	81.2
4	English	52.1	79.2
5	English	45.8	83.3
6	English	56.2	81.2
7	English	58.3	85.4
8	English	58.3	87.5
9	English	45.8	75.0
10	English	60.4	83.3
11	English	52.1	72.9
12	English	50.0	81.2
13	Spanish	50.0	75.0
14	Hindi	45.8	66.7
15	Arabic	50.0	79.2
16	Arabic	47.9	79.2
17	Turkish	47.9	66.7
18	Malayali	60.4	75.0
19	Kanala	58.3	79.2
20	Italian	41.7	54.2
21	Turkish	45.8	72.9

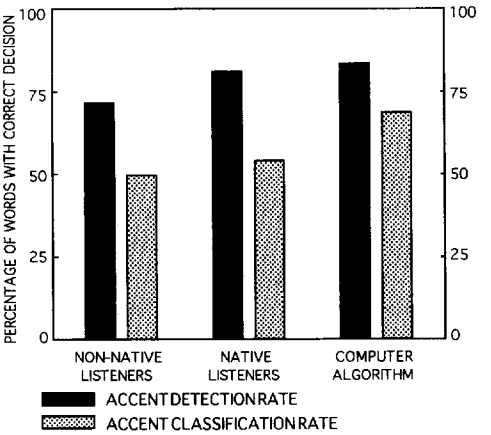

Fig. 9. The average accent classification and detection rates for native speakers, non-native speakers, and computer algorithm.

can detect and classify accent better than non-native speakers. Most decision errors were committed for words spoken by Turkish speakers. For example the words "student" and "teeth" were judged as neutral, and tokens of the word "bird" were judged as German accented by most listeners even though they were produced by Turkish speakers. The majority of the words spoken by American speakers were identified correctly as neutral by the listeners. However, some listeners judged single syllable words such as "root" and "hear" as accented. Only the words "communication" and "target" spoken by American speakers were unanimously classified correctly by all the listeners. In general there was a higher level of decision agreement among native listeners than for non-native listeners. Listeners identified the accent of the word "change" correctly in a majority of the cases, however the computer algorithm could not identify the accent for this word as reliably. The judgement of a majority of the listeners agreed with the computer algorithm's decision 69% of the time.

4.4. Application to speech recognition

One application of accent classification is to assist or direct speaker independent speech recognition algorithms. It is well known that speech recognition algorithms tend to degrade in performance when accented speech is encountered. A study was conducted in order to investigate the response of a speech recognition algorithm to accented speech, and

a method was proposed to improve the recognition rate. Generally, speech recognition algorithms work well for American speakers, especially for small vocabulary isolated word systems. For this study, the 20-word database was used for training. Isolated word HMMs for neutral, Turkish, Chinese and German accents were generated. The topology was a left-to-right HMM with no state skips allowed. The number of states for each word 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 the evaluations, a balanced amount of microphone and telephone data was used. In addition, the microphone speech was bandpass filtered between 100 Hz and 3800 Hz in order to simulate the telephone channel distortion.

First, the speech recognition system was tested with American, Chinese, German and Turkish speakers using neutral accented HMMs in the Viterbi scoring algorithm. The resulting error rates are shown in Fig. 10 by the black colored boxes. The average error rate for the American speakers was 0.3%, whereas for Chinese speakers it was 11.3%, for Turkish speakers it was 7.5%, and for German speakers it was 4.7%. The error rates for non-native speakers were substantially higher than for native speakers. Next, open set non-native speakers were tested with the HMMs generated from other speakers of their accent group (i.e., each Turkish speaker was tested with HMMs generated with speech data from the remaining Turkish speakers). The resulting error rates for true accented HMMs are shown in Fig. 10 (as grey boxes). Using accent knowledge, the error rate was reduced to 3.7% for Chinese speakers (a 67.3% decrease from the original), 2.0% for Turkish speakers (a 73.3% decrease from the original), and 1.3% for German speakers (a 72.3% decrease from the original). The dramatic reduction in error rate suggests that training separate models for each accent can be very useful in speaker independent speech recognition systems.

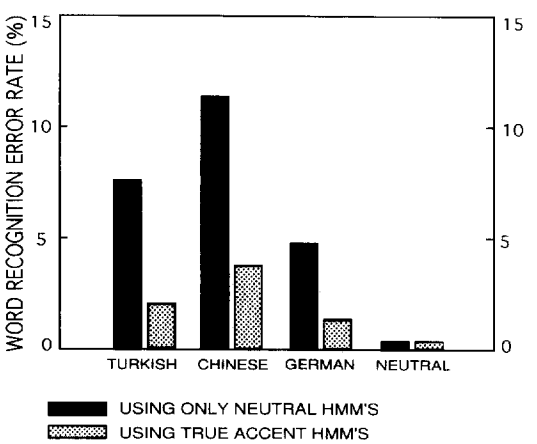

Fig. 10. The error rates for speakers of Turkish, Chinese, German and neutral accents, for two conditions: (i) only neutral accent word models are used, (ii) true-accent word models of the speaker are used.

5. Conclusion

In this paper, the problem of accent classification for American English is considered. An investigation was first conducted using a number of prosodic and spectral features for foreign accent discrimination. Energy and spectral information were found to be useful parameters in accent classification. Other prosodic features such as pitch/intonation contour were also shown to vary significantly. However, due to inter-speaker variability, it may be more difficult to isolate that portion of the parameter variation due to accent alone. Next, spectral and energy based features were used to formulate an HMM based accent classification algorithm. Three different scenarios employing monophone and whole word models were considered. It was shown that whole word models capture accent information better than monophone models. Another observation was that the classification performance increases as test word count increases. Using input test strings of 7–8 words the accent classification system is able to identify the accent among four different accent classes with a 93% accuracy. The system also achieves correct accent detection (i.e., whether the speaker has any foreign accent or not) with a 96% accuracy. It was also clearly illustrated that some words are better relayers of accent than others. Such

vocabulary sensitive information can be utilized to improve accuracy for future classification systems.

In order to compare computer performance with human performance, a subjective listening test was established. Listeners were able to classify accent with 52.3% accuracy, and detect accent with 77.2% accuracy. The computer classification algorithm (IW-FS) was tested on the same word set presented to listeners. Its performance was superior to listener performance, with 68.7% classification and 83.3% detection rates. Finally, in order to investigate the application of accent classification for robust speech recognition, an experiment using accent dependent hidden Markov models was performed. By using accented word models, error rate reductions were achieved as follows: from 11.3% to 3.7% for Chinese accented words, from 7.5% to 2.0% for Turkish accented words, and from 4.7% to 1.3% for German accented words. This dramatic reduction confirms that knowledge estimated from accent classification can be a useful source of information to improve the robustness in speaker independent speech recognition systems.

References

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

W.J. Barry, C.E. Hoequist and F.J. Nolan (1989), "An approach to the problem of regional accent in automatic speech recognition", *Computer Speech and Language*, Vol. 3, pp. 355–366.

K.M. Berkling and E. Barnard (1994a), "Language identification of six languages based on a common set of broad phonemes", *Proc. Internat. Conf. on Spoken Language Processing*, Vol. 4, pp. 1891–1894.

K.M. Berkling and E. Barnard (1994b), Analysis of phoneme-based features for language identification, *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Vol. 1, pp. 289–292.

F.M. Chreist (1964), *Foreign Accent* (Prentice-Hall, Englewood Cliffs, NJ, USA).

J. Deller, J. Proakis and J.H.L. Hansen (1993), *Discrete Time Processing of Speech Signals*, Macmillan Series for Prentice-Hall (Prentice-Hall, Englewood Cliffs, NJ, USA).

J.E. Flege (1984), "The detection of French accent by American listeners", *J. Acoust. Soc. Amer.*, Vol. 76, No. 9, pp. 692–707.

J.E. Flege (1988), "Factors affecting degree of perceived foreign accent in English sentences", *J. Acoust. Soc. Amer.*, Vol. 84, No. 6, pp. 70–79.

C. Grover, D.G. Jamieson and M.B. Dobrovolsky (1987), "Intonation in English, French and German: Perception and production", *Language and Speech*, Vol. 30, No. 3, pp. 277–295.

V. Gupta and P. Mermelstein (1982), "Effects of speaker accent on the performance of a speaker-independent, isolated-word recognizer", *J. Acoust. Soc. Amer.*, Vol. 71, pp. 1581–1587.

J.H.L. Hansen and L.M. Arslan (1995), "Foreign accent classification using source generator based prosodic features", *ICASSP-95: IEEE Proc. Internat. Conf. Acoust. Speech Signal Process.*, Detroit, MI, May 1995, pp. 836–839.

T.J. Hazen and V.W. Zue (1994), "Recent improvements in an approach to segment-based automatic language identification", *Proc. Internat. Conf. on Spoken Language Processing*, Vol. 4, pp. 1883–1886.

A.S. House and N. Neuberg (1977), "Toward automatic identification of the language of an utterance. 1. Preliminary methodological considerations", *J. Acoust. Soc. Amer.*, Vol. 62, No. 3, pp. 708–713.

A. Ljolje and F. Fallside (1987), "Recognition of isolated prosodic patterns using hidden Markov models", *Computer Speech and Language*, Vol. 2, pp. 27–33.

Y.K. Muthusamy, E. Barnard and R.A. Cole (1994), "Reviewing automatic language identification", *IEEE Signal Process. Mag.*, October, pp. 33–41.

T. Piper and D. Cansin (1988), "Factors influencing foreign accent", *Canadian Modern Language Review*, Vol. 44, No. 2, pp. 334–342.

L.R. Rabiner and J.G. Wilpon (1977), "Speaker-independent isolated word recognition for a moderate size (54 word) vocabulary", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 27, pp. 583–587.

S. Tahta and M. Wood (1981), "Foreign accents: Factors relating to transfer of accent from the first language to a second language", *Language and Speech*, Vol. 24, No. 3, pp. 265–272.

M.A. Zissman (1993), "Automatic language identification using Gaussian mixture and hidden Markov models", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, pp. 399–402.

M.A. Zissman (1995), "Language identification using phoneme recognition and phonotactic language modeling", *Proc. Internat. Conf. on Spoken Language Processing*, Vol. 5, pp. 3503–3506.