Foreign Accent Classification Using Source Generator Based Prosodic Features

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

ABSTRACT

Speaker accent is an important issue in the formulation of robust speaker independent recognition systems. Knowledge gained from a reliable accent classification approach could improve overall recognition performance. In this paper, a new algorithm is proposed for foreign accent classification of American English. A series of experimental studies are considered which focus on establishing how speech production is varied to convey accent. The proposed method uses a source generator framework, recently proposed for analysis and recognition of speech under stress[5]. An accent sensitive database is established using speakers of American English with foreign language accents. An initial version of the classification algorithm classified speaker accent from among four different accents with an accuracy of 81.5% in the case of unknown text, and 88.9% assuming known text. Finally, it is shown that as accent sensitive word count increases, the ability to correctly classify accent also increases, achieving an overall classification rate of 92% among four accent classes.

1. INTRODUCTION

A number of recent studies have considered the problem of language identification based on various speech production features [6, 8]. A related problem, which has not been explored in detail is the issue of foreign accent identification. Second language learning requires a modification in the patterns of intonation, lexical stress, rhythm, grammar, as well as the use of additional distinctive phonemes which are perhaps unfamiliar to the speaker. Thus, the person with a foreign accent can be said to possess deviations in articulation, rhythm, voice and symbolization. The role of intonation in foreign accent has been studied. An experiment verified that French, English and German speakers differ in the slopes (fundamental frequency(f_0) divided by time) of their continuative intonation [3].

Accent is also a challenging problem in speech recognition. It is one of the most important factors, aside from gender, that creates undesirable variability in speaker independent speech recognition schemes. Studies have attempted to normalize regional accent prior to speech recognition. In [1], a two stage procedure was proposed for British English accent normalization. In stage one, an accent identification 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.

The use of prosody in speech recognition systems has received much interest recently. Ljolje and Fallside used hidden Markov models to represent the prosodic features (i.e., f_0, $\frac{df_0}{dt}$, energy, and smoothed energy) of isolated words [7]. An interesting outcome of their experiment was that the largest error rate resulted for non-native speakers of English. This suggests a correlation between speaker accent and prosodic features.

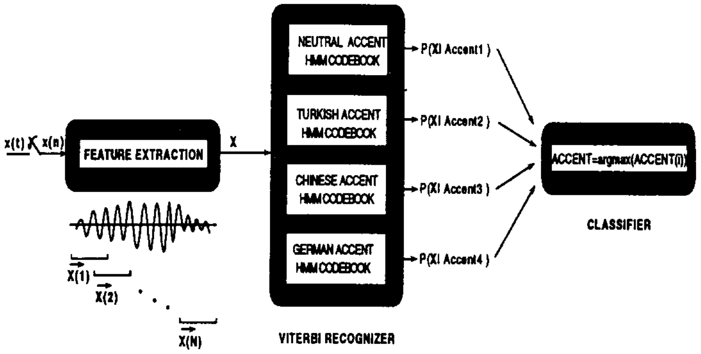

Figure 1: Framework for the accent classification algorithm.

In this paper, we consider a series of experiments whose results motivate the formulation of an algorithm for foreign accent classification. The evaluation considers a range of acoustic parameters as potential accent relayers. The motivation is based on discovering what traits in speech production are responsible for conveying speaker accent. An accent classification algorithm is formulated based on those features which are determined to be reliable accent relayers (Fig. 1). A useful application for such an algorithm would be for selection of alternate pronunciation models in a multiple entry dictionary for a speech understanding system.

2. SOUND PROBLEMS IN FOREIGN ACCENT OF ENGLISH

Previous research studies in language education have been directed at improving the intelligibility of American English spoken by non-native speakers [2]. These studies have shown that one develops a speaking style while acquiring language skills up to the age of 16, which consists of phoneme production, articulation, tongue movement and other physiological phenomena related to the vocal tract. In general, a foreign speaker preserves this speaking style while learning a second-language. Therefore, the speaker will substitute phonemes from his/her native language when they encounter a new phoneme in their second language. As an example, the /ae/ sound as in *cat* is not a valid phoneme in most languages. Norwegian and Chinese are two of the few languages excluding English that use the sound /ae/. Speakers whose native language is Arabic,

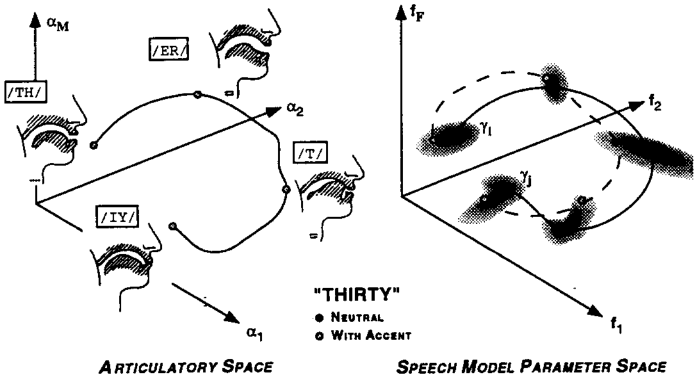

Figure 2: Sample Source Generator paths for American English under neutral and foreign accent conditions as projected from an articulatory feature space.

substitute /a/ for the /ae/ phoneme consistently in such words as: *add, and, bat, dad.* Another sound that creates problems is the phoneme /uh/ as in *look, book.* The majority of European languages possess no /uh/ phoneme. Certain languages such as Japanese have no diphthongs. When two vowels appear together in the same word as separate syllables, there is no change of position or glide in producing the sounds. For the example word "eat," the tendency is to pronounce "eat" as "it". Also, for the words *boy, how, line,* the duration of their native diphthong character is shortened to a single pure vowel sound. Given these and other accent/language dependent cases, a continuous speech recognizer is likely to fail in finding a match with a pronunciation dictionary that is designed for native American English speakers.

3. SOURCE GENERATOR FRAMEWORK

It is suggested that the changes which occur in American English speech production under foreign accent can be represented using a recently proposed framework for analysis and modeling of speech under stress. This method, proposed by Hansen [4, 5], suggests that normal speech production consists of a sequence of movements in some articulatory space from one source generator to another. Actual speech production consists of a "neutral" speech feature production path which must be traversed to produce a given word or utterance. In [4, 5], it was proposed that the presence of a task induced stress condition causes speech production to deviate from this neutral path. As a result, speech recognition algorithms were formulated with the goal of minimizing the deviation between stress and neutral speech feature paths (i.e., a source generator equalization approach). Here, it is suggested that the perturbation in speech production under accent can also be represented using a similar framework (see Fig. 2), and that the deviation in neutral versus accent word production can be exploited to achieve reliable accent classification.

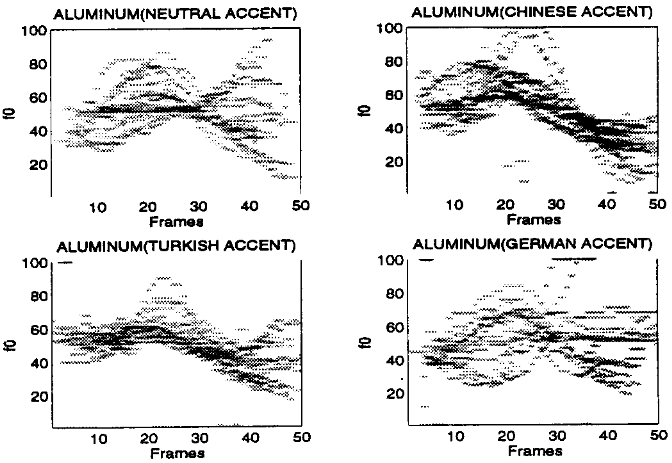
ALUMINUM(NEUTRAL ACCENT)
ALUMINUM(CHINESE ACCENT)
ALUMINUM(TURKISH ACCENT)
ALUMINUM(GERMAN ACCENT)

Figure 3: Gray-scale histograms of pitch contours for 4 different accents for the word *aluminum.*

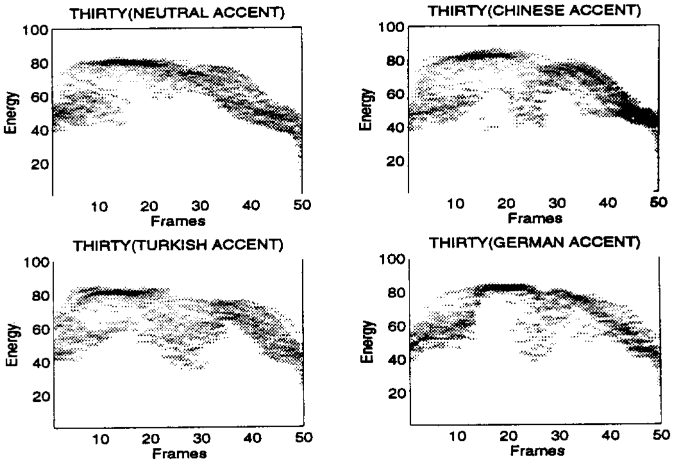
THIRTY(NEUTRAL ACCENT)
THIRTY(CHINESE ACCENT)
THIRTY(TURKISH ACCENT)
THIRTY(GERMAN ACCENT)

Figure 4: The gray-scale histograms of energy (in dB) contours for 4 different accents for the word *thirty.*

4. ACCENT DATABASE

In order to investigate accent, a vocabulary and phrase wordlist was established which contains accent sensitive phonemes or phoneme combinations. Vocabulary choice was based on our literature review of language education of American English as a second language. The data corpus was collected across an online telephone setup (ISDN network), using speakers from the general Duke University community. The test vocabulary consists of twenty isolated words (sample words include: aluminum, thirty, bringing, target, bird), and four test sentences. Available speech includes neutral American English, and English under the following accents: German, Chinese, Turkish, French, Persian, Spanish, Italian, Hindi, Romanian, Japanese, Russian, and others. For the studies conducted here, we focus on American English speech from twenty-seven speakers across the following accents: Turkish, Chinese and German.

5. PROSODIC FEATURE ANALYSIS

In order to characterize the change in speech production due to accent in an articulatory space, a series of experimental evaluations were performed on speech modeling features (i.e., see Fig. 2). The acoustic features which were considered include: frame power, zero-crossing rate, LP reflection coefficients, autocorrelation lags, log-area-ratios, line-spectral pair frequencies, LP cepstrum coefficients, FFT cepstrum coefficients, f_0, formants locations and bandwidths. The first-order differenced parameters which represents the "velocity" information was also considered. These features are represented as $\vec{\mathbf{X}}_i, i = 1, \ldots, N$ in the general accent classification algorithm from Fig. 1. An examination of these features across a generator partition was conducted to informally determine the effectiveness of these features as accent relayers. These results will be discussed in conjunction with classification evaluations.

6. ACCENT CLASSIFICATION SYSTEM

Next, an accent classification system was formulated as shown in the flow-diagram in Fig. 1. The algorithm assumes that continuous speech is sampled at 8kHz, high-frequency preemphasis is performed, Hamming windowed, followed by acoustic feature extraction on a frame-by-frame basis. Classification was based on a sequence of 3-state hidden Markov models (HMM's) having single Gaussian densities. The Baum-Welch algorithm was used to train source generator parsed monophone models for a set of 4 different accents (i.e., neutral, Turkish, Chinese, German). The HMM codebooks generated after training are used in the classification phase in order to obtain a score of how likely each accent codebook is 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. $P(\mathbf{X}|Accent_i), i = 1, \ldots, 4)$. Finally, the accent resulting in the maximum probability is selected as the test speaker's accent (i.e., $argmax(P(\mathbf{X}|Accent_i), i = 1, \ldots, 4)$.

7. EVALUATIONS

In order to investigate the influence of accent on source generator parsed prosodic features, we considered an analysis of normalized pitch and energy (in dB) contours, and spectral structure for a sequence of accent sensitive words. Fig. 3 illustrates 4 gray-scale histograms of pitch contour for the word *aluminum* under 4 different accents for the speaker set. The histograms were generated after time-aligned of each waveform. Although the histograms can be distinguished from each other visually, there is measurable inter-speaker variability caused by factors other than accent, such as stress, emotion, and general intonation. It is clear that pitch contour variation differs between American English and across accent in varying degrees.

In Fig. 4, gray-scale histograms of energy contours are shown for the word *thirty* under the same four accents. The differences in these contours are more clearly defined than those for pitch contours. The energy level drop towards the middle of each contour indicates the pause duration between /er/ and /d/-/iy/ sounds while pronouncing the word *thirty*. We can conclude that the American English speakers produce very little pause, while Turkish (and to a lessor degree Chinese and German) speakers insert longer pause duration between these syllables.

Next, we considered the change in mean formant structure for vowels under neutral and accented speech conditions. Vowel codebooks we obtained from source generator parsed speech for the 4 accents under consideration. The codebooks contain only those vowels which are available in the accent sensitive vocabulary set. Fig. 5 shows mean F_1 vs. F_2 frequencies for neutral American English and accent classes. The associated vowel bar graph shows a quantitative measure of how accent effects the centroid in the F_1, F_2 space. Here, the standard deviation among the four accents is summarized. It can be seen that the English and Chinese accent /ae/ sounds are separated from German and Turkish accented /ae/ sounds. This confirms the fact that there is no /ae/ phoneme present in Turkish and German, and these speakers tend to substitute the /e/ sound. The /uh/ sound is found to be the most sensitive phoneme for the four accents considered in this study.

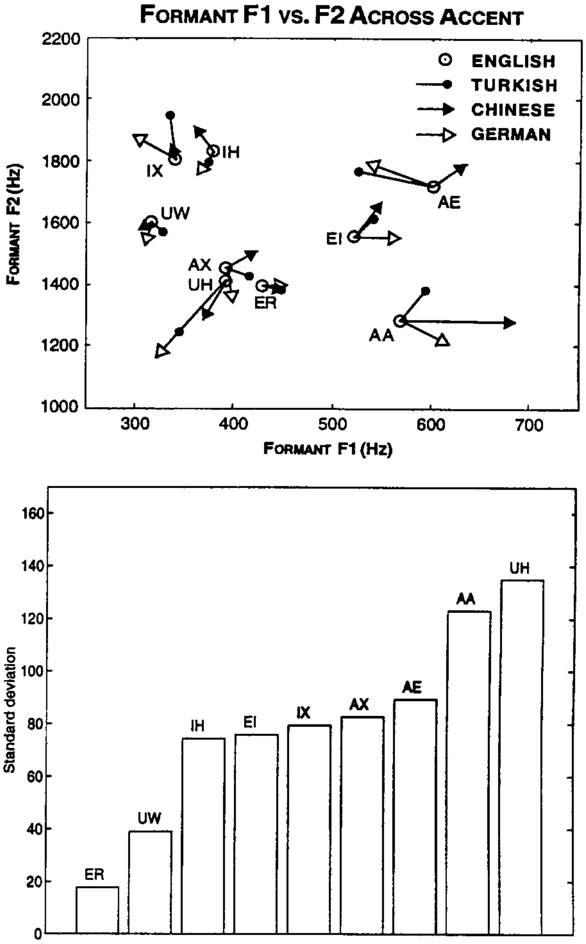

Figure 5: A first formant vs. second formant plot for selected vowels in the accent sensitive vocabulary. A summary of the standard deviation from the neutral accent is also shown.

Next, a series of accent classification experiments were performed to illustrate which speech features might be the most successful accent relayers for the given vocabulary set. These studies focused on a five word vocabulary set, and a simple five-state continuous distribution hidden Markov model recognition system. Table 1 summarizes the general overall classification ability of time domain frame energy (RMS Energy), f_0, and individual formant structure for neutral versus Turkish accented American English. The original feature, its' deriva-

EVALUATION OF PROSODIC PARAMETERS FOR ACCENT DETECTION			
Features	*Original*	*Derivative*	*Orig. w. 0-mean*
RMS Energy	+	+	-
Pitch	o	o	o
First Formant	+	+	-
Second Formant	+	-	+
Third Formant	+	+	o
Fourth Formant	+	-	+

Table 1: A summary of the evaluation of prosodic parameters, their derivatives, and their zero means in detecting neutral accent vs. Turkish accent over a set of 5 words(*aluminum, boy, catch, communication, would*).

CLASSIFICATION SCORES BETWEEN NEUTRAL AND TURKISH ACCENTS	
WORDS	*SCORES*
communication	98
thirty	94
teeth	86
there	100
line	92

Table 2: Scores for classifying neutral vs. Turkish accent in English using 12 prosodic parameters(*pitch, energy, first 4 formants and bandwidths, max. value of cross-correlation, ratio of first and second cross-correlations*) over a set of 5 words.

tive (sometimes referred to as 'velocity'), and its' original with zero mean normalization are indicated. These results suggest that specific knowledge of formant location for a predefined vocabulary set can be successful in classifying accented speech. Also, that particular words in the chosen vocabulary set are better predictors of accent. In Table 2, classification scores are presented for American vs. Turkish accent in English using 12 prosodic parameters (*HMM-S.G.Pro12*: *pitch, energy, first 4 formants and bandwidths, maximum value of cross-correlation, ratio of first and second cross-correlations*) over the set of 5 words. Some phoneme sequences are therefore more effective in indicating the presence of these accents.

Finally, the *HMM-S.G.Pro12* algorithm was evaluated for three conditions (results summarized in Table 3). First, (i) a single sentence test was conducted which include phonemes outside the source generator accent training set, with most phonemes contextually different than the trained monophone set. In this case, the summation of the probabilities across the test sentence was used in the final accent decision process. Here, the accent classifier achieved a recognition rate of 81.5%. For the second evaluation, (ii) the utterances tested were 3 known isolated words (communication, target, change) present in the available vocabulary. The resulting overall accent classification rate was 88.9%, which was obtained by summing the output probabilities for the three words. In a final evaluation, (iii) a sequence of randomly selected isolated words were submitted to the accent classifier (word count ranging from 1 to 15). After, 6-7 words, the accent classification rate levels off (at 6 words: 92.1%). This suggests that as additional text is made available, the ability to reliably classify accent will increase to some limit, which may be dependent on accent type.

EVALUATION OF PROPOSED ACCENT CLASSIFICATION ALGORITHM	
Features	*Overall %*
Test case (i)	81.5%
Test case (ii)	88.9%
Test case (iii)	92.1%

Table 3: A summary of the evaluation of the proposed accent classification algorithm: (i) unknown text sequence, (ii) known three word sequence, and (iii) random selected 7 word sequence.

8. CONCLUSIONS

In this study we have considered the effects of accent on prosodic features derived under a source generator framework. An accent sensitive speech database was established, and an HMM classification algorithm formulated to establish effective prosodic and acoustic features for accent classification. While a high degree of variability across speakers exists for pitch structure, some global aspects can be used for accent classification. It is also suggested that energy, duration, and spectral information are better features for accent detection. The most distinct features of accent are at the phonemic level, with phoneme substitution being a consistent identifier of language accent. Finally, the accent classification system was evaluated for American English under neutral, German, Turkish, and Chinese accent. For an unknown open speech sequence, an accent classification rate of 81.5% was achieved. This rate increases to 88.9% when the test is limited to a known three-word set. Finally, it has been shown that as accent sensitive word count increases, the ability to correctly classify accent also increases.

References

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

[2] F.M. Chreist. *Foreign accent.* Prentice-Hall, Eaglewood Cliffs, N.J., 1964.

[3] C. Grover, D.G. Jamieson, and M.B. Dobrovolsky. "Intonation in English, French and German: perception and production". *Language and Speech*, 30-3:277–295, 1987.

[4] J.H.L. Hansen, "Adaptive Source Generator Compensation and Enhancement for Speech Recognition in Noisy Stressful Environments," *IEEE ICASSP-93*, vol.2, pp.95-98, April 1993.

[5] J.H.L. Hansen, "Morphological Constrained Enhancement with Adaptive Cepstral Compensation (MCE-ACC) for Speech Recognition in Noise and Lombard Effect," *IEEE Trans. on Speech & Audio Processing, SPECIAL ISSUE: Robust Speech Recognition*, vol.2, no.4, pp. 598-614, Oct.1994.

[6] A.S. House, E.P. Neuburg, "Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations," *J. Acoust. Soc. Am.*, 62(3):708–713, Sept.1977.

[7] A. Ljolje and F. Fallside. "Recognition of isolated prosodic patterns using hidden Markov models". *Computer Speech and Language*, 2:27–33, 1987.

[8] Y.K. Muthusamy, R.A. Cole, "Automatic Segmentation and Identification of Ten Languages Using Telephone Speech," *ICSLP-92*, pp. 1007-1010, Alberta, Canada, Oct.1992.