

FREQUENCY CHARACTERISTICS OF FOREIGN ACCENTED SPEECH

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

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

<http://www.ee.duke.edu/Research/Speech> arslan@epiwr12.entropic.com jhlh@ee.duke.edu

ABSTRACT

In this study, frequency characteristics of foreign accented speech is investigated. Experiments are conducted to discover the relative significance of different resonant frequencies and frequency bands in terms of their accent discrimination ability. It is shown that second and third formants are more important than other resonant frequencies. A filter bank analysis of accented speech supports this statement, where the 1500-2500 Hz range was shown to be the most significant frequency range in discriminating accented speech. Based on these results, a new frequency scale is proposed in place of the commonly used Mel-scale to extract the cepstrum coefficients from the speech signal. The proposed scale results in better performance for the problems of accent classification and language identification.

1. INTRODUCTION

In general, the presence of foreign accent can degrade the quality and intelligibility of speech. In a speech communication system, it can be regarded as a type of interference such as actual background noise. Foreign accent causes changes in intonation and lexical stress patterns as well as variation of the acoustic structure in time and spectral domains. In addition, phoneme substitutions, additions, and deletions are often observed in non-native speaker's speech. As a result, a significant amount of degradation in intelligibility can result due to the presence of foreign accent. In general, noise and channel characteristics are constant over short time intervals and can be characterized reasonably well in a communication scenario. However, distortion of speech due to foreign accent is highly dependent on context which requires a complete understanding of acoustic variations at both the linguistic and the semantic levels [1].

In this study, we analyze accented speech in terms of its frequency characteristics. First, a description of the foreign accent database is given where the acoustic analysis is performed. Secondly, we present a study on the frequency analysis of foreign accent. In this last section, the validity of using the Mel-scale for parameterization in accent classification is questioned, and a more appropriate scale for accent classification is proposed.

2. ACCENT DATABASE

In order to investigate accent, a vocabulary of words was established which contains accent sensitive phonemes or

phoneme combinations [3, 4, 2]. Vocabulary choice was based on a literature review of language education of American English as a second language. The data corpus was collected using a head-mounted microphone, from speakers among the general Duke University community. The test vocabulary consists of twenty isolated words (sample words include: aluminum, thirty, bringing, target, bird). 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 forty-eight speakers across the following accents: neutral, Turkish, Chinese and German.

3. FREQUENCY CHARACTERISTICS

In order to formulate better speech recognition algorithms, it is 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 its relative frequency. Empirical evidence suggests that the human ear is more sensitive to absolute changes in low frequency signals. After extensive experimental analysis, the Mel-scale was formulated for the sampling of the frequency axis based on perceptual criteria [7]. Speech features derived using the Mel-scale have also resulted in superior speech recognition performance when compared to parameters obtained from a linear scale [5].

In this study, 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 a native speaker as possible. 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 resonance frequencies and frequency bands for both speech recognition

*Dr. Arslan was with the Robust Speech Processing Lab when this work was performed. He has since joined the Speech Research Group at Entropic Corp., Washington, D.C.

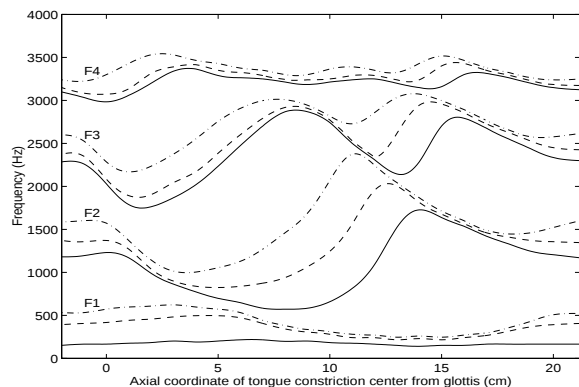


Figure 1: The influence of the place of tongue constriction and constriction area on formant frequencies. Each of the three curves for each formant frequency represent cross sectional contours corresponding to different cross sectional areas at the place of constriction (- - : $A = 8.0 \text{ cm}^2$ - . : $A = 2.0 \text{ cm}^2$ — : $A = 0.16 \text{ cm}^2$). Adopted from Fant (1970).

and accent classification. The following sections discuss the experimental set-up followed by their results.

3.1. Formant Frequencies

In general, non-native speakers have difficulty in changing learned articulatory movements of their own language when learning a second language. Fant [6] 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 Fig. 1, the measurements of formant frequencies based on an electrical line analog (LEA [6]) 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 value 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 have more problems with detailed tongue movements, and 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 for the /R/ sound. In Fig. 2, 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 /R/ sound which suggests early oral cavity closure resulting from the tip of the tongue touching the hard palate and sliding back. However, for most non-native speakers the tongue does not touch the hard palate until the very last moment in the production

of the /R/ sound, which causes some degree of separation between these two formant frequencies.

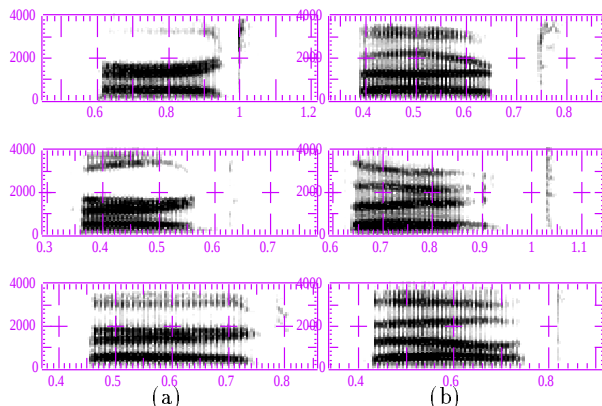


Figure 2: Illustration of the influence of accent on F_2 - F_3 separation in /ER/ sound in *bird*: (a) three native speakers (b) three non-native speakers.

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 (48 speakers x 20 words x 5 tokens). Next, the first four formant frequencies were estimated for each time frame. A hidden Markov model (HMM) was generated 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 ΔF_1 parameters of the word *thirty* from the Turkish training speaker set). Next, the HMM based accent recognizer using formant structure was evaluated. Open test accent classification results for the first four formant frequencies are shown in Fig. 3a. Using the HMM set trained with 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 Fig. 3b. When accent classification and speech recognition performance are compared, F_2 was 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.

3.2. Filter Banks

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 Fig. 4. The energy in each frequency band was weighted with a triangular window. Next, the output of each filter bank was 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 neutral, Turkish, Chinese, and German accents 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

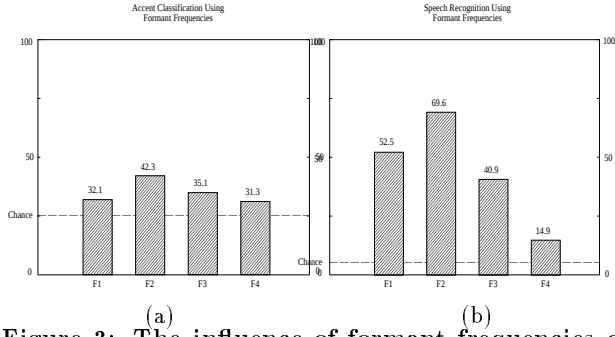


Figure 3: The influence of formant frequencies on the performance of (a) accent classification and (b) speech recognition.

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.

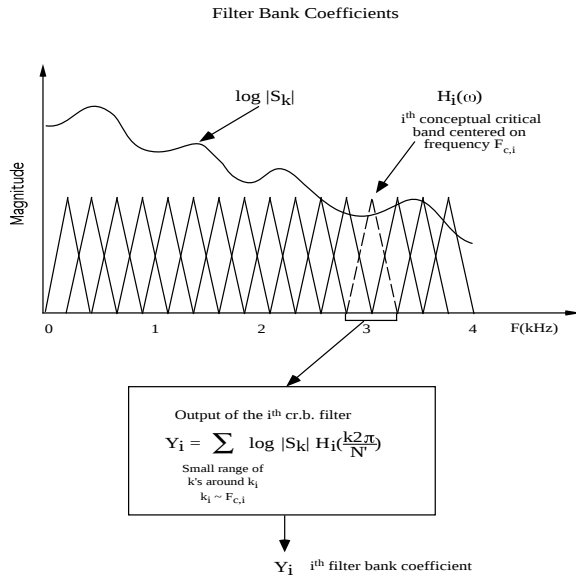


Figure 4: The extraction of filter bank coefficients.

In Fig. 5, plot (a) shows accent classification performance across the 16 linearly spaced frequency bands. In order to compare accent classification performance to speech recognition performance across the linear frequency band, a second experiment was performed. Using only neutral trained American English HMMs 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 Fig. 5b. From the graphs, it can be concluded that the impact of high frequencies on both speech recognition and accent classification performance is limited. However, mid-range frequencies (1500-2500 Hz) contribute more to accent classification performance than to speech recognition whereas low frequencies improve speech

recognition performance more than accent classification performance. 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. In addition, the first formant F_1 which was shown not to be as significant in accent discrimination corresponds to lower frequencies in Fig. 5a.

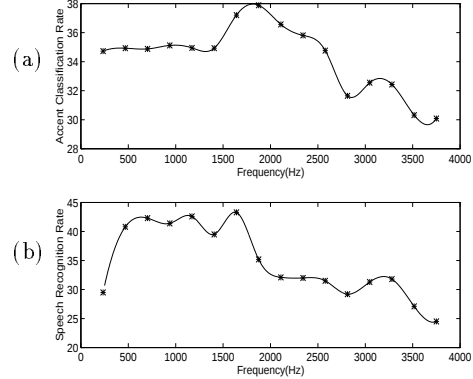


Figure 5: Comparison of accent classification versus speech recognition performance based on the energy in each frequency band.

The Mel-scale, which is approximately linear below 1 kHz and logarithmic above [7], is more appropriate than a linear scale for speech recognition performance across frequency bands. However, the results in Fig. 5a suggest that 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 Fig. 6. Since a larger number of filter banks are concentrated in the mid-range frequencies, the output coefficients are better able to emphasize accent-sensitive features. The 16 center frequencies of the filter bank which range between 0-4 kHz are also given in Fig. 6 (note, a symmetric triangular filter bank is employed).

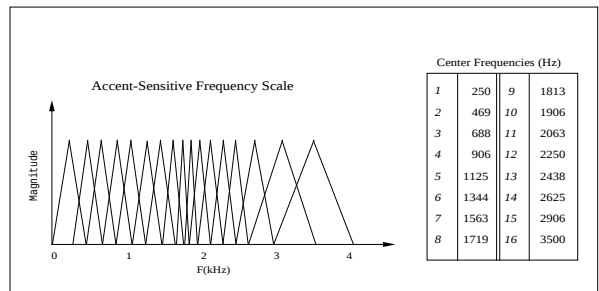


Figure 6: A new sampling scheme for the filter banks which is more sensitive to accent characteristics.

4. EVALUATIONS

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 (neutral, German, Turkish, Chinese): i) cepstrum coefficients derived from a linear uniformly sampled filter bank, ii) Mel-frequency cepstrum coefficients, and iii) cepstrum coefficients derived from the new accent-sensitive filter bank sampling. 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 1. 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 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 1.

COMPARISON OF DIFFERENT FREQUENCY SCALES IN ACCENT DISCRIMINATION			
	Linear	Mel	Accent-sens.
Accent Classification %	55.4	57.1	58.3
With Delta	60.0	60.7	61.9

Table 1: Comparison of the linear scale, Mel-scale, and accent-sensitive scales in terms of their accent classification performance among neutral, Chinese, Turkish, and German accents.

Since accent is a result of differences in first language background, the new accent-sensitive scale developed here might also be useful in language discrimination. The Oregon Graduate Institute Multi-Language Telephone Speech (OGI-TS) Corpus [8] was used to evaluate the performance of the accent-sensitive scale on a language identification task. In our evaluations, the initial training set (about 50 speakers per language) was used in training, and the development test set (about 20 speakers per language) was used for testing.

In the language ID system, a Gaussian Mixture Model with 64 mixtures is employed for each language. The feature set used in the system was comprised of 8 cepstrum coefficients, 8 delta cepstrum coefficients, and delta energy. Table 2 includes language classification error rates comparing the use of the Mel-scale and accent-sensitive scale in cepstrum coefficient calculation. The average error rate was reduced from 16.7% to 15.7% (a 6% reduction) after accent-sensitive scale was used in parameter estimation. The performance degraded for German, Spanish, and Tamil, whereas it improved for the rest of the languages (for French, performance did not change). It is interesting to note that in general, the improvement was achieved on languages that did not belong to the same language family as English. Therefore the proposed feature set might be useful for discriminating between language families which will reduce the number of possible languages in a multi-stage

language ID system.

<i>Error rates in classification of English versus other languages in OGI multi-language database</i>		
<i>Test pair</i>	<i>MFCC</i>	<i>ASCC</i>
English-Farsi	16.2	16.2
English-French	15.8	15.8
English-German	21.6	29.7
English-Japanese	14.3	8.6
English-Spanish	16.2	21.6
English-Korean	16.7	8.3
English-Mandarin	15.2	8.1
English-Tamil	13.5	16.2
English-Vietnamese	19.4	16.7
Overall	16.7	15.7

Table 2: Language ID performance improvement after using accent-sensitive scale cepstrum coefficients (ASCC) instead of Mel-frequency cepstrum coefficients (MFCC) for pairwise English-Other experiments.

5. CONCLUSION

In this study, we performed a detailed frequency analysis of accented speech. A new frequency scale is formulated for cepstrum coefficient calculation based on frequency analysis of the accent database. The new scale resulted in better performance for accent classification and language identification problems. In the future, a filter bank analysis for language identification problem will be conducted, and the significant frequencies for this problem will be identified.

References

- [1] L.M. Arslan. *Automatic Foreign Accent Classification in American English*. Ph.D. thesis, Duke University, Durham, N.C., 1996.
- [2] L.M. Arslan, J.H.L. Hansen. "Foreign Accent Classification using Source Generator Based Prosodic Features". *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 836–839, Detroit, USA, May 1995.
- [3] L.M. Arslan, J.H.L. Hansen. "A study of temporal features and frequency characteristics in american english foreign accent". accepted to *J. Acoust. Soc. Am.*, December 1996.
- [4] L.M. Arslan, J.H.L. Hansen. "Language Accent Classification in American English". *Speech Communication*, 18(4):353–367, August 1996.
- [5] S.B. Davis, P. Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". *IEEE Trans. Acoust., Speech, Sig. Proc.*, 28(4):357–366, Aug. 1980.
- [6] G. Fant. *Acoustic Theory of Speech Production*. Mouton, Paris, France, 1970.
- [7] W. Koenig. "A new frequency scale for acoustic measurements". *Bell Telephone Laboratory Record*, 27:299–301, 1949.
- [8] Y.K. Muthusamy, R.A. Cole, and B.T. Oshika. "The OGI multilanguage telephone speech corpus". *Proc. Inter. Conf. Spoken Lang. Proc.*, pp. 895–898, Oct. 1992.