Phonetic Discriminative Power of /r/

Ryan J. Amundsen, Carol Espy-Wilson, Daniel Garcia-Romero, Xinhui Zhou

***Abstract*— It is well known that American English speakers use a variety of vocal tract shapes to produce the /r/ sound. The two extreme postures are the "retroflex" /r/ and the "bunched" /r/. Recent research has indicated that the variability in tongue posture can be captured by the relative spacing of the fourth and fifth formants (F4 and F5). In this study, the discriminative powers of the mel-scale filter bank energies that correspond to F4 and F5 were shown to be stronger than those of other filter banks. The existence of a correlated relationship between tongue posture and F4 and F5 is implicit in this observation.**

I. Introduction

SPEAKER recognition by machines requires speaker specific information from both the glottal source and the vocal tract. Research is being conducted to understand the acoustic signatures of this speaker specific information and has shown that speakers can produce certain American English sounds, in particular /r/, using a variety of different tongue postures [1]. It has also shown that acoustic signatures carry speaker specific information related to these variations. Figure 1 shows a block diagram of the speech production system with the vocal tract as the transfer function.

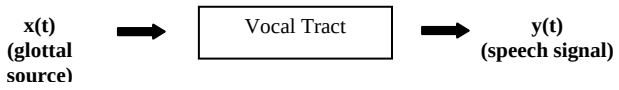
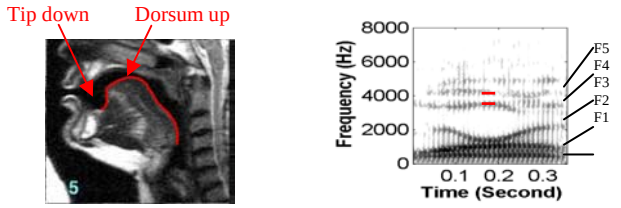


Figure 1. Speech Production System

This transfer function is altered depending on the tongue position utilized for an /r/ utterance. The most extreme articulations of /r/ are “bunched” and “retroflexed”. Figures 2 and 3 contrast the magnetic resonance images (MRI) and their respective spectrograms of two different speakers: the former employing the “bunched” tongue posture and the latter employing “retroflexed”. Figures 2a and 3a were attained from speaker 5 and speaker 22 in the University of Cincinnati database[2]. Vocal tract modeling studies show that the spacing between the fourth and fifth formants (F4 and F5) captures the difference in tongue shape [3]. This spacing is marked in red on each spectrogram.

In this study, we look at the discriminative power of vocal tract information from the American English /r/, particularly between the frequencies of 2.5 kHz and 4.5 kHz: the standard range of F4 and F5. Similar studies have been conducted focusing on only the first three formants of /r/ [4].



a.) b.)

Figure 2. (a)The MRI image of a “bunched” /r/ articulation of nonsense word “warav” is pictured on the left. It labels the tongue tip and dorsum positions. (b)The corresponding spectrogram is pictured on the right. The fourth and fifth formants are marked in red.

a.) b.)

Fig 3. (a)The MRI image of a “retroflex” /r/ articulation of nonsense word “warav” is pictured on the left. It labels the tongue tip and dorsum positions. (b)The corresponding spectrogram is pictured on the right. The fourth and fifth formants are marked in red.

II. Methodology

In this study, the mel-scale filter bank energies of every utterance of /r/ in the Ohio State University Buckeye Corpus are computed [5]. Organizing these by speaker, the discriminative power was then determined. The Mel-scale filter bank energy calculation process is shown in Figure 4.

After the magnitude of the short-term Fourier transform (STFT) of the input waveform is found, the resulting frequency spectrum is sent through a pre-emphasis filter and a mel-scale filter bank. The STFT employed has a window size of 20 ms with a 10 ms overlap. Each filter is spaced according to the plot in Figure 5.

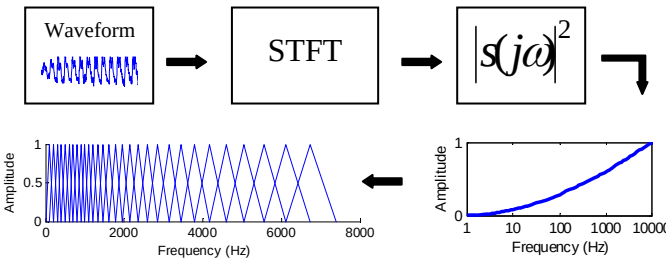


Figure 4. Mel-Scale Filter Bank Energy Calculation

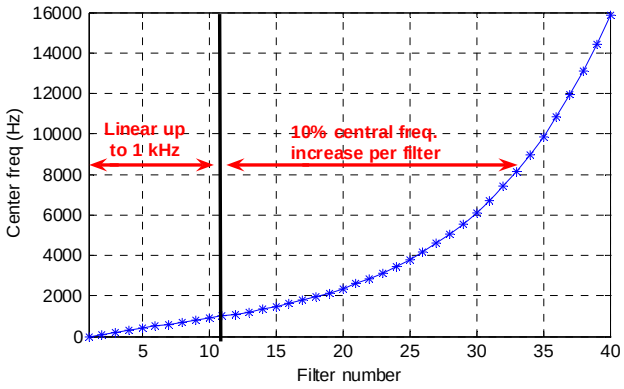


Figure 5. Mel-Frequency Scale

The signal energy of the output of each filter is one of the mel-scale filter bank energies for that utterance. For this study, only the first 31 energies were taken into account because the Buckeye Corpus is band limited to 8 kHz. A set of 31 energies is calculated for every frame of every utterance of /r/.

After sets of energies were calculated for each utterance, the discriminative power was determined for each energy band. Discriminative power is calculated using the following formulas where P_i is the probability of speech per speaker or class, M is the number of speakers, S_w is the within-class scatter matrix, S_i is the covariance matrix of class i, S_b is the between class scatter matrix, μ_0 is the global mean vector, and D_P is the discriminative power [6]:

$$S_w = \sum_{i=1}^{M} P_i S_i \quad (1)$$

$$S_i = E[(x-\mu_i)(x-\mu_i)^T] \quad (2)$$

$$S_b = \sum_{i=1}^{M} P_i(\mu_i-\mu_0)(\mu_i-\mu_0)^T \quad (3)$$

$$\mu_0 = \sum_{i=1}^{M} P_i \mu_i \quad (4)$$

$$D_P = \frac{diag(S_b)}{diag(S_w)} \quad (5)$$

Figure 6-8 illustrate the meaning of discriminative power. In these figures, three examples of a set of two feature data with three classes have been randomly generated. Table 1 details the discriminative power of each.

	Case 1	Case 2	Case 3
Feature 1	63.473	2.076	680.023
Feature 2	85.452	0.682	764.557

Table 1. Discriminative power of scatter plots in figures 6-8

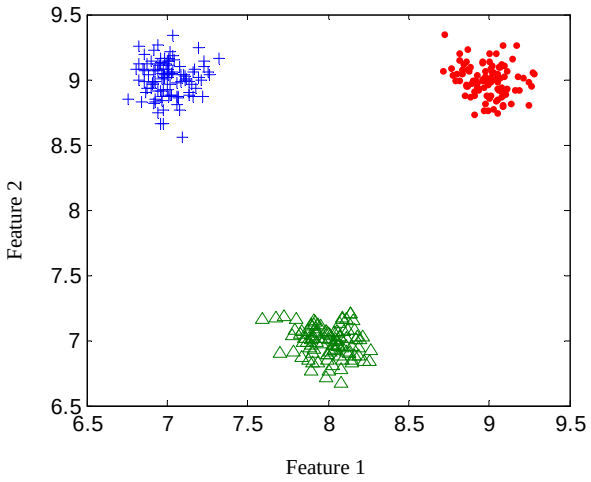


Figure 6. Case 1: Scatter plot of three classes with moderate between-class variability and moderate within-class variability

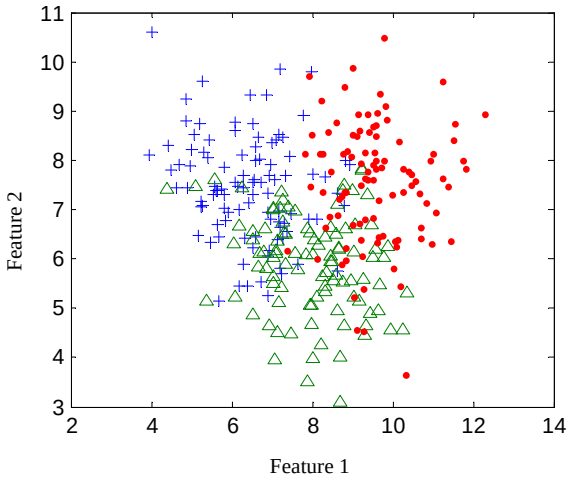


Figure 7. Case 2: Scatter plot of three classes with low between class variability and high within-class variability

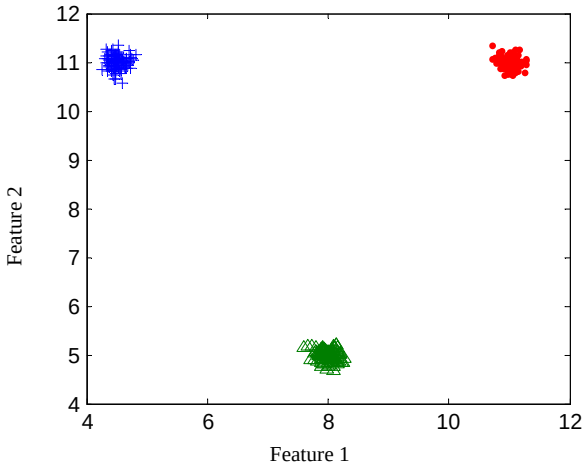


Figure 8. Case 3: Scatter plot of three classes with high between class variability and low within-class variability

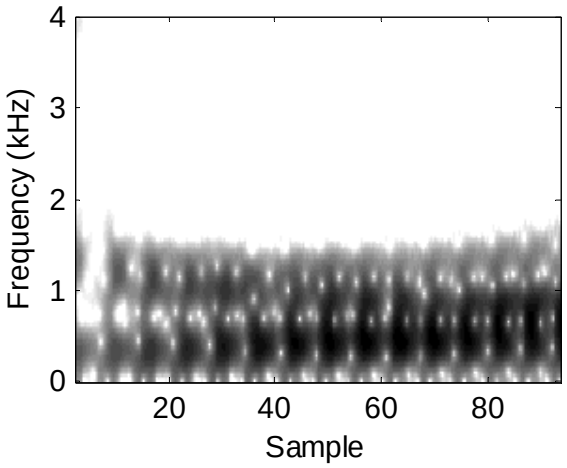


Figure 9. Spectrogram of male speaker's /r/ utterance in the word "destroying"

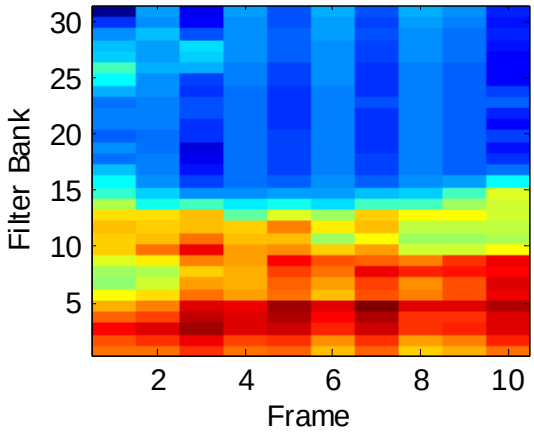


Figure 10. Mel-scale filter bank energies of male speaker's /r/ utterance in the word "destroying"

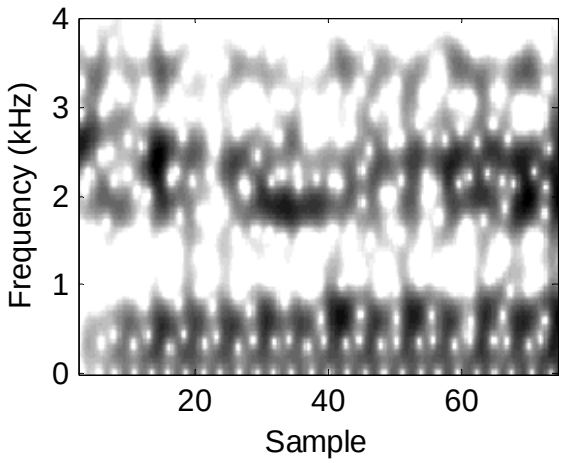


Figure 11. Spectrogram of female speaker's /r/ utterance in the word "injury"

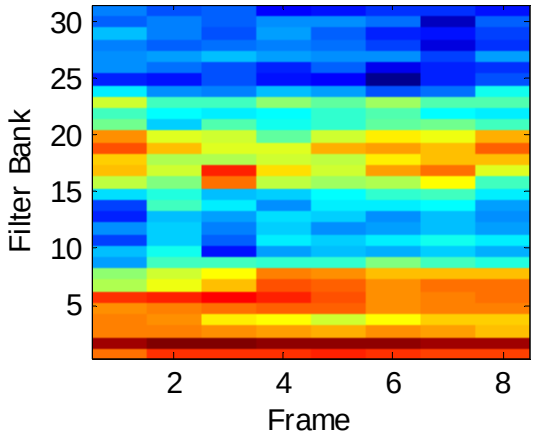


Figure 12. Mel-scale filter bank energies of female speaker's /r/ utterance in the word "injury"

III. Database

Ohio State University's Buckeye Corpus was chosen for this study for its unrehearsed conversational American English and its hand corrected phonetic transcriptions. It has over 300,000 recorded words from 40 speakers, 20 male and 20 female. It was also recorded at a 16 kHz sampling rate which includes more information than the typical 8 kHz telephone recording.

IV. Results

A visualization of the mel-scale filter bank energies of a male speaker's /r/ utterance in the word "destroying" is shown in Figure 10. Its corresponding spectrogram is shown in Figure 9. Notice the similarities in signal intensity. This comparison shows the mel-scale filter bank energies are an accurate representation of intensity content for a signal. The only difference is that the energies are parsed into 31 separate values per frame whereas a spectrogram has many more values per frame.

A similar comparison is given in Figures 11 and 12. Figure 11 shows the spectrogram of a female speaker's /r/ utterance in the word "injury" and Figure 12 gives the corresponding energies.

The discriminative power of /r/ in male speakers is given in Figure 13 and that of female speakers is given in Figure 14. The discriminative power is divided by sex because of the consistent difference in average pitch of male speakers versus female speakers.

V. Discussion

From Figure 13, the highest /r/ utterance discriminative power for male speakers lays between the filter banks 16 and 25. For female speakers, the most discriminative information lays between filter banks 18 and 27 (Figure 14). Respectively, these values correspond to the frequency ranges of 1611 Hz – 3797 Hz and 1949 Hz – 4595.

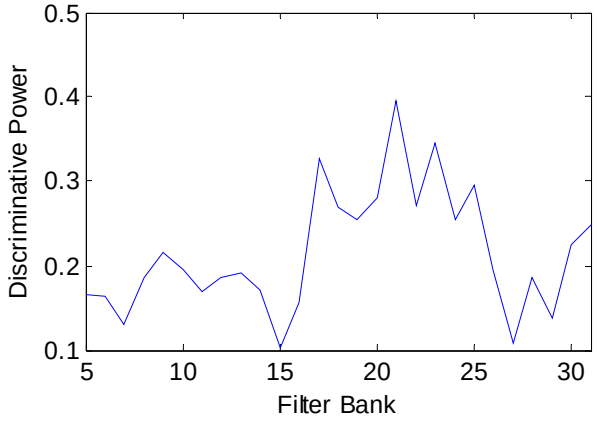


Figure 13. Discriminative power of /r/ in male speakers

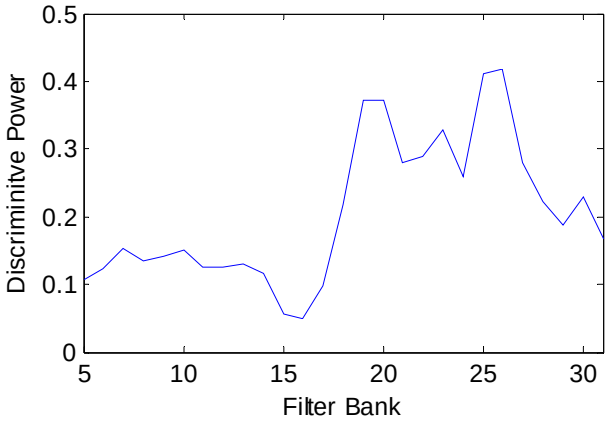


Figure 14. Discriminative power of /r/ in female speakers

According to Figure 15, the majority of fourth formants of the Buckeye Corpus range from 2644 Hz and 3594 Hz and the majority of fifth formants range from 3740 Hz and 4508 Hz. The range of the discriminative powers for both male and female speakers is roughly aligned to 2644 Hz – 4508 Hz, the combined range of the fourth and fifth formants of the Buckeye Corpus.

VI. Conclusion

Even though the average utterance of /r/ carries most of its energy in the regions of F1, F2 and F3, our results show that F4 and F5 have the most discriminative power among the first five formants for speaker recognition. These are the formants that show high variability depending on tongue posture as well. Hence, it can be inferred that there exists a strong relationship between tongue shape and F4 and F5

VII. Future Work

This conclusion paves the direction for many new studies. An understanding of how F4 and F5 directly relate to articulatory configurations of /r/ will be reached. This relationship will be quantified. A study similar to the study in this paper will be conducted with /l/ and its varying tongue postures. Afterwards

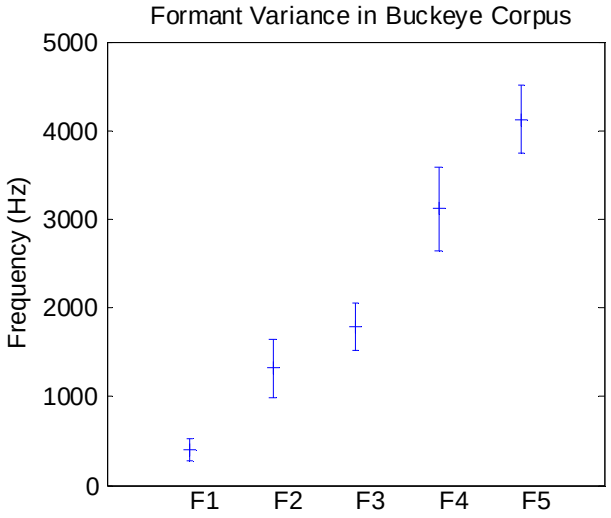


Figure 15. Means and standard deviations of the first five formants of /r/ utterances across all forty speakers in the Ohio State University Buckeye Corpus

these relationships, provided useful feasibility, will be integrated into a speaker recognition algorithm as one of many acoustic parameters included to improve accuracy.

Acknowledgment

Thank you to Dr. Carol Espy-Wilson for her guidance, Xinhui Zhou and Daniel Garcia-Romero, for their assistance and expertise, and Alec Colvin for a crash course in Unix programming. Without the preceding individuals, this project would not have been possible.

References

[1] Espy-Wilson, C. Y., Boyce, S. E., Jackson, M., Narayanan, S., Alwan, A., 2000. Acoustic modeling of American English /r/. Journal of the Acoustical Society of America 108 (1), 343-356.

[2] Tiede, M., Boyce, S. E., Holland, C., Chou, A., 2004. A new taxonomy of American English /r/ using MRI and Ultrasound. Journal of the Acoustical Society of America 115 (5), 2633-2634.

[3] Zhou, X. H., Espy-Wilson, C., Boyce, S., Tiede, M., 2007. An articulatory and acoustic study of "retroflex" and "bunched" American English rhotic sound based on MRI. In: Interspeech 2007

[4] Goldstein, U. G., 1976. Speaker-identifying features based on formant tracks. Journal of the Acoustical Society of America 59 (1), 176-182.

[5] Pitt, M. A, Johnson, K., Hume, E., Kiesling, S., Raymond, W., 2005. The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication*, 45: 89-95.

[6] Theodoridis, Sergios, and Konstantinos Koutroumbas. Pattern Recognition, 3rd ed. San Diego: Elsevier, 2006. 228-231.