Audio Engineering Society

Convention Paper 7272

Presented at the 123rd Convention
2007 October 5–8 New York, NY, USA

Experiment in Computational Voice Elimination using Formant Analysis

Durand R. Begault

[1] Audio Forensic Center, Charles M. Salter Associates, 130 Sutter St., San Francisco, CA 94104 USA
Durand.Begault@cmsalter.com

ABSTRACT

This study explores the use of a computational approach to the elimination of a known from an unknown voice exemplar in a forensic voice elimination protocol. A subset of voice exemplars from 11 talkers, taken from the TIMIT data base, were analyzed using a formant tracking program. Intra- versus inter-speaker mean formant frequencies are analyzed and compared.

1. DESCRIPTION OF EXPERIMENT

This study is an informal exploration of the use of a "computational approach" to differentiation between two or more voice exemplars from different or identical speakers, for use in a forensic voice elimination protocol. Normally, in a voice elimination protocol as described in Begault and Poza [1] and Gruber and Poza [2], an aural-spectrographic approach is used. In that approach, spectrographs are compared in a manner similar to that addressed by Tosi et al. [3] except that "critical listening" comparison between unknown and known exemplars is also performed. Aural-spectrographic analysis is common to most speaker-identification protocols; in reference [1,2] the known exemplars are elicited in a manner to have the talker imitate the unknown exemplar as closely as possible, so as to allow elimination on the basis of the spectrographic comparison without confounds resulting from inconsistent declamatory styles.

It would be extremely useful to use a "computational approach" to further buttress the results of such an analysis. The computational approach addressed here pools the locus of formants F1, F2 and F3 and creates a mean value and standard deviation for each speaker. In addition, the fundamental frequency is also analyzed. Plotting the loci of F1-F2 and F2-F3 mean frequencies provides an additional means of visualizing the data. The general proposition examined here is notion that the average value of formant frequencies, in terms of F1-F2

and F2-F3 relationships, will be significantly different when compared from different talkers, and more similar when compared to speech from the same talker.

This informal exploration of this so-called computational approach was inspired in part by the recent work of of Nolan and Grigoras [4] who have stated that "formants, whose frequencies and dynamics are the product of the interaction of an individual vocal tract with the idiosyncratic articulatory gestures needed to achieve linguistically agreed targets, are so central to speaker identity that they must play a pivotal role in speaker identification." The implementation of F1-F2 and F2-F3 analysis is embodied within the software developed by Grigoras (*Catalina*), that uses the formant tracking and spectrum analysis output of speech analysis software from the Speech, Music and Hearing department of the Royal Institute of Technology in Stockholm (KTH *Wavesurfer*) [5]. The present investigation has used both *Catalina* and 'manual' analysis based on the Wavesfurfer fomant data export.

The visual inspection of comparison spectrographs is a gestalt pattern matching activity on the part of the examiner that is inherently subjective. A computational approach may infer that an examiner can remove subjective interpretation from the process of comparing two samples, compared to an aural-spectral examination, by simply applying the analysis (Figure 1). Nevertheless, the forensic process still involves multiple levels of examiner interpretation (lower part of Figure 1), including decisions whether the speech recordings are of sufficient quality or adequate quantity; how material should be edited; and how to address the error rate of the technique (at this point unknown) and the criteria used for excluding or not excluding a match between exemplars. [1]

To explore how the computational approach might be helpful for forensic analysis, perhaps as a supplement to the aural-spectrographic technique, comparisons were made using voice exemplars that are at once similar and very different from actual forensic contexts. Male voice exemplars were harvested from the DR7 (American Midwest English) set from TIMIT speech data base developed some years ago by DARPA for development and evaluation of automatic speech recognition systems [6].

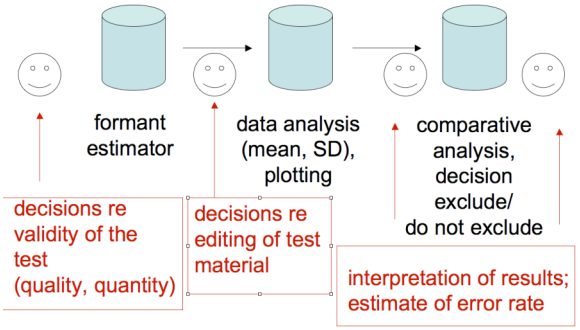

Figure 1. Abstraction of computational approach to speaker identification ('discs') with the intervention of subjective judgment within each stage ('faces').

The TIMIT exemplars used contain, for each speaker, two common spoken phrases and eight different spoken phrases, for a total duration of about 30 seconds. These voice exemplars differ from normal forensic exemplars in that the recording quality is quite good. They are further characterized by having very similar aural quality, which is often the case when making forensic comparisons. Common sentences to all speakers: "She had your dark suit and greasy wash water all year", and "Don't ask me to carry an oily rag like that". In addition, each of the eight speakers had 8 unique sentences (e.g., "she sounded as though they already existed"). Ideally, for computational analysis much longer samples would be available, but this is frequently not the case for forensic applications.

Analysis of TIMIT material has potential for forensic analysis investigation in that assumptions regarding forensic speech samples can be tested without the confounding effects of low quality, while allowing comparison of different voices with similar dialect. The current test was especially challenging from the standpoint that speakers MCHH and MESR are very similar aurally and in some cases spectrographically.

Other analyses are of interest. For instance, the fundamental frequency for eight speakers is shown in Figure 2. The similarity of the mean F0 and magnitude of the standard deviations illustrates the difficulty in basing speaker identification or discrimination on the basis of the mean fundamental frequency alone. By

[1] Tosi et al. [3] found a 6.4 % false identification and 11.8% false elimination error rate under restricted conditions for spectrographic, non-aural speaker analysis, based on 250 speakers, 29 examiners, and 34, 996 trials.

contrast, a temporal investigation may or may not reveal that MGRT has a more 'sung' quality (*Sprechstimme*) compared to a lower pitched, more monotone MKDR.

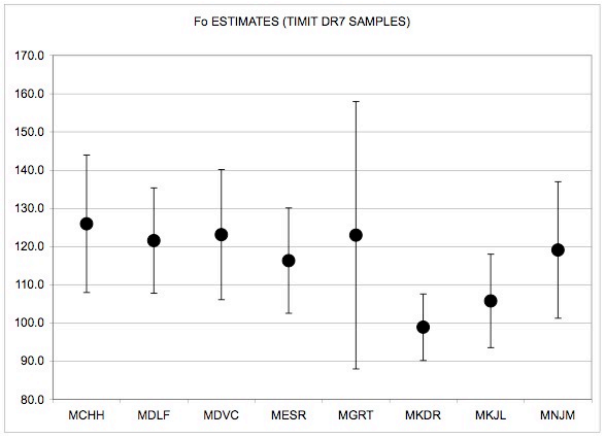

Figure 2. F0 plots (mean and standard deviation) for eight talkers from the TIMIT DR7 speech exemplars.

2. ESTIMATE BIAS DUE TO UNVOICED SPEECH (ESTIMATOR ERROR)

The following analyses illustrate the importance of removing unvoiced or silent portions of the waveform prior to analysis. Unvoiced phonemes such as stopped consonants *(p, t, k)* do not involve vibration of the vocal cords. A formant frequency estimator will produce erroneous values during these instances, as shown in Figure 3. These data should be removed prior to analysis by editing. [2]

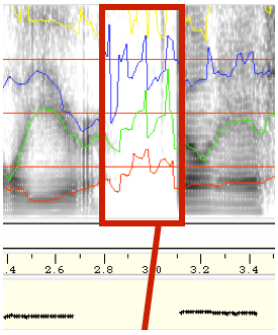

Figure 3. Errors produced by formant tracking analysis (boxed area).

[2] Grigoras' approach [4] cleverly bypasses this problem by only accounting for formant values when a fundamental frequency can be analyzed.

Figure 4 shows for a 6 s segment from one speaker the effect of not excluding unvoiced segments, on each analyzed frame, for the F1-F2 estimate. The solid triangles result when unvoiced segments are removed; the open circles are 'raw' data. The effect of removing unvoiced segments on removing outliers and tightening of the data distribution can be clearly seen. The same effect can be seen with three other speakers that were analyzed.

Figure 5 shows the effect on the mean values estimated for F2 (horizontal axis) versus F3 (vertical axis), for the same 6 phrase as analyzed in Figure 4. The filled symbols are for the condition with unvoiced speech present, the open symbols for the condition with unvoiced speech removed. The effect is a significant upward biasing of the values for F2 and F3

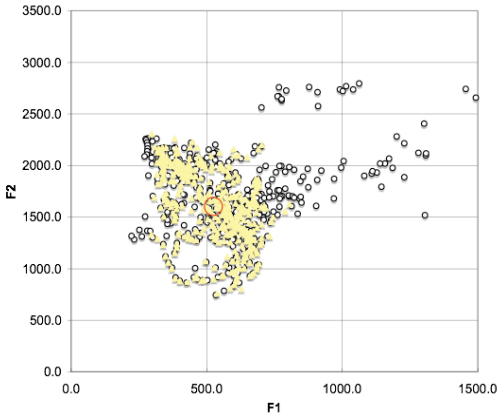

Figure 4. See text.

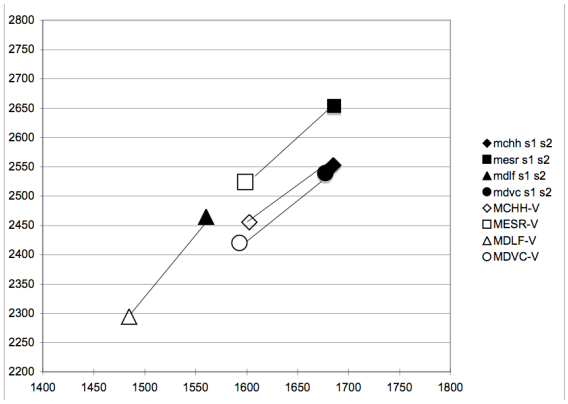

Figure 5. See text.

A similar upward biasing of the mean value occurs when comparing a relatively short exemplar (6 s) to a longer exemplar (30 s) for a single speaker. Figure 6 shows an overlay of Figure 4 of the long sample analysis. Figure 7 shows a comparison of short versus long sample effect on four speakers for F2-F3 analysis. Each symbol represents the same four speakers indicated in Figure 5. The lower mean value for each speaker occurs with the longer sample analysis, with the effect primarily on the estimation of the mean value of F3.

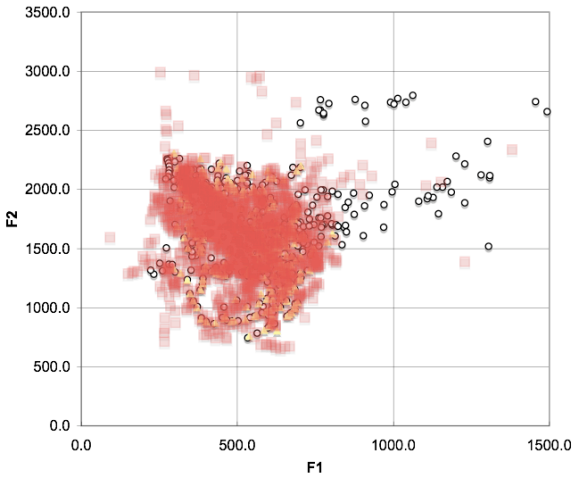

Figure 6. See text.

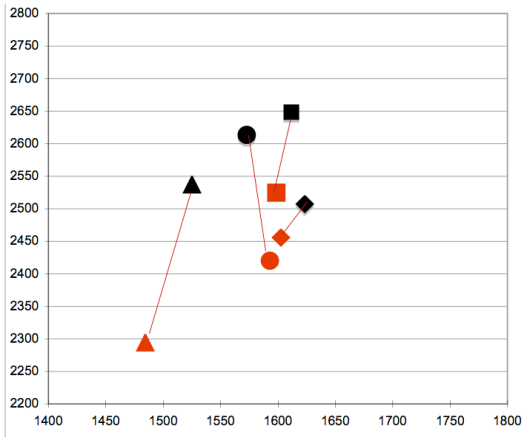

Figure 7. Horizontal axis = F2; vertical axis = F3. The lower value of F2 and F3 for each speaker results from a relatively longer sample (30 versus 6 s). See text.

3. DIFFERENT SPEAKER COMPARISON

Figure 8 shows a spectrographic comparison of the same material spoken by MCHH and MDLF. These two speakers are aurally more distinct than, for example, MCHH and MESR. Differentiation in the formant pattern is particularly evident in the phonemes corresponding in the words "had" and "greasy". Mean formant analysis of the entire 30 s exemplar for "vowel categories" *o, a, e* and *i* are shown in Figure 9[3].

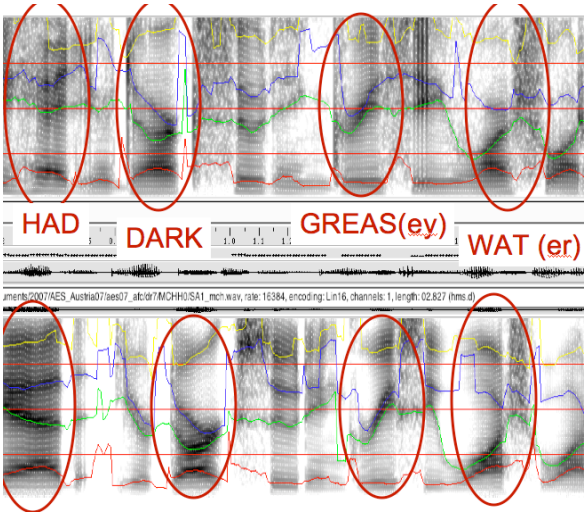

Figure 8. Inter-speaker variability, MCHH and MDLF, spectrogram (frequency range is 0-4 kHz in all spectrograms shown; horizontal red lines cross at 1, 2 and 3 kHz).

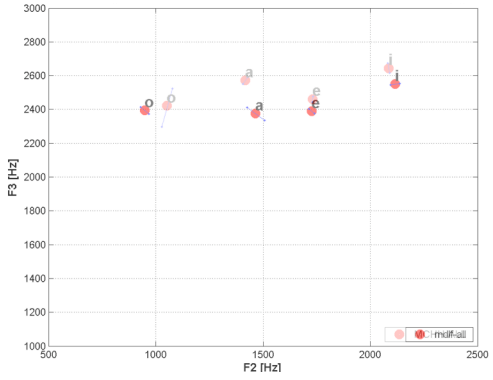

Figure 9. Inter-speaker variability, MCHH and MDLF, F2-F3 comparison plot for all 10 exemplars (apx. 30 s).

[3] Plots of F2-F3 for vowels were produced using Grigoras' *Catalina* software. Vowel 'frequency ranges' are prescribed and are therefore categorical as opposed to 'true' vowels.

Figure 10 shows a spectrographic comparison of the same material spoken by MCHH and MESR, who are aurally similar. Mean formant analysis of the entire 30 s exemplar for vowel categories *o, a, e* and *i* are shown in Figure 11. Differentiation in the formant pattern is less evident in the spectrograph (e.g., "dark", "water") and greater overlap occurs between vowel categories *a* and *e* in Figure 11, compared to Figure 9. A similar result occurs when investigating the F1-F2 intersections graphically.

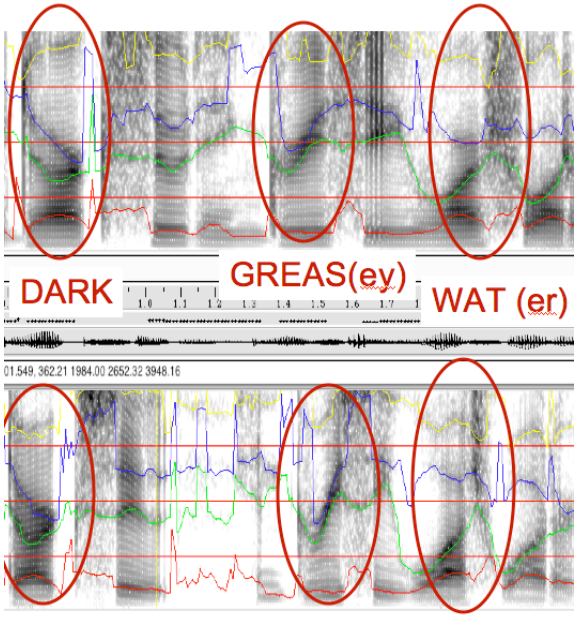

Figure 10. Inter-speaker variability, MCHH and MESR, spectrogram (frequency range is 0-4 kHz in all spectrogams shown).

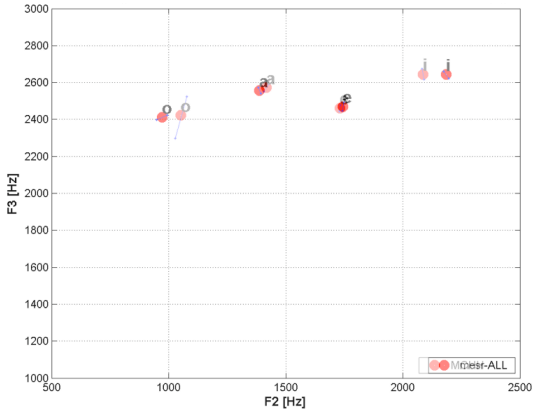

Figure 11. Inter-speaker variability, MCHH and MESR. F2-F3 comparison plot.

4. SAME VERSUS DIFFERENT SPEAKERS

Next, the performance of computational analysis of speech exemplars from different versus identical speakers was compared. The ten TIMIT phrases used previously for MCHH and the ten used for MESR were again used. These phrases were divided into separate sound files comprised of five exemplars for each speaker, each approximately 15 s in duration. Between MCHH and MESR, there was one identical phrase and four unique phrases between each group. As a result, the following sound files were formed:

- MCHH group 1 (exemplars 1-5)
- MCHH group 2 (exemplars 6-10)
- MESR group 1 (exemplars 1-5)
- MESR group 2 (exemplars 6-10)

The hypothesis tested was that within-speaker comparisons (MCCH group 1 compared to MCHH group 2; and MESR group 1 compared to MESR group 2) would exhibit greater similarity than between-speaker comparisons (MCCH group 1 compared to MESR group 1; and MCHH group 2 compared to MESR group 2). Alternatively, the inter-speaker versus intra-speaker variability would appear to be of about the same magnitude. F2-F3 intersections were examined since F3 is commonly assumed to indicate individual characteristics.

Figures 12-13 indicate the intra- and inter-speaker results for comparisons of each group. Figure 12 shows that, within speakers, there is either a slight difference between the two groups for some vowel categories, while other vowel categories (*e, i*) appear nearly the same. In Figure 13, MCHH versus MESR exemplars are compared for each group. There are only slightly greater differences for group 1 (top of Figure 13) than between the intra-speaker comparisons in Figure 12. However, the results for inter-speaker variability for group 2 in Figure 13 are nearly identical for the vowel categories *o, a,* and *e*; and in fact appear more 'dead on' than found in the intra-speaker comparisons.

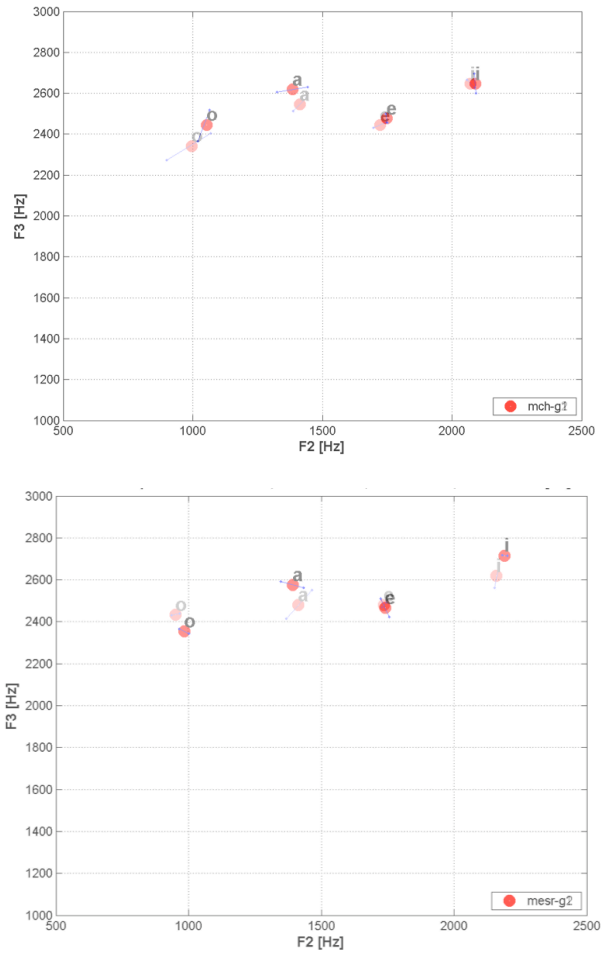

Figure 12. Intra-speaker variability, Top: MCCH group 1 compared to MCHH group 2. Bottom: MESR group 1 compared to MESR group 2.

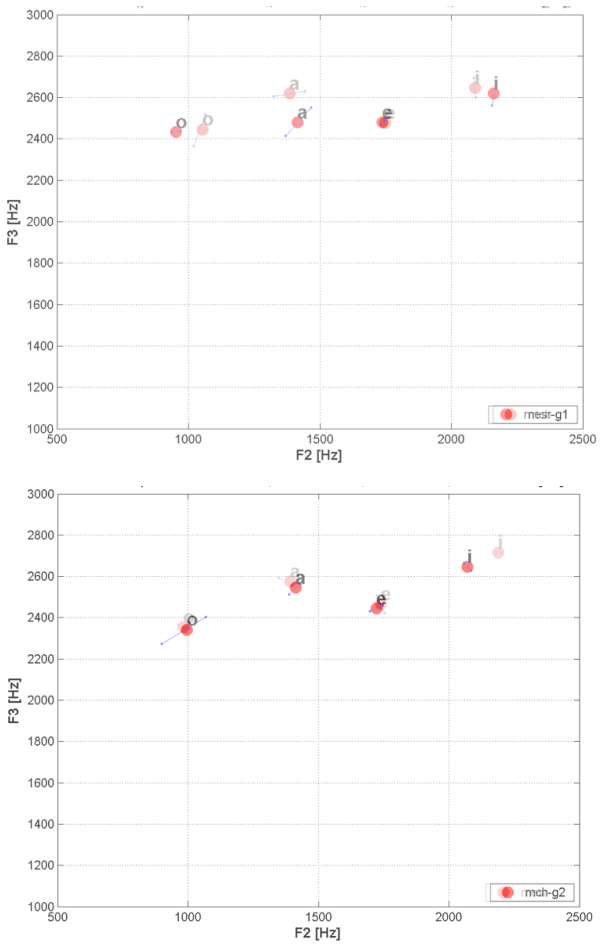

Figure 13. Inter-speaker variability. Top: MCCH group 1 compared to MESR group 1. Bottom: MCHH group 2 compared to MESR group 2.

5. CONCLUSION

Overall, within the very restricted material analyzed here, the results overall do not support the hypothesis that the statistical approach to determining mean formant frequencies will allow discrimination of same versus identical speakers. This can partially be explained by inadequate sample size. Figure 7 showed that longer versus short samples (6 s versus 30 s) result in a consistent upward bias for the estimation of F3. Here, 15 s of material was used. Longer samples may converge more towards 'true' means, and more precise means of analyzing formants for specific vowels as opposed to vowel categories may or may not improve the technique.

This paper was meant to give an exploratory look at the use of computational analysis of mean formant analysis using controlled material (aurally similar speakers with full bandwidth). Other examples not shown here have given confirmatory support to exclusion based on aural-spectral analyses. Further study is recommended, particularly with material more representative of forensic recordings.

6. REFERENCES

[1] Poza, F., and Begault, D. R. "Voice Identification and Elimination using Aural-Spectrographic Protocols" *Audio Engineering Society 26th International Conference, "Audio Forensics in the Digital Age", July 2005,* pp. 21-28

[2] Gruber, J., and Poza, F. "Voicegram Identification Evidence", In *American Jurisprudence: Trials*, 54. Thompson (1995).

[3] Tosi, O., Oyer, H. Lashbrook, W., Pedrey, C., Nicol, J. and Nash, E. "Experiment on Voice Identification". J. Acoust. Soc. Am. 51, 2030 (1972).

[4] Nolan, F. and C. Grigoras, C. "A case for formant analysis in forensic speaker identification" *International Journal of Speech, Language and the Law* 12(2), 143-173 (2005).

[5] http://www.speech.kth.se/wavesurfer

[6] http://www.ldc.upenn.edu/Catalog/readme_files/timit.readme.html