# Robust Formant Estimation: Increasing the Reliability by Comparison among Three Methods

A. Kabir, J. Barker, and M. Giurgiu

***Abstract*—**This research assesses the reliability of three widely used methods in order to estimate formant frequencies of eleven monophthongs of British pronunciation of English for 18 male and 16 female speakers whose speech is available in the GRID corpus. In particular, we are interested in a technique that is robust enough to estimate accurate and reliable formant frequencies over a large speech corpus. This paper provides a solution whether to trust a single method or to consider multiple methods while robust estimation of formant frequencies. Reliability of each method is judged by comparison with the standard formant values accepted previously in the science literature.

***Keywords*—**Formant Estimation, GRID Corpus.

## I. Introduction

THE formants, the resonance frequencies of the vocal tract, and their trajectories which describe the contours of energy concentrations in time and frequency, are believed to be one of the most useful features in speech signal, and used for experiments in many areas, including speech recognition, speech characterization, speech synthesis, and speaker identification [1, 2, 3]. It is well known that first two or three formant frequencies are adequate for perceptual identification of vowels. The formant representation is attractive because of its simplicity to represent the speech signal with a very small number of parameters [4].

The speech signal can be described in many ways that are related to speech production and speech perception, and the formants have been successfully applied to both speech production and speech perception [4]. Numerous attempts have been made to employ formant representations in speech technology applications such as speech synthesis, speech coding and automatic speech recognition (ASR) [4].

There is another exceptional reason exist why the formant representation of the speech signal appeals the speech community. It is because of their relation to spectral maxima [4]. Therefore a speech signal represented by formant parameters could be robust against additive noise because the lower energy regions of the spectrum could be masked by the noise energy, but the format regions might stay up above the noise level even if the average signal to noise ratio measure in dB becomes zero or negative [4].

The potential interest of formant data has led to numerous signal processing techniques for formant estimation over the past two decades [2, 3]. The dominant method of formant frequency estimation is based on modeling the speech signal as if it is generated by a particular kind of source and filter. This kind of analysis is called source-filter separation [6].

It is difficult to separate to the source and filter because of the spectral shape of the vocal tract excitation strongly influences the observed spectral envelope, such that it can not be guaranteed that all vocal tract resonance will cause maxima in the observed spectral envelope, nor that all the maxima in the observed spectral envelope are caused by the resonance of the vocal tract [6].

There is also very little published data on acoustic measurements of large speech corpora. Despite numerous attempts to build accurate and reliable automatic formant extractors, there are still no tools available that can automatically extract the “true” formants from the speech in the very large corpora that have become the standard in developing speech technology systems [4:4]. Therefore, it is fair enough to say that accurate formant estimation is a notoriously difficult task for which many works must be committed in order to develop an accurate formant tracking algorithm.

However the objective of the research reported in this paper is not to develop an accurate formant estimating technique but to examine the performance of three widely used fully automatic formant tracking algorithms (classical LPC method, the burg algorithm implemented in popular and widely used software PRAAT, and Auto Regressive method) which are applied to GRID corpus [9]. The questions here we try to answer whether it is better to apply a single method or multiple methods for robust formant estimation. Reliability of each method is determined by comparison with formant frequencies

Manuscript received August 12, 2010. This work was supported by the the EC-funded Project Marie Curie Research Training Network MRTN-CT-2006-035561 S2S (“Sound to Sense”, www.sound2sense.eu).

A. Kabir, is with the Department of Telecommunications, Technical University of Cluj-Napoca, 400027 Cluj-Napoca, Romania. (phone: +40-264-401807; e-mail: ahsanul.kabir@com.utcluj.ro).

J. Barker, is with the Department of Computer Science, Sheffield S1 4DP, United Kingdom. (e-mail: j.barker@dcs.shef.ac.uk).

M. Giurgiu, is with the Department of Telecommunications, Technical University of Cluj-Napoca, 400027 Cluj-Napoca, Romania. (e-mail: mircea.giurgiu@com.utcluj.ro).

available in [7] and [8]. We demonstrated by experimental results that no method is reliable enough for robust formant extraction.

## II. SPEECH MATERIALS

Speech data has been taken from the GRID corpus which is a large multi talker audio visual sentence corpus to support joint computational-behavioral studies in speech perception and automatic speech recognition [5]. It contains a total of 34,000 sentences of high quality audio and video (facial) recordings, 1000 sentences spoken by each of 34 speakers (18 male speakers, 16 female speakers). All speak British English as their first language. Though GRID is not suitable for large vocabulary systems, it provides is large enough to meet the training requirements of ASR systems.

## III. THE BACKGROUND OF METHODS

The vocal tract is a tube with time varying shape, and it has resonance frequencies like any other tubes. These resonances are called formants. Formant estimation usually involve in two different stages. In the first stage, speech signal is analyzed and formant candidates are obtained. In the second stage, most likely formant candidates are chosen by applying restraints. First three vowel formant frequencies (F1, F2, and F3) are obtained at the mid-point of the time interval corresponding to each vowel instance by the following methods.

### A. *Linear Prediction Coding (LPC)*

Liner prediction analysis based method is more common in formant estimation because the resolution can be set by the order of prediction, provides fine estimates of the spectrum (especially at the peaks which belong to formants), and able to produce acceptable spectrum estimates even for short speech segments. Transfer function of the LPC is given by,

$$H(z)=\frac{1}{1-\sum_{i=1}^{p} a_i z^{-i}} \qquad (1)$$

We applied hamming window at the mid-point of the time interval corresponding to each vowel. Then we computed linear prediction coefficients, $a_i$ of windowed speech and obtained the LPC polynomial (denominator of the transfer function). The roots of the LPC polynomial represent the poles of the vocal tract system and formants candidates are associated to the respective poles of the vocal tract system.

### B. *PRAAT (Burg Algorithm)*

The Burg algorithm implemented in PRAAT for formant estimation also works via LPC. First the sound is resampled to a sampling frequency of twice the value of maximum formant. Then a pre-emphasis is applied and finally, PRAAT applies a Gaussian like window to compute linear prediction coefficients through the Burg algorithm.

This algorithm can initially find formants at very low or high frequencies. In order to identify the F1 and F2, all formants below 50 Hz and all formants above maximum formants minus 50 Hz, are removed.

### C. *Auto Regressive Method (AR)*

AR is the simplest model for the vocal tract, consisting of linked cylindrical tubes producing an-all pole vocal tract transfer function. AR models try to predict the output of a system by the previous output and by the previous input. Notably, AR model is based on frequency domain analysis and needs to be windowed. In this case, we used hamming window. The transfer function of the AR model is given by,

$$H(z)=\frac{b_1+b_2 z^{-1}+........+b_{m+1} z^{-m}}{a_1+a_2 z^{-1}+........+a_{m+1} z^{-m}} \qquad (2)$$

Where *a*, and *b* are the filter coefficients and *m* is the order of the filter.

## IV. RELIABITLITY EXPERIMENT

The reliability of each method is evaluated by comparison with the formant frequencies published in [7] and [8]. We labeled these values as the expected formant frequencies ($\mu_{\text{expected}}$ in Figure 1). Firstly, for each method and for each vowel independently, average formant frequencies for F1, F2, and F3 across all instances are computed for each speaker as well as for the entire population. Secondly, for each vowel independently, difference between the formant frequencies generated by each method and expected formant frequencies are computed across all instances.

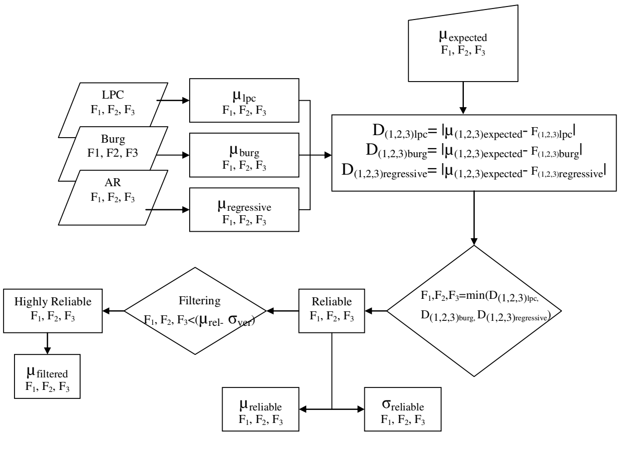
μexpected F1, F2, F3
LPC F1, F2, F3
Burg F1, F2, F3
AR F1, F2, F3
μlpc F1, F2, F3
μburg F1, F2, F3
μregressive F1, F2, F3
D(1,2,3)lpc= |μ(1,2,3)expected- F(1,2,3)lpc|
D(1,2,3)burg= |μ(1,2,3)expected- F(1,2,3)burg|
D(1,2,3)regressive= |μ(1,2,3)expected- F(1,2,3)regressive|
F1,F2,F3=min(D(1,2,3)lpc, D(1,2,3)burg, D(1,2,3)regressive)
Reliable F1, F2, F3
Filtering F1, F2, F3<(μrel- σver)
Highly Reliable F1, F2, F3
μfiltered F1, F2, F3
μreliable F1, F2, F3
σreliable F1, F2, F3

Fig 1. schematic diagram of the reliability experiment

Thirdly, the method which provides the minimum difference among the methods is picked up as the reliable formant (RF) estimation for the specific vowel. Therefore, it is possible that F1 for a vowel is labeled as reliable computed by LPC where F2 for the same vowel is labeled as reliable computed by AR. Fourthly, for each vowel independently, average formant frequencies for F1, F2, and F3 and their standard deviation across all instances are computed for each speaker as well as for the entire population after the selection of reliable formants. Fifthly, for each vowel and for each speaker independently, any formant lies beyond the standard deviation (for the same vowel and same speaker) from the expected values are discarded. The remaining formants which have been kept after this process, labeled as highly reliable formant (HRF). And finally, for each vowel independently, average formant frequencies for F1, F2, and F3 and their standard deviation are computed again for each speaker as well as for the entire population. Figure 1 shows the schematic diagram of the reliability experiment for robust formant estimation.

## V. Experimental Results

### A. *Measurements*

The average values for the first three formants estimated by LPC, Burg Algorithm, and Auto Regressive (AR) method for male and female speakers are shown in Table I, Table II and Table III. Average F1 for male speakers are greater than the average F1 for female speakers estimated by LPC and AR. Therefore, measurements given by these two methods are not accurate and robust enough for large speech corpora. Burg algorithm shows the opposite characteristics where F1 for male speakers are less than the F1 for female speakers.

TABLE I. Average F1, F2, and F3 by LPC

| Vowel | Male | | | Female | | |
|---|---|---|---|---|---|---|
| | *F1* | *F2* | *F3* | *F1* | *F2* | *F3* |
| aa | 603 | 1122 | 2205 | 543 | 1058 | 1687 |
| ae | 497 | 1358 | 2264 | 452 | 1039 | 2064 |
| ah | 482 | 1161 | 2198 | 456 | 1034 | 1966 |
| ao | 433 | 865 | 2060 | 472 | 1000 | 2023 |
| ax | 379 | 1439 | 2294 | 377 | 1016 | 2254 |
| eh | 513 | 1414 | 2253 | 497 | 981 | 2081 |
| ey | 401 | 1285 | 2237 | 446 | 1046 | 2361 |
| ih | 373 | 1312 | 2256 | 417 | 1054 | 2241 |
| iy | 305 | 1165 | 2284 | 361 | 1160 | 2457 |
| ow | 491 | 1346 | 2193 | 425 | 966 | 1919 |
| uw | 346 | 1364 | 2225 | 380 | 1042 | 2157 |

TABLE II. Average F1, F2, and F3 by Burg Algorithm (PRAAT)

| Vowel | Male | | | Female | | |
|---|---|---|---|---|---|---|
| | *F1* | *F2* | *F3* | *F1* | *F2* | *F3* |
| aa | 650 | 1129 | 2617 | 776 | 1266 | 2897 |
| ae | 543 | 1608 | 2579 | 599 | 1970 | 2906 |
| ah | 541 | 1336 | 2657 | 661 | 1597 | 2868 |
| ao | 470 | 852 | 2608 | 511 | 1031 | 2905 |
| ax | 422 | 1883 | 2506 | 480 | 2251 | 2868 |
| eh | 570 | 1729 | 2611 | 723 | 1987 | 2812 |
| ey | 434 | 1997 | 2697 | 522 | 2342 | 2981 |
| ih | 407 | 1853 | 2590 | 475 | 2221 | 2955 |
| iy | 339 | 2161 | 2788 | 403 | 2634 | 3150 |
| ow | 497 | 1360 | 2260 | 580 | 1693 | 2538 |
| uw | 363 | 1664 | 2415 | 429 | 2049 | 2815 |

### B. *Reliability Measurements*

Obviously formants estimated by these three methods have a high difference among them and very difficult to pick up the reliable estimation because of the robustness of the corpus. A reliability experiment, described in section IV has been carried out in comparison to standard formant values presented in Deterding [7] and Hawkins [8]. The average value for the first three reliable formants and highly reliable formants for male and female speakers are shown in Table IV and Table V. It can be noted that measurement of the third formant and for female speakers are not available in Hawkins [8].

TABLE III. Average F1, F2, and F3 by Auto Regressive Model

| Vowel | Male | | | Female | | |
|---|---|---|---|---|---|---|
| | *F1* | *F2* | *F3* | *F1* | *F2* | *F3* |
| aa | 565 | 1047 | 2161 | 434 | 937 | 1416 |
| ae | 492 | 1437 | 2357 | 432 | 1056 | 2179 |
| ah | 457 | 1188 | 2216 | 396 | 924 | 1823 |
| ao | 439 | 826 | 2437 | 386 | 709 | 1540 |
| ax | 394 | 1722 | 2440 | 320 | 927 | 2379 |
| eh | 505 | 1459 | 2331 | 463 | 968 | 2112 |
| ey | 403 | 1699 | 2522 | 386 | 868 | 2460 |
| ih | 385 | 1647 | 2471 | 360 | 955 | 2373 |
| iy | 317 | 1785 | 2659 | 283 | 796 | 2690 |
| ow | 473 | 1289 | 2150 | 390 | 852 | 1911 |
| uw | 354 | 1551 | 2346 | 317 | 863 | 2284 |

TABLE IV. Average F1, F2, and F3 Labeled as Reliable Formants by Comparison to [7] & [8]

| Vowel | Male | | | | | Female | | |
|---|---|---|---|---|---|---|---|---|
| | *F1* | | *F2* | | *F3* | *F1* | *F2* | *F3* |
| | [7] | [8] | [7] | [8] | | | | |
| aa | 643 | 631 | 1136 | 1103 | 2487 | 777 | 1263 | 2870 |
| ae | 548 | 555 | 1591 | 1579 | 2555 | 594 | 1898 | 2875 |
| ah | 551 | 552 | 1220 | 1191 | 2553 | 664 | 1279 | 2804 |
| ao | 449 | 438 | 841 | 789 | 2603 | 424 | 956 | 2856 |
| ax | 430 | 423 | 1385 | 1356 | 2477 | 478 | 1108 | 2733 |
| eh | 548 | 552 | 1650 | 1625 | 2581 | 652 | 1901 | 2797 |
| ey | 440 | 444 | 1835 | 1986 | 2629 | 523 | 2297 | 2933 |
| ih | 393 | 402 | 1823 | 1881 | 2578 | 412 | 2216 | 2922 |
| iy | 308 | 306 | 2177 | 2184 | 2764 | 321 | 2640 | 3100 |
| ow | 473 | 481 | 1333 | 1347 | 2291 | 413 | 1516 | 2546 |
| uw | 351 | 343 | 1448 | 1617 | 2395 | 356 | 1411 | 2726 |

TABLE V. Average F1, F2, and F3 Labeled as Highly Reliable Formants by Comparison to [7] & [8]

| Vowel | Male | | | | | Female | | |
|---|---|---|---|---|---|---|---|---|
| | *F1* | | *F2* | | *F3* | *F1* | *F2* | *F3* |
| | [7] | [8] | [7] | [8] | | | | |
| aa | 642 | 605 | 1151 | 1041 | 2470 | 658 | 984 | 2181 |
| ae | 670 | 884 | 1552 | 1526 | 2559 | 659 | 1239 | 1938 |
| ah | 616 | 627 | 1144 | 1141 | 2565 | 762 | 1260 | 2584 |
| ao | 422 | 390 | 830 | 659 | 2647 | 391 | 917 | 2910 |
| ax | 535 | 469 | 1133 | 840 | 2530 | 181 | 270 | 702 |
| eh | 497 | 512 | 1500 | 1438 | 2549 | 624 | 1769 | 2726 |
| ey | 482 | 577 | 1723 | 1909 | 2609 | 602 | 1874 | 2543 |
| ih | 375 | 394 | 1772 | 2097 | 2546 | 395 | 2196 | 2929 |
| iy | 288 | 284 | 2245 | 2295 | 2803 | 305 | 2708 | 3135 |
| ow | 389 | 429 | 1190 | 1298 | 2305 | 408 | 1406 | 2496 |
| uw | 326 | 304 | 1279 | 1628 | 2364 | 341 | 1358 | 2728 |

### C. *Evaluation of Formant Estimation Algorithms*

Figure 2 shows the contributing percentage of methods in order to select reliable formants in comparison to Deterding [7]. Clearly, Burg algorithm is working better for both male and female speakers in order to select reliable formants. Burg algorithm is followed by LPC for selecting F1 and F2, but followed by AR for selecting F3. Notably, Burg algorithm is working much better for female speakers rather than male speakers in order to select reliable formants. It is providing 65% of reliable formants for female speakers where only 43% for the male speakers.

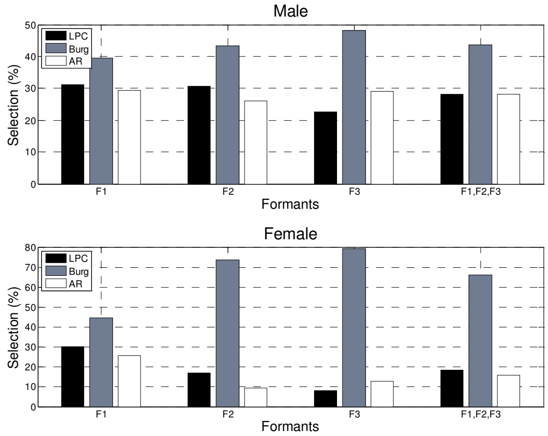

Fig 2. reliable formants selection from three methods in comparison to Deterding [7]

Figure 3 shows the contribution of different methods in order to select highly reliable formants in comparison to Deterding [7]. Overall performance (F1, F2, F3 in Figure 2 and Figure 3 where F1, F2 in Figure 4) of Burg algorithm is better for both male and female speakers but did not outperform other two methods in each case. Burg algorithm is working reliably well over LPC and AR by supplying the highest number of highly reliable F2 and F3. But the scenario is somehow different for F1 as LPC provides the highest contribution which is followed by Burg Algorithm for the male speakers and by AR for the female speakers.

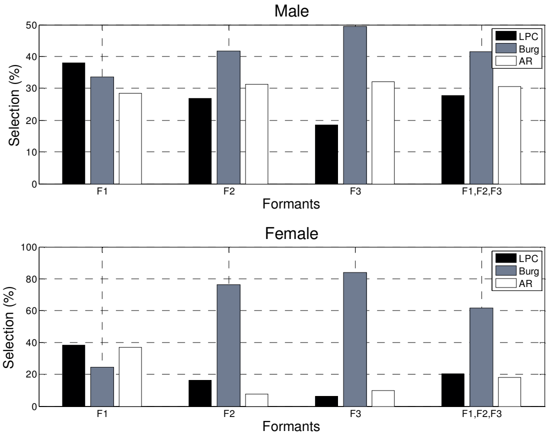

Fig 3. highly reliable formants selection from three methods in comparison to Deterding [7]

Figure 4 shows the contribution of different methods in order to select reliable and highly reliable formants in comparison to Hawkins [8]. For reliable formants selection, Burg algorithm outperformed other methods in each case. It is followed by LPC in selecting F1 where followed by AR in selecting F2. For highly reliable formant selection, overall performance of Burg algorithm is better but did not outperformed other two methods. LPC is followed by Burg algorithm in selecting F1 where Burg algorithm is followed by AR in selecting F2.

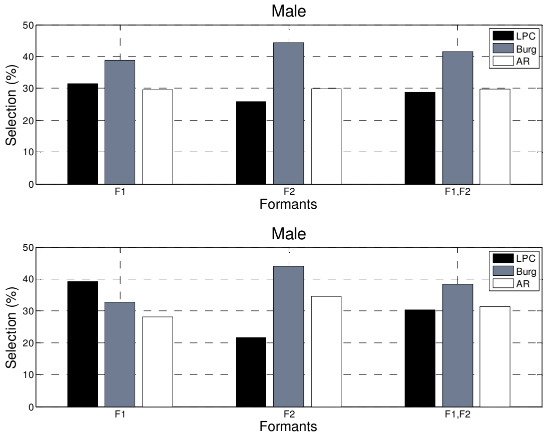

Fig 4. reliable and highly reliable formants selection from three methods in comparison to Hawkins [8]

## VI. Conclusion

In this paper, we presented formant frequencies of the eleven monophthongs of British pronunciation of English extracted from a large speech corpus by three different methods and classified according to gender. It can be noted that no method is robust and accurate enough for reliable formant estimation. A single method might work reliably for a small speech corpus but multiple methods should be considered for large speech corpus and especially for connected speech. This research could serve as the reference for future research and formant based robust vocal tract length normalization.

## References

[1] C.Kim, K.Seo and W.Sung. "A Robust Formant Extraction Algorithm Combining Spectral Peak Picking and Root Polishing," *EURASIP Journal on Applied Signal Processing,* vol. 2006, pp. 1-16, 2006.

[2] Z.B.Messaoud, D.Gargouri, S.Zribi and A.B.Hamida. "Formant Tracking Linear Prediction Model using HMMs for Noisy Speech Processing," *International Journal of Signal Porcessing,* vol. 5, pp. 291-296, 2009.

[3] Q.Yan, S.Vaseghi, E.Zavarehei B.Milner, J.Darch, P.White and I.Andrianakis. "Formant Tracking Linear Prediction Model using HMMs and Kalman Filters for Noisy Speech Processing," *Computer Speech and Language,* vol. 21, pp. 543-561, Jul. 2007.

[4] F.D.Wet, K.Weber, L.Boves, B.Cranen, S.Bengio and H.Bourlard, "Evaluation of Formant-Like Features for Automatic Speech Recognition." *Journal of the Acoustical Society of America*, vol. 116, pp. 1781-1791, 2004.

[5] M. Cooke, J. Barker, S. Cunningham and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition." *Journal of the Acoustical Society of America*, vol. 120, 2006.

[6] "Introduction to Computer Programming, Speech Signal Analysis." Internet: http://www.phon.ucl.ac.uk/courses/spsci/matlab/lect10.html, Dec. 10, 1990 [Jul. 21, 2010].

[7] D.Deterding. "The Formants of Monophthong Vowels in Standard Southern British English Pronunciation," Journal of the International Phonetic Association, vol. 27, pp. 47–55, 1997.

[8] S.Hawkins and J.Midgley. "Formant frequencies of RP monophthongs in four age groups of speakers," Journal of the International Phonetic Association, vol. 35, pp. 183–199, 2005.

[9] J.P.Burg. "Maximum entropy spectral analysis," in *Proc. Meet. Society of Exploration Geophysicists*, 1967, pp. 34-41.