STATISTICAL ANALYSIS OF AMPLITUDE MODULATION IN SPEECH SIGNALS USING AN AM–FM MODEL

Pirros Tsiakoulis

School of Electrical and Computer Eng.,
National Technical University of Athens,
Athens 15773, Greece
ptsiak@ilsp.gr

Alexandros Potamianos

Dpt of Electronics and Computer Eng.,
Technical University of Crete,
Chania 73100, Greece
potam@telecom.tuc.gr

ABSTRACT

Several studies have been dedicated to the analysis and modeling of AM–FM modulations in speech and different algorithms have been proposed for the exploitation of modulations in speech applications. This paper details a statistical analysis of amplitude modulations using a multi-band AM-FM analysis framework. The aim of this study is to analyze the phonetic- and speaker-dependency of modulations in the amplitude envelope of speech resonances. The analysis focuses on the dependence of such modulations on acoustic features such as, fundamental frequency, formant proximity, phone identity, as well as, speaker identity and contextual features. The results show that the amplitude modulation index of a speech resonance is mainly a function of the speaker's average fundamental frequency, the phone identity, and the proximity between neighboring formant resonances. The results are especially relevant for speech and speaker recognition application employing modulation features.

Index Terms— AM–FM, modulations, speech analysis

1. INTRODUCTION

It is well-known that speech production exhibits various nonlinear and time-varying phenomena, due to the nature of the underlying physics. The AM–FM model is a nonlinear model that describes a speech resonance as a signal with a combined amplitude modulation (AM) and frequency modulation (FM) structure [1]

$$r(t) = a(t)\cos(2\pi[f_c t + \int_0^t q(\tau)d\tau] + \theta) \qquad (1)$$

where $f_c \stackrel{\triangle}{=} F$ is the "center value" of the formant frequency, $q(t)$ is the frequency modulating signal, and $a(t)$ is the time-varying amplitude. The instantaneous formant frequency signal is defined as $f(t) = f_c + q(t)$. The speech signal $s(t)$ is modeled as the sum $s(t) = \sum_{k=1}^{K} r_k(t)$ of K such AM–FM signals, one for each formant. Based on the AM–FM model, and utilizing the Teager-Kaiser energy operators, a number of efficient algorithms have been introduced for energy separation and demodulation of the speech signal in its AM–FM components [1, 2, 3].

The AM–FM model has been successfully applied in various areas of signal processing including speech, music and image processing. Specifically in speech processing, the AM–FM model has been used for speech analysis and modeling [4, 5, 6], speech synthesis [4], speech recognition [7, 6] and speaker identification [8, 9]. Significant improvement in speech recognition accuracy has been shown in [7], where features measuring amplitude and frequency modulation percentage are included in the speech recognition front-end. In [9], similar features have been applied to speaker identification. In [8], the ratio of amplitude between the primary and secondary pulse within a pitch period (a measure of amplitude modulation) is also used as a feature for speaker identification.

Despite the considerable amount of work on the AM–FM model, and its successful utilization in different areas of speech processing, there are still various aspects related to the presence of AM–FM modulations in speech that need to be further investigated. This paper aims towards a better understanding of why and when amplitude modulations appear in speech; the end goal is to improve the use of modulations in speech applications based on the AM–FM model. We perform a statistical analysis of amplitude modulations on bandpassed speech signals along the formant tracks. The analysis focuses on the dependence of such modulations on linguistic and/or non-linguistic properties of speech, such as fundamental frequency, formant proximity, etc. Preliminary investigations have shown both speaker and phone dependency of amplitude modulation patterns [4]. Next the AM-FM demodulation analysis method is described.

2. AM–FM DEMODULATION ANALYSIS

The AM-FM multiband demodulation approach outlined in [10] is followed in this paper. Specifically, the formant tracks are estimated using the multiband demodulation formant tracking algorithm (MDA). The speech signal is then decomposed into resonances using Gabor filtering along estimated formant tracks (typical effective RMS Gabor filter bandwidth is 400 Hz). Each speech resonance is assumed an AM–FM signal (Eq. 1).

The amplitude envelope $|a(t)|$ and instantaneous frequency $f(t)$ signals can be obtained by applying the energy separation algorithm on the speech resonance signal $r(t)$. The *energy separation algorithm* (ESA) [1] is based on the nonlinear differential Teager–Kaiser energy operator. The energy operator tracks the energy of the source producing an oscillation signal $r(t)$ and is defined as $\Psi[r(t)] = [\dot{r}(t)]^2 - r(t)\ddot{r}(t)$ where $\dot{r} = dr/dt$. The frequency and amplitude estimates are [1]

$$\frac{1}{2\pi}\sqrt{\frac{\Psi[\dot{r}(t)]}{\Psi[r(t)]}} \approx f(t)\ , \qquad \frac{\Psi[r(t)]}{\sqrt{\Psi[\dot{r}(t)]}} \approx |a(t)|. \qquad (2)$$

Similar equations and algorithms exist in discrete time (DESA).[1]

[1] Alternatively, the Hilbert transform demodulation (HTD) algorithm can be used to estimate $|a(t)|$, $f(t)$ (as the modulus and the phase derivative of the analytic signal).

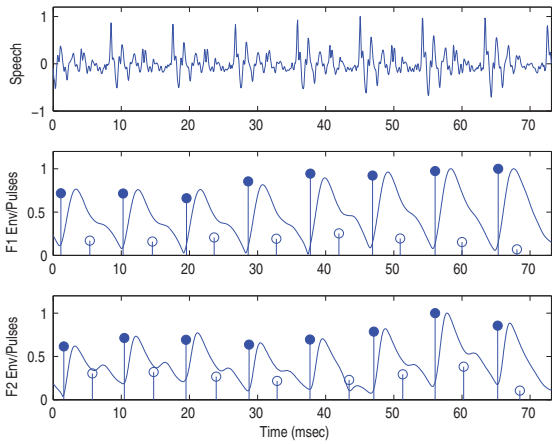

Fig. 1. Speech signal (phoneme /ae/, male speaker) and the amplitude envelopes for the first two formants and superimposed the corresponding primary and secondary excitation pulses.

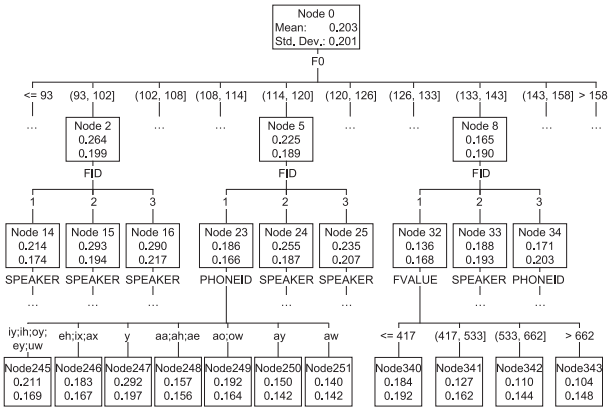

Fig. 2. CART of amplitude modulation index (AMI) estimated for all male vowels and diphthongs. Each node shows the corresponding mean value and standard deviation of AMI, while the each edge shows the corresponding splitting criterion. The tree has been pruned for better readability.

The amplitude envelope signals $|a(t)|$ have a specific structure which is modeled using a multipulse model and an analysis-by-synthesis loop [4]. The amplitude envelope signals $a(n)$ are modeled as

$$a(n) = u(n) * g(n) * h(n), \quad u(n) = \sum_{k=1}^{K} b_k \delta(n - n_k) \qquad (3)$$

where the impulse sequence $u(n)$ is the excitation signal, $g(n)$ is the impulse response of a critically damped second-order system, $h(n)$ is the baseband impulse response of the filter used for extracting the corresponding resonance signal and $\delta(n)$ the Kronecker delta function. The pulse positions n_k are computed from an analysis-by-synthesis loop, while the amplitudes b_k have a closed form solution so that the mean square modeling error is minimized.

The mathematical model outlined above can efficiently capture amplitude modulation patterns. In this study, we apply a two-pulse per pitch period analysis, since we focus on a quantitative statistical analysis of amplitude modulation, and not on detailed amplitude modulation pattern modeling. The pulse with the maximum amplitude a_p within a pitch period, is characterized as *primary*, while the stronger pulse till the next primary pulse a_s, if any, is characterized as *secondary*. The fundamental frequency was estimated using the algorithm described in [4]. The estimate of the amplitude modulation index (AMI) is defined as the ratio of the secondary to the primary pulse, i.e., AMI $= a_s/a_p$. The AMI is computed for each pitch period and statistics are computed.

In Fig. 1, the amplitude envelope signals and the corresponding primary and secondary pulses are shown. The excitation pulses were computed as described above for the speech segment from an instance of the phoneme /ae/ (shown at the top).

3. STATISTICAL ANALYSIS

The TIMIT database was analyzed using the procedure described in the previous section and data was collected for the statistical analysis (both train and test sets were included). For each sentence in the database, the primary and secondary pulse amplitudes were computed for each formant resonance amplitude envelope signal (F1, F2 and F3) for instances of vowels and diphthongs. The ratio of the secondary to primary excitation pulse amplitude was computed as a rough estimate of the amplitude modulation index (AMI).

For the evaluation of the AMI estimation process we manually inspected a small set of analysis examples, randomly chosen from the analysis data set. As far as the location of the estimated pulses concerned, the resulting success rates (within few hand-labeled samples) were: 96%, 96%, 90% for the male speakers F1, F2, F3, respectively, and 93%, 94%, 89% for female speakers.

Next, a CART analysis [11] was performed using the AMI estimate as the dependent variable, while various linguistic and nonlinguistic features were used as independent variables, in order to find possible dependencies of AMI estimate on such features. Specifically, the independent variables included the following: speaker's identity, gender, dialect, phone identity, previous and next phone identities, previous and next phonetic class, the relative position within phone, fundamental frequency (F0), formant identity and value, and distances from previous and next formant.

CART analysis reported here is on all TIMIT vowel and diphthong instances and is broken down by gender. In Fig. 2, the most important independent variables are shown as computed from the CART algorithm from all male vowel and diphthong regions (tree depth is set to three). The most important variable is F0 which is used as the first split criterion. Note that there are 10 nodes in the first level with decreasing AMI as the corresponding F0 increases. In the second and third level, the chosen independent variables are the following: formant identity, speaker identity, formant value, difference of the formant value from the previous formant value, the phoneme identity, the relative position within the phone and the phonetic context.

The CART analysis clearly shows a correlation between the AMI estimate and F0, as well as, with phone identity, speaker identity and formant values. These relationships are further investigated below.

3.1. Fundamental Frequency

In Fig. 3, the mean AMI estimated for the first three formants is plotted versus F0. There is a clear decreasing trend of the AMI estimate

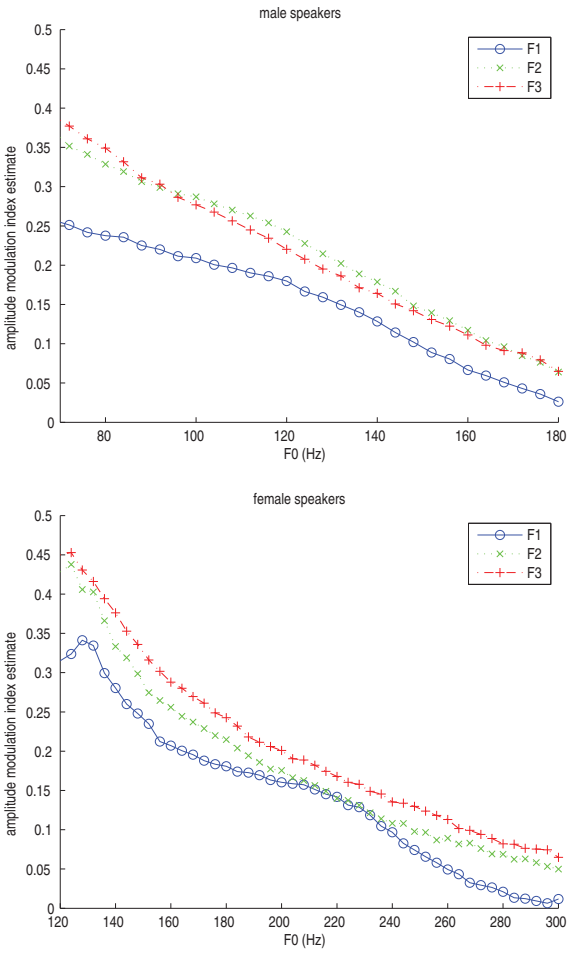

Fig. 3. The amplitude modulation index (AMI) estimated for each formant resonance amplitude envelope signal (F1, F2, F3) versus fundamental frequency (F0).

for all formants, both for male and female speakers. Moreover the decreasing pattern is very similar across formants and across gender for the F0 range of interest.

The results show that amplitude modulation phenomena are clearly more prominent in low F0 conditions. This could be due to the fact that lower F0 allows for more time to achieve complete closure of the vocal cords, leading to a more prominent secondary excitation(s) within the pitch period. There is also a significant difference in AMI for F1 (compared to F2 and F3). This trend is also evident in Figs. 4 and 5.

3.2. Phone Identity

The relationship between the AMI estimate and the phone identity is investigated in Fig. 4 for vowels and diphthongs. The average AMI is shown for the first three formants of each phone. ANOVA analysis shows that the AMI differences across different phones are statistically significant (significance level less than 0.01).

The main conclusion drawn from inspecting the phone identity's plots is that for high vowels (vowels produced with high tongue position) the AMI is higher than for the rest. This behavior is especially pronounced for the first formant and is consistent for both genders. A

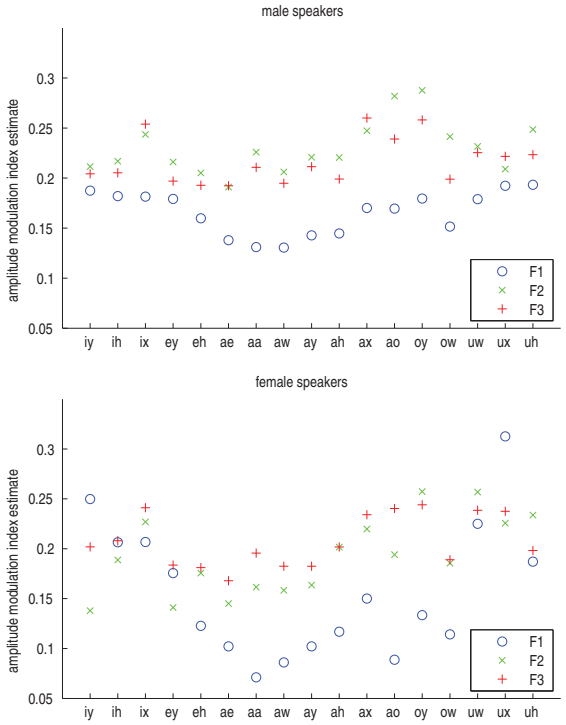

Fig. 4. The average amplitude modulation index (AMI) estimate for the first three formants (F1, F2 and F3) shown for male and female speakers as a function of phone identity.

possible explanation for this phenomenon is that the high tongue position creates an additional constriction along the vocal tract, which in turn enhances non-linear airflow phenomena such as turbulence and vorticity. Such phenomena could affect the modulation index, especially for F1, which is formed behind this constriction[2]. Modulations in F2 and F3 have a more erratic behavior compared to F1, especially for male speakers. This indicates that modulations for the higher formants are affected by other factors in addition to tongue position.

Diphthongs, being a compound phonemic class, have a mixed behavior. The behavior of diphthongs is also hard to analyze from a non-linear production point of view due to the transient nature of these sounds. A more detailed look at the steady state and transient parts of those sounds is needed to draw concrete conclusions. All in all, phone identity is a strong correlate of modulation phenomena.

3.3. Formant Proximity

In Fig. 5, the average AMI is shown for F1-F3 as a function of formant proximity (difference of adjacent formants' values) for both male and female speakers.

All plots show an increased modulation index for formants that are far from each-other, especially for F1 modulations. This indicates that modulations are affected by the spectral valley between the two formants. This could be due to spectral zeros in vocal tract trans-

[2]It is interesting to note that this plot indicates that one can "guess" a vowel's identity simply from F1 value and F1 modulation index. This could offer a possible explanation to the puzzle as to why natural speech lowpassed below 1kHz retains some phonetic information, while LPC vocoded speech has no such information.

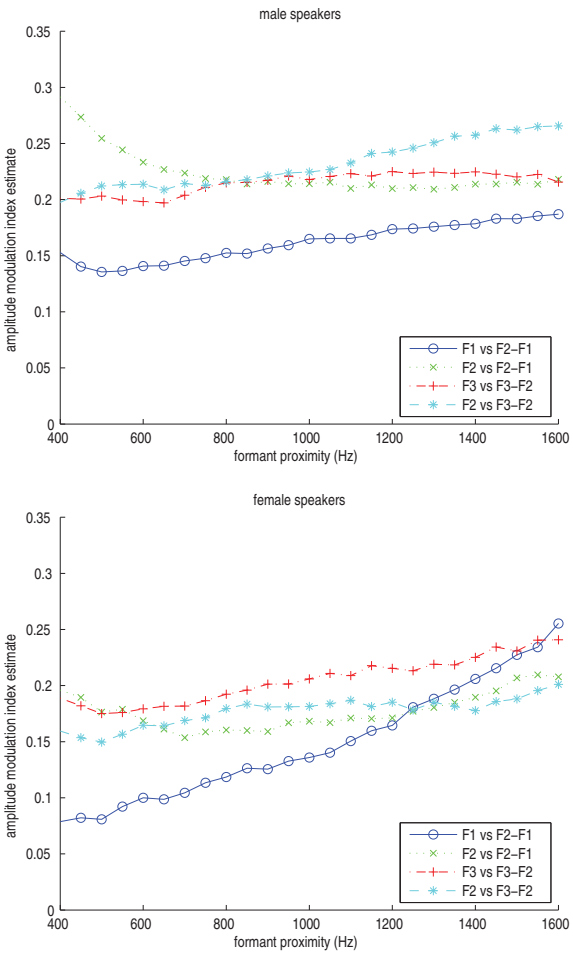

Fig. 5. The amplitude modulation index (AMI) estimated for F1-F3 versus formant proximity for male and female speakers.

fer function. The average AMI for F1, for both genders, increases as F2 moves further away from F1. Note that a large distance between F1 and F2 implies that the vowel is high-front, in which case, there is an increased F1 modulation index as it can be seen in Fig. 4.

The AMI for F2 and F3 shows a less steep increment in relation to formant proximity. This is because in these cases, there is also the effect from the other adjacent formant. A large distance, e.g., between F1 and F2, usually implies proximity between F2 and F3. Specifically, for F2 in relation to F1-F2 proximity, there is a significant rise bellow 600Hz. This could be due to nonlinear interaction between F1 and F2, as F2 closes on F1. All in all, formant proximity is a factor that affects AMI especially for F1 and F2; formant distance usually implies higher modulation index.

4. CONCLUSIONS

The analysis of the amplitude modulation index (AMI) for vowels and diphthongs shows that modulations are mainly a function of speaker identity, gender and phone identity. The three most important correlates are F0, phone identity (especially high vowels vs. the rest) and formant proximity (especially for F1 and F2). Gender is also an important correlate that manifests itself both through F0 and formant proximity analysis. Overall, higher modulations are evident for low F0, formants that are far from each-other and high tongue position. All these factors tend to enhance non-linear phenomena during speech production.

As far as, speech applications are concerned, it is clear from this analysis that modulations can help in both speech and speaker recognition, since both speaker characteristics and phone identity correlate well with amplitude modulation index. Features that measure amplitude and frequency modulation, e.g., FMP proposed in [7] have been used successfully for both applications [7, 6, 8, 9]. The analysis proposed in this paper can help us extract better features for speech applications, as well as, normalize such features to limit the effect of extraneous factors, e.g., effect of F0 (or in general speaker characteristics) in speech recognition applications.

This paper is a first step towards a more detailed analysis of modulation phenomena in speech production. More research is needed on the speech analysis, modeling and recognition fronts to better understand (i) the properties of modulations, (ii) the physics behind non-linear speech production, and (iii) the relevance of modulation for speech processing applications.

5. ACKNOWLEDGEMENTS

This research was co-financed partially by E.U.-European Social Fund (80%) and the Greek Ministry of Development-GSRT (20%) under Grant PENED-2003-ED866. The authors would like to thank Prof. Petros Maragos, Dr. Dimitrios Dimitriadis and Athanassios Katsamanis for many useful discussions.

6. REFERENCES

[1] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Signal Processing*, vol. 41, no. 10, pp. 3024–3051, October 1993.

[2] P. Maragos, T. F. Quatieri, and J. F. Kaiser, "Speech nonlinearities, modulations and energy operators," in *ICASSP-91*, Toronto, Canada, May 1991.

[3] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On amplitude and frequency demodulation using energy operators," *IEEE Trans. Signal Processing*, vol. 41, no. 4, pp. 1532–1550, April 1993.

[4] A. Potamianos and P. Maragos, "Speech analysis and synthesis using an AM–FM modulation model," *Speech Communication*, vol. 28, pp. 195–209, July 1999.

[5] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 5, pp. 569–586, September 1999.

[6] D. Dimitriadis and P. Maragos, "Continuous energy demodulation methods and application to speech analysis," *Speech Communication*, vol. 48, no. 7, pp. 819–837, July 2006.

[7] D. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM–FM features for speech recognition," *IEEE Signal Processing Letters*, vol. 12, no. 9, pp. 621–624, September 2005.

[8] C. R. Jankowski Jr., T. F. Quatieri, and D. A. Reynolds, "Measuring fine structure in speech: Application to speaker identification," in *ICASSP-95*, Detroit, USA, May 1995.

[9] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 6, pp. 1097–1111, August 2008.

[10] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *Journal of Acoustical Society of America*, vol. 99, pp. 3795–3806, June 1996.

[11] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, CRC Press, 1998.