



# On the Effect of Fundamental Frequency on Amplitude and Frequency Modulation Patterns in Speech Resonances

Pirros Tsiakoulis<sup>1</sup> and Alexandros Potamianos<sup>2</sup>

<sup>1</sup>School of Electrical and Computer Engineering, National Technical University of Athens, Greece

<sup>2</sup>Department of Electronics and Computer Engineering, Technical University of Crete, Greece

ptsiak@ilsp.gr, potam@telecom.tuc.gr

## Abstract

Amplitude modulation (AM) and frequency modulation (FM) in speech signals are believed to reflect various non-linear phenomena during the speech production process. In this paper, the amplitude and frequency modulation patterns are analyzed for the first three speech resonances in relation to the fundamental frequency (F0). The formant tracks are estimated, and the resonant signals are extracted and demodulated. The *Amplitude Modulation Index* (AMI) and *Frequency Modulation Index* (FMI) are computed, and examined in relation to the F0 value, as well as the relation between F0 and the first formant value (F1). Both AMI and FMI are significantly affected by pitch, with modulations being more frequently present in low F0 conditions. Evidence of non-linear interaction between the glottal source and the vocal tract is found in the dependence of the modulation patterns on the ratio of F1 over F0. AMI is amplified when pitch harmonics coincide with F1, while FMI shows complementary behavior.

**Index Terms:** speech analysis, AM–FM, modulation, fundamental frequency, pitch harmonics

## 1. Introduction

It is well-known that speech production exhibits various non-linear and time-varying phenomena, due to the nature of the underlying physics. Several studies report various experimental results on the vocal tract aeroacoustics, as well as numerical simulations, that provide strong evidence of such non-linear phenomena [1, 2]. Furthermore, various other studies have shown that there exists non-linear coupling between the glottal source and the vocal tract [3, 4, 5, 6].

The AM–FM speech model was proposed as a non-linear alternative, by modeling the speech signal as a sum of AM–FM components [7]. Modulation patterns were found to be speaker, phoneme and context depended [8], making the use of AM–FM modeling suitable for a variety of speech applications. Significant improvement in speech recognition accuracy has been shown in [9], when features measuring AM and FM percentage extend the standard acoustic feature vector. In [10] we investigated short-time estimates of instantaneous frequency and bandwidth as stand-alone feature sets for speech recognition. Modulation based features have also been proposed for speaker identification [11, 12]. 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

This work was supported in part by the E.U.-European Social Fund (80%) and by the Greek Ministry of Development-GSRT (20%) under Grant PENED-2003-ED866.

that need to be further investigated. In a recent study, we have detailed a statistical analysis of amplitude modulation index of speech resonant signals [13]. In this paper, we extend our analysis considering both amplitude and frequency modulation metrics. Conclusions drawn from this analysis, are of high interest for speech applications, and especially for speech recognition.

## 2. AM–FM analysis framework

The speech analysis framework used in this work consists of an AM–FM model, and set of accompanying tools. The AM–FM model is a non-linear representation of the speech signal. The speech signal is modeled as a composition of signals that combine both amplitude and frequency modulation. A filterbank is used to decompose the speech signal into its resonant components. The extracted the resonant signals, are then demodulated into the instantaneous amplitude and instantaneous frequency signals utilizing the *Teager-Kaiser Energy Operator* – (TEO), and the *Energy Separation Algorithm* – (ESA) [7, 14].

### 2.1. The AM–FM Model

The AM–FM model used in this work describes a speech resonance as a signal with a combined amplitude modulation (AM) and frequency modulation (FM) structure [7, 14]

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

where  $f_c$  is the “center value” frequency,  $q(t)$  is the frequency modulating signal,  $f(t) = f_c + q(t)$  the is instantaneous frequency signal, and  $a(t)$  is the time-varying amplitude signal. The speech signal  $s(t)$  is modeled as the sum of  $K$  such AM–FM resonant signals  $s(t) = \sum_{k=1}^K r_k(t)$

### 2.2. Demodulation

The demodulation of an AM–FM signal can be efficiently achieved employing ESA. The ESA demodulation is based on the TEO, which is defined via the first and second order derivatives of the signal  $x(t)$  as follows

$$\Psi[x(t)] = [\dot{x}(t)]^2 - x(t)\ddot{x}(t) \quad (2)$$

The ESA<sup>1</sup> instantaneous amplitude  $a(t)$  and frequency  $f(t)$  components are defined as [14]

$$\frac{1}{2\pi} \sqrt{\frac{\Psi_c[\dot{x}(t)]}{\Psi_c[x(t)]}} \approx f(t), \quad \frac{\Psi_c[x(t)]}{\sqrt{\Psi_c[\dot{x}(t)]}} \approx a(t) \quad (3)$$

<sup>1</sup>Usually the discrete ESA algorithms (DESA) are used, which are based on similar equations and the discrete TEO [7].

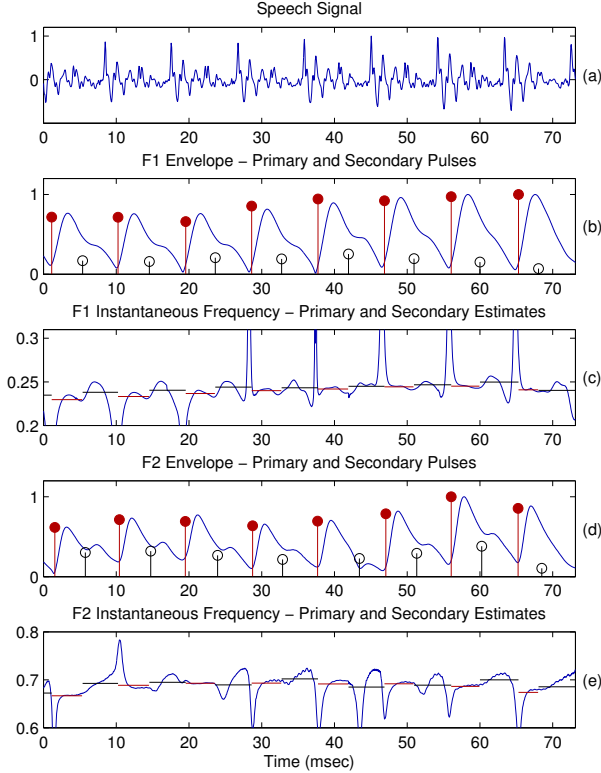


Figure 1: (a) Phoneme /ae/ from a male speaker, (b) the F1 instantaneous amplitude, and superimposed the primary and secondary pulses, (c) the F1 instantaneous frequency, and the  $F_w^p$  and  $F_w^s$  estimates, (d),(e) the corresponding F2 estimates.

### 2.3. Amplitude Modulation Index

The modulation patterns of instantaneous amplitude signals, as shown in Fig 1, have a specific structure, that can be exploited to estimate  $AMI$ . This structure of amplitude envelope signals is modeled using a multi-pulse model [8, 13]. The amplitude envelope signals  $a(n)$  are modeled as

$$a(n) = u(n) * g(n) * h(n), \quad u(n) = \sum_{k=1}^K b_k \delta(n - n_k) \quad (4)$$

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  $\delta(n)$  is 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. For the purpose of  $AMI$  estimation, the analysis is performed with two pulses per pitch period. The pulse with the maximum amplitude  $a_p$  within a pitch period, is characterized as *primary*, while the next stronger pulse, if any, is characterized as *secondary*. The amplitude modulation index is defined as the ratio of the secondary to the primary pulse

$$AMI = a_s / a_p \quad (5)$$

In Fig. 1 (b),(d) the amplitude envelope signals and the corresponding primary and secondary pulses are shown for the first and second formant (F1 and F2). The excitation pulses were computed as described above for a male vowel segment.

### 2.4. Frequency Modulation Index

The *Frequency Modulation Index (FMI)* is an estimate of the degree of divergence of the instantaneous frequency  $f(t)$  from its formant frequency. As we can see in Fig. 1 (c),(e) the instantaneous frequency varies within a pitch period, excluding the local spikes which are algorithmic side-effects. In order to capture this variation, we exploit the two-pulse modeling of the instantaneous amplitude. The primary and secondary pulses are used to define the primary and secondary regions respectively. The primary pulse region is defined as the region between the primary and secondary pulse positions (from  $n_p$  to  $n_s$ ), while the secondary pulse region is the rest of the pitch period. The frequency estimation is performed by averaging on the two regions, while incorporating an amplitude weighting to eliminate the spikes in the raw instantaneous frequency signals, i.e.

$$F_w = \frac{\sum_{k=0}^N f(k) a^2(k)}{\sum_{k=0}^N a^2(k)} \quad (6)$$

The ratio of the absolute difference between the frequency estimates in the two regions over the estimate in the whole period is used as an estimate of frequency modulation index

$$FMI = \frac{|F_w^p - F_w^s|}{F_w} \quad (7)$$

where  $F_w^p$  and  $F_w^s$  are the frequency estimates in the primary and secondary regions respectively.

## 3. Results

The TIMIT database is analyzed using the methodology described in the previous section. The analysis is performed on both train and test sets, and data is collected for male and female speakers. For each sentence, the multiband demodulation formant tracking algorithm (MDA) is applied for the estimation of the tracks of the first three formants (F1, F2 and F3) [8]. Next, the estimated tracks are used to extract the corresponding speech resonant signals, using Gabor bandpass filters along each track. The statistics of  $AMI$  and  $FMI$  are collected over vowel and diphthong regions, and examined in relation to F0.

### 3.1. AMI and FMI vs F0

In Fig. 2 (a),(b) the mean  $AMI$  estimated for the first three formants is plotted versus F0. The horizontal axis represents the value of F0 in Hz, which is sampled using bins of length 4 Hz. The vertical axis shows the mean  $AMI$  estimated over all vowel and diphthong regions, per formant and gender. The mean  $FMI$  is shown in Fig. 2 (c),(d). There is a clear decreasing trend of  $AMI$ , as well as,  $FMI$  estimates for all formants and both genders. Moreover, the decreasing pattern is very similar across formants and across gender.

### 3.2. Occurrence of amplitude modulation patterns

Fig. 3 (a),(b) show the mean  $AMI$  as a function of F0, including only the cases where amplitude modulations exist. For the cases where no amplitude modulation patterns are identified, there is an absence of a secondary pulse. The plots are significantly different than the corresponding plots in Fig. 2 (a),(b). The decreasing trend is to a large degree reduced, or even reversed. This shows that the decreasing trend is due to an increased number of zero values for high F0. This is confirmed in Fig. 2 (c),(d), where the ratio of the number of instances without secondary pulse over the total instances is plotted vs F0.

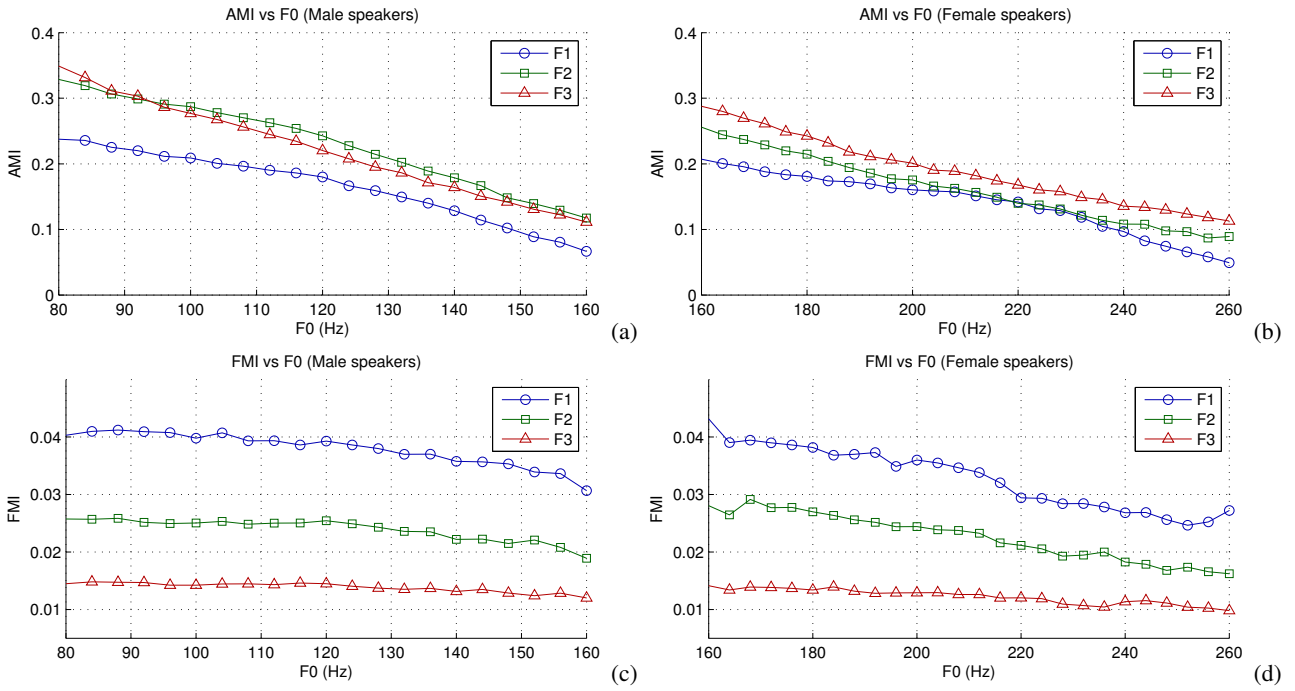


Figure 2: The *AMI* and *FMI* estimated for the F1, F2, and F3 resonant signals as a function of F0 value for male and female speakers.

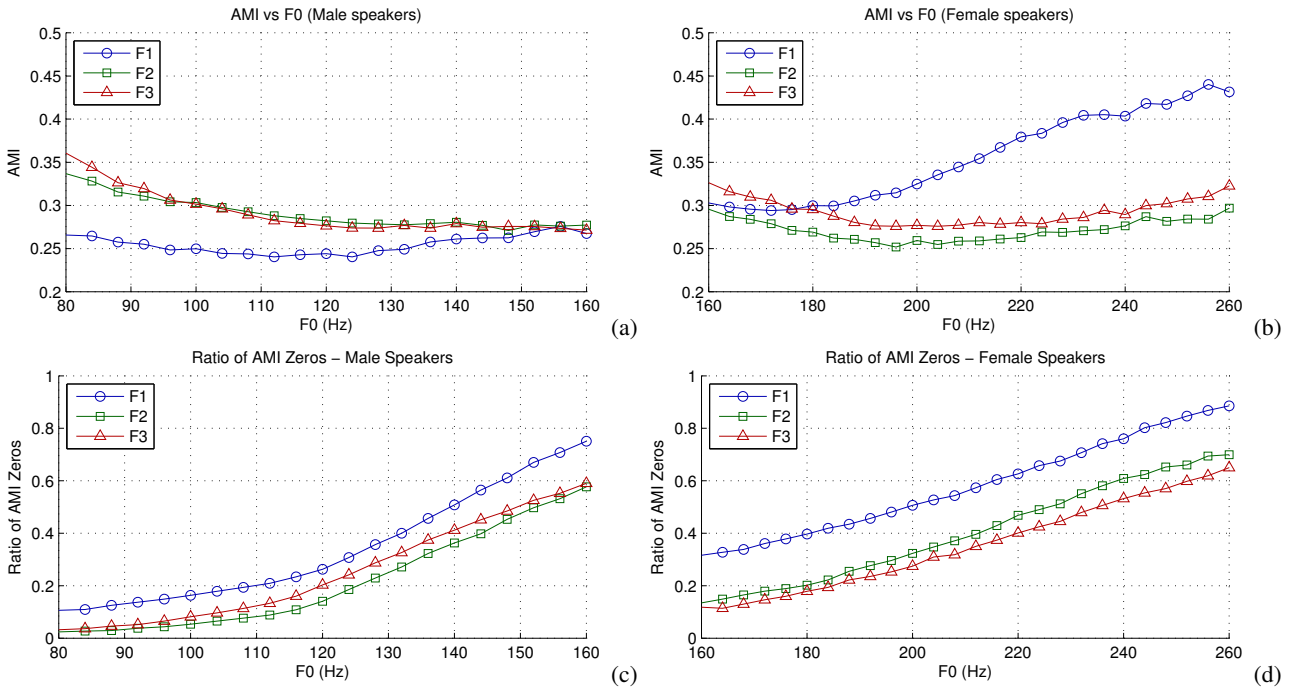


Figure 3: The *AMI* versus F0, excluding cases without secondary pulse, for male (a) and female speakers (b). Percent of cases with absence of secondary pulse are shown as a function of F0 (c),(d).

### 3.3. The F0-F1 tuning

The mean *AMI* and *FMI* estimates for the first formant are further examined in more detail for both genders. More specifically, they are examined in relation to the ratio of the F1 value over the F0 value ( $F1/F0$ ). Fig. 4 (a) shows *AMI* versus the

$F1/F0$  ratio. One can see strong peaks for integer ratio values, both for male and female speakers, which shows that amplitude modulation is amplified when the formant value coincides with a pitch harmonic. Fig. 4 (b) shows the corresponding *FMI* plots. *FMI* has the exact opposite behavior, having valleys in the integer ratio regions, and strong peaks in between.

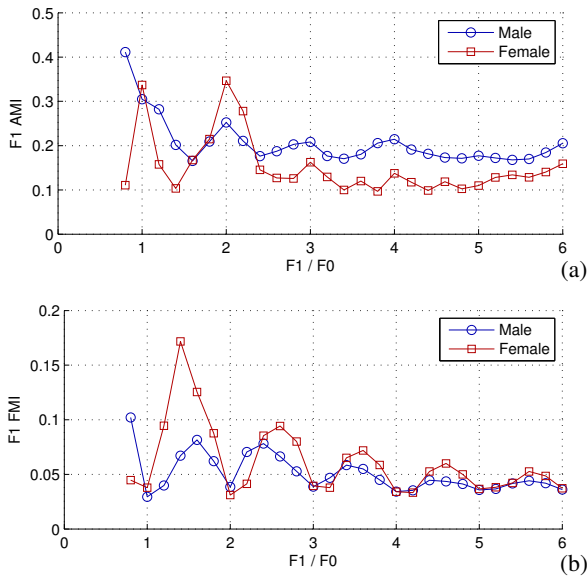


Figure 4: The  $AMI$  and  $FMI$  estimated for  $F1$  versus the ratio between  $F1$  and  $F0$  values.

#### 4. Discussion

The results show that amplitude modulation is clearly more prominent in low  $F0$  conditions. This could be due to the fact that lower  $F0$  allows for more time to achieve complete glottal closure, leading to a more prominent secondary excitation(s) within the pitch period [15]. The range of  $AMI$  and  $FMI$  values is consistent across gender, which suggests that the trends are related to phenomena arising during the glottal cycle (i.e., AM is mostly generated at the source). It is possible that the secondary excitations, that are reflected as AM patterns, arise more frequently when the tension of the vocal cords recedes.

It is interesting to note that although, the percent of vowel instances without modulation increases as a function of  $F0$ , once modulations are present the value of  $AMI$  remains relatively constant as a function of  $F0$ . Thus, although modulation phenomena are less frequent as  $F0$  increases, the level of modulation remains equally strong. In fact, for female speakers and high  $F0$ ,  $AMI$  increases, which may be due to increased source-tract coupling as  $F0$  (or multiples of  $F0$ ) approaches  $F1$ .

$FMI$  shows little variation as a function of the  $F0$  value. This suggests that FM patterns are not seriously affected by glottal source phenomena. This is expected, since  $FMI$  mostly measures variations of the resonant frequencies, which are mostly affected by the vocal tract shape and the subglottal pressure, and less so by the source characteristics.

Figure 4 shows strong evidence of mode locking between  $F0$  and  $F1$ . More specifically, when the  $F1/F0$  ratio is integer, i.e. when  $F1$  coincides with a pitch harmonic,  $AMI$  is amplified while  $FMI$  is reduced. These complementary tendencies of  $AMI$  and  $FMI$  reflect the different phenomena that drive the AM and FM patterns. The AM patterns are more related to the glottal source, while FM patterns are related to the vocal tract. The locking patterns between  $F1$  and  $F0$  are strong indications for non-linear interaction between the glottal source and the vocal tract. It is possible that secondary sources of excitation are amplified when pitch harmonics are at the vicinity of the formant frequency. Other studies also report non-linear phenomena when harmonics and formants coincide [4, 6].

#### 5. Conclusions

The *Amplitude Modulation Index (AMI)* and *Frequency Modulation Index (FMI)* are defined and estimated for the first three formants for vowels and diphthongs. Both  $AMI$  and  $FMI$  are significantly affected by  $F0$ , as well as its relation with  $F1$ . Modulation patterns are directly related to the pitch value, with AM being more frequent in low  $F0$  conditions. Moreover, the tuning of pitch harmonics with the first formant affect both amplitude and frequency modulation patterns. Amplitude modulation is amplified when a harmonic coincides with the  $F1$  value, while at the same time frequency modulation recedes. Overall, this work targets the better understanding of modulation patterns in speech, their relation to the physics behind non-linear phenomena, and the relevance of modulation for speech applications. The conclusions will help us devise new modulation based features, to target the specifics of each application. Further research is needed towards the theoretical modeling of the physics of speech production, as well as extra experimentation, that would better explain the observed non-linearities of speech.

#### 6. References

- [1] H. M. Teager and S. M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," *Speech production and speech modelling*, vol. 55, 1990.
- [2] W. Zhao, C. Zhang, S. H. Frankel, and L. Mongeau, "Computational aeroacoustics of phonation, Part I: Computational methods and sound generation mechanisms," *J. Acoust. Soc. Amer.*, vol. 112, pp. 2134–2146, 2002.
- [3] T. V. Ananthapadmanabha and G. Fant, "Calculation of true glottal flow and its components," *Speech Communication*, vol. 1, no. 3/4, pp. 167–184, 1982.
- [4] I. R. Titze, "Nonlinear source-filter coupling in phonation: Theory," *J. Acoust. Soc. Amer.*, vol. 123, no. 5, pp. 2733–2749, 2008.
- [5] M. Zañartu, L. Mongeau, and G. R. Wodicka, "Influence of acoustic loading on an effective single mass model of the vocal folds," *J. Acoust. Soc. Amer.*, vol. 121, pp. 1119–1129, 2007.
- [6] H. Hatzikirou, W. Fitch, and H. Herzog, "Voice instabilities due to source-tract interactions," *Acta Acustica united with Acustica*, vol. 92, no. 3, pp. 468–475, 2006.
- [7] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Sig. Proc.*, vol. 41, no. 10, pp. 3024–3051, October 1993.
- [8] A. Potamianos and P. Maragos, "Speech analysis and synthesis using an AM-FM modulation model," *Speech Communication*, vol. 28, pp. 195–209, July 1999.
- [9] 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.
- [10] P. Tsiakoulis, A. Potamianos, and D. Dimitriadis, "Short-time instantaneous frequency and bandwidth features for speech recognition," in *ASRU*, 2009.
- [11] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 16, no. 6, pp. 1097–1111, August 2008.
- [12] C. R. Jankowski Jr., T. F. Quatieri, and D. A. Reynolds, "Fine structure features for speaker identification," in *ICASSP*, 1996.
- [13] P. Tsiakoulis and A. Potamianos, "Statistical analysis of amplitude modulation in speech signals using an AM-FM model," in *ICASSP*, 2009.
- [14] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On amplitude and frequency demodulation using energy operators," *IEEE Trans. Sig. Proc.*, vol. 41, no. 4, pp. 1532–1550, April 1993.
- [15] H. M. Hanson and E. S. Chuang, "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data," *J. Acoust. Soc. Amer.*, vol. 106, no. 2, pp. 1064–1077, 1999.