INDIRECT ESTIMATION OF FORMANT FREQUENCIES THROUGH MEAN SPECTRAL VARIANCE WITH APPLICATION TO AUTOMATIC GENDER RECOGNITION

U. K. Laine[1], O. J. Räsänen[1]

[1]Helsinki University of Technology, Department of Signal Processing and Acoustics, Espoo, Finland

Abstract: **A novel approach for estimation of speaker specific vocal tract properties is presented in this paper. Instead of using the well-known long-term average spectrum (LTAS) of speech, it is shown that the variance of the magnitude of the spectrum in each band is also suitable for estimation of formant frequencies. This representation, called mean spectral variance (MSV), is applied to an automatic gender classification task, where it is shown to achieve good classification accuracy in combination with the fundamental frequency of speech. The MSV is compared with LTAS and their similarities and differences are discussed.**

Keywords: **Formant estimation, gender classification, long-term feature averaging**

I. INTRODUCTION

Speaker dependent variability in vocal apparatus properties has a notable impact on the acoustic properties of speech signals. Cross-speaker variation in characteristic formant frequencies poses a difficult challenge for speech processing systems designed to work independently of speaker identity, while it also plays an important role in speaker identity detection [1] and gender classification (e.g., [2]).

One possible approach for analyzing speaker and gender specific properties of the vocal apparatus is through long-term averaging of the acoustic parameters [3]. The long-term average spectrum (LTAS) has been widely studied in speaker recognition, and although its performance falls behind state-of-the-art Gaussian-mixture models (GMM) using Mel-cepstral coefficients (MFCCs), the computational simplicity of LTAS is appealing for many applications [4-5]. In addition to LTAS, averaging of, e.g., autocorrelation-, LPC-, cepstral-, and reflection coefficients, have also been studied [6].

However, all these studies have concentrated on the averaging of feature vectors per se, but none to our knowledge have studied modeling of feature variance in isolation of the spectral mean. In this paper we show that instead of utilizing the long-term spectrum directly, the spectral variability of speech signals also reflects the speaker and gender specific average formant structure. For estimation of speaker specific acoustic parameters, we introduce a straightforward method for estimating average formant frequencies (AFF) indirectly from continuous speech. More specifically, we show that the AFFs can be easily obtained by computing the mean spectral variance (MSV) separately for each frequency band during voiced speech. The basic idea behind our method is simple; while each formant is moving mainly around its mean value these movements should cause the largest spectral variance to occur around the mean as well.

The MSV representation is compared to the well-known LTAS, and it is shown that the methods contain complementary information regarding speech signals. The general quality and usability of the MSV method is assessed in a classification task where MSV templates and pitch of the speaker are combined as cues to perform automatic gender detection.

II. METHODS

A. Computation of mean spectral variance (MSV)

The speech signal (f_s = 16 kHz) is first pre-emphasized with a standard 1st order FIR-filter. Voicing is estimated using standard cepstral analysis and only voiced frames are preserved for further analysis. The signal is then windowed using a 6 ms Hamming window with 2 ms window shifts. The small window length causes the absence of pitch periodicity in spectral representations and leads to regularly good matches between window position and the maximal excitation of vocal tract resonances during glottal closure. Spectral tilt and mean are removed from each frame by fitting a line to the spectrum and the frames are normalized into unit vectors. All spectral frames are collected into a spectrogram and the mean spectral variance for each frequency band is computed over the entire set of frames to produce the MSV representation. The tilt and mean of the MSV are removed and then this vector is normalized to a unit vector. In addition to MSV, the long-term average spectrum (LTAS) is extracted from the speech material. The procedure for LTAS is identical to MSV except that the mean of the spectrum is taken over the spectrogram instead of the variance.

Figures 1 and 2 illustrate the LTAS and MSV representations computed over several speakers from the TIMIT corpus. The average, gender specific, formant structure is readily seen. The AFF estimates provided by both methods are relatively close to each other as predicted. Two general observations can be made; first, both genders most actively utilize the frequency band of 300-3400 Hz that was historically selected to be the band of analog telephone systems (see Fig. 1 bottom frame), and secondly, the shape of MSV between genders is very contrastive in the 1000-5000 Hz frequency band.

B. *Automatic gender detection based on formant structure and pitch*

There are notable structural differences in the vocal tracts for men and women, and therefore the average formant information can be utilized for automatic detection of speaker gender (e.g., [7]). In addition, vocal fold structure can be considered as at least partially independent of vocal tract length (cf., e.g., source-filter modeling), and it also serves as a reliable cue to gender identity. Therefore the mean pitch of a speaker is also utilized in the recognition process.

In the training of the recognizer, MSV vectors $\boldsymbol{v}_\mathrm{m}$ and $\boldsymbol{v}_\mathbf{f}$ are computed across several speakers from the TIMIT training set (N = 560 for both genders) in order to estimate the average male and female spectral structures with formant peaks. The common mean $\boldsymbol{v}_\mathrm{c}=(\boldsymbol{v}_\mathbf{m}+\boldsymbol{v}_\mathbf{f})/2$ of the vectors is subtracted from both $\boldsymbol{v}_\mathrm{m}$ and $\boldsymbol{v}_\mathrm{f}$ in order to maximize contrast:

$$v_g^{'} = v_g - v_c \qquad (1)$$

Finally, the obtained templates are normalized to unit vectors.

Only variation in the frequency band of 1000-5166 Hz is used for recognition, since it was found to lead to maximal performance. The use of this frequency band is also in line with the work of Mendoza et al. [5], who performed a statistical discriminant analysis of male and female voices and found that gender specific differences in LTAS are concentrated in the frequency region of 0.8 – 5 kHz.

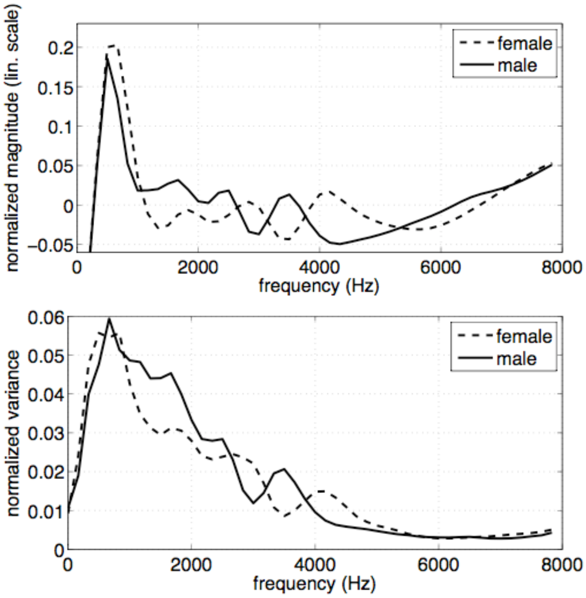

Figure 1: Average LTAS (top) and MSV (bottom) according to gender from the TIMIT training corpus.

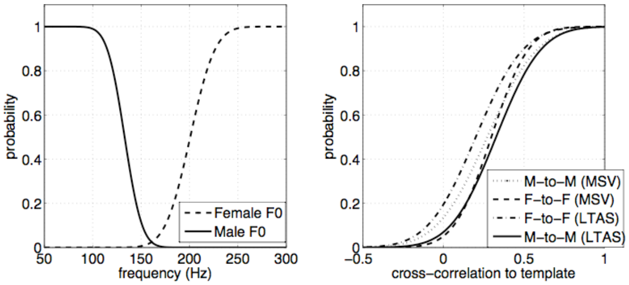

Figure 2: Gender specific cumulative probability distributions for pitch (left) and MSV & LTAS (right).

Once the template vectors for both genders are created, the training set is processed again and the distance d_g between MSV of the analyzed utterance and the templates is measured by cross-correlation. Distributions of d_g values from male utterances to the male template and female utterances to the female template are modeled as a cumulative normal distribution (fig. 2, right). Pitch is also modeled for both genders as two separate cumulative Gaussian distributions estimated from the training data (Fig. 2, left).

In the classification phase, MSV is computed from the input utterance according to section 2.A and vector $\boldsymbol{v}_\mathrm{c}$ is again subtracted from the representation. The mean pitch of the utterance is also extracted. Ultimately, the probability for a gender is estimated using the trained probability distributions and by assuming the independence of probabilities:

$$P(gender) = P(f_0|gender) * P(d_g|gender) \qquad (2)$$

where f_0 is the mean pitch of the utterance and d_g is the cross-correlation between the gender specific MSV template $\boldsymbol{v}'_g$ and the MSV representation estimated from the utterance. When LTAS is used for comparison, the same training and classification procedure is used to obtain gender templates and respective cross-correlation distributions.

III. RESULTS

A. The templates

Fig. 3 shows the obtained limited-band templates used for gender classification for both MSV and LTAS. The structure of both features clearly differentiates between male and female speakers. Although the behavior at higher frequencies is quite similar for both MSV and LTAS, there are notable differences in the region between 1 and 2.5 kHz. One major difference is that the male LTAS contains two peaks at approximately 1300 Hz and 2300 Hz, whereas the male MSV has only one wide peak in between centered around 1800 Hz. Since the range of male F2 is usually between 900 Hz and 2300 Hz, and F3 receives values between 1700 Hz and 3000 Hz [7], this may suggest that MSV computed from sub-pitch periodic windows reacts more strongly to the movement of formants (describing their frequency range) whereas LTAS indicates mean formant locations. MSV peaks are slightly wider than LTAS peaks also at higher frequencies, thus supporting this assumption.

It is also well known that active articulation mainly affects the three lowest formants (especially the second), whereas higher formants are more stationary, reacting relatively passively to articulatory movements. This is also reflected in both the LTSA and MSV templates, where the shape of normalized mean and variance models approach each other at higher frequencies.

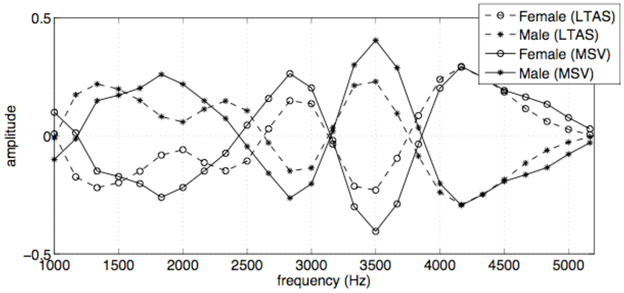

Figure 3: MSV and LTAS templates used in recognition.

B. Baseline classification results

When gender classification is evaluated with the TIMIT test set (56 males and 56 females, 10 utterances per speaker, 1120 utterances in total), a correct classification rate of 98.6 % is achieved (Table 1). This compares well with the approaches reported in the literature. For example, Zeng et al. [8] achieved a 98.2 % gender classification accuracy using a GMM based approach. Vergin et al. [2] achieved a classification rate of 85 % with a different corpus by using the average values of the two first formants as reference values for gender classification. Interestingly, they reported that no improvement was gained by including the higher formants, whereas the current approach leads to optimal results when the analyzed frequency region includes formants F2-F4 (1000 Hz – 5166 Hz) but not F1.

Table 1: Gender classification results for the full TIMIT test set (560+560 utterances).

gender	F0+MSV	F0+LTAS	LTAS	MSV	F0
male	99.3	98.8	82.9	85.7	98.6
female	97.9	97.3	87.0	84.3	95.7
mean	***98.60***	98.05	84.95	85.00	97.15

While MSV and LTAS both carry information regarding gender identity, their overall effect is small compared to F0, which alone leads to an over 97 % classification rate.

C. Feature combinations and noise

To gain a better insight of feature performance in different signal conditions, the gender classification task was performed separately for each possible combination of the three features (F0, MSV and LTAS) using a subset of 300 + 300 utterances (30 + 30 speakers) from the TIMIT test set. Three different noise conditions were used: the clean signal, and SNRs of 20 dB and 10 dB (Table 2).

The results indicate that MSV + F0 again yield the best recognition results (98.5 %), although the differences to LTAS + F0 and MSV + LTAS + F0 are not large. Although the recognition result at 10 dB SNR is still above 90 %, the noise robustness of this approach falls behind a GMM-model using F0 and RASTA-PLP features, where gender recognition rates of 97.9 % for an SNR = 20 dB and 97.5 % for an SNR = 10 dB have been reported [8]. The results obtained with solely LTAS are in line with previous gender recognition systems (e.g., [9], where the LTAS above 1 kHz was used for classification).

Table 2: Gender recognition results for different feature sets in noise (TIMIT test, 300 + 300 utterances).

Features	Male	Female	Mean
Clean speech (SNR = ∞)			
MSV + LTAS + F0	99.1	97.0	98.05
LTAS + F0	98.7	96.3	97.50
MSV + F0	100.0	97.0	**98.50**
MSV + LTAS	89.0	88.3	88.65
MSV	89.7	86.7	88.20
LTAS	87.0	85.3	86.15
F0	99.0	93.3	96.15
White noise (SNR = 20 dB)			
MSV + LTAS + F0	98.0	97.3	**97.65**
LTAS + F0	98.0	97.0	97.50
MSV + F0	99.0	96.0	97.50
MSV + LTAS	91.7	83.3	87.50
MSV	90.0	76.7	83.35
LTAS	88.7	82.3	85.50
F0	97.3	94.0	95.65
White noise (SNR = 10 dB)			
MSV + LTAS + F0	86.0	97.7	**91.85**
LTAS + F0	86.0	97.3	91.65
MSV + F0	87.3	96.0	91.65
F0	87.7	95.3	91.50
MSV + LTAS	80.0	79.3	79.65
MSV	78.0	72.3	75.15
LTAS	77.7	83.3	80.50

A closer error analysis revealed that while MSV and LTAS have a similar overall performance on clean speech, they do not always make errors in the same utterances. In 76 cases of the total 600 utterances (clean speech), MSV and LTAS were giving contradictory information, i.e., one of the two was supporting the wrong gender hypothesis. However, the probabilistic framework used in the recognition compensates for this by assigning small probabilities to features that do not match either of the models. When the SNR drops to 10 dB, MSV performs significantly worse than LTAS, which is a reasonable result since white noise has a larger impact on the variance than the mean.

IV. CONCLUSIONS

A straightforward and efficient method for estimating the average formant frequencies (AFF) through mean spectral variance (MSV) from continuous speech was presented in this paper. As predicted, the MSV method provides comparable AFF estimates compared with those of long-term average spectrum (LTAS).

The usefulness of this approach was demonstrated in a gender classification task where speaker-specific MSV-information and pitch were combined in a straightforward manner as cues for gender identity. In addition, MSV was compared and combined with LTAS. The achieved gender classification rate compares well to other approaches reported in the literature (e.g., [2], [8]) and MSV performance was slightly higher than LTAS for clean speech. However, and as can be expected, MSV is not a particularly robust feature for long-term averaging in severe white noise. The obtained gender classification results are also in line with previous literature, showing that F0 alone is a very strong cue to gender identity in speech.

REFERENCES

[1] Faundez-Zanuy M. & Monte-Moreno E.: State-of-the-art in speaker recognition. *IEEE Aerospace and Electronic Systems Magazine,* Vol. 20, No. 5, pp. 7-12, 2005

[2] Vergin R., Farhat A. & O'Shaughnessy D.: Robust gender-dependent acoustic-phonetic modeling in continuous speech recognition based on a new automatic male/female classification. *Proc. ICLSP'96,* pp. 1081-1084, 1996

[3] Markel J. D., Oshika B. T. & Gray A. H.: Long-Term Feature Averaging for Speaker Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing.* Vol. 25, No. 4, pp. 330-337, 1977

[4] Kinnunen T., Hautamäki V. & Fränti P.: On the use of long-term average spectrum in automatic speaker recognition. *International Symposium on Chinese Spoken Language Processing (ISCSLP'06)*, Singapore, pp. 559-567, 2006

[5] Mendoza E., Valencia N., Muñoz J. & Trujillo H.: Differences in Voice Quality Between Men and Women: Use of the Long-Term Average Spectrum (LTAS). *Journal of Voice*, Vol. 10, pp. 59-66, 1996

[6] Wu K. & Childers D. G.: Gender recognition from speech. Part I: Coarse analysis. *J. Acoustical Society of America*, Vol. 90, No. 4, pp. 1828-1840, 1991

[7] Hillenbrand J., Getty L. A., Clark M. J. & Wheeler K.: Acoustic characteristics of American English vowels. *Journal of Acoustical Society of America*, Vol. 97, No. 5, pp. 3099-3111, 1995

[8] Zeng Y.-M., Wu Z.-Y., Falk T. & Chan W.-Y.: Robust GMM based gender classification using pitch and RASTA-PLP parameters of speech. *Proceedings of* 5th *Int. Conference on Machine learning and Cybernetics*, pp. 3376-3379, Dalian, 2006

[9] Hertrich I. & Ziegelmayer G.: Sexual dimorphism in the long term speech spectrum. *Human Evolution,* Vol. 2, No. 3, 1987