RECOGNITION OF VOICED SPEECH FROM THE BISPECTRUM

Anastasios Delopoulos, Maria Rangoussi *and* *Janne Andersen*

Division of Computer Science,
Department of Electrical Engineering,
National Technical University of Athens,
GR-15780, Athens, Greece
e-mail: {delo, maria}@softlab.ece.ntua.gr

Digital Signal Processing Group,
Institute for Electronic Systems,
Aalborg University,
DK-9220, Aalborg, Denmark

ABSTRACT

Recognition of voiced speech phonemes is addressed in this paper using features extracted from the bispectrum of the speech signal. Voiced speech is modeled as a superposition of coupled harmonics, located at frequencies that are multiples of the pitch and modulated by the vocal tract. For this type of signal, nonzero bispectral values are shown to be guaranteed by the estimation procedure employed. The vocal tract frequency response is reconstructed from the bispectrum on a set of frequency points that are multiples of the pitch. An AR model is next fitted on this transfer function. The AR coefficients are used as the feature vector for the subsequent classification step. Any finite dimension vector classifier can be employed at this point. Experiments using the LVQ neural classifier give satisfactory classification scores on real speech data, extracted from the DARPA/TIMIT speech corpus.

1 INTRODUCTION

In a composite continuous speech recognition system, successful operation of the baseline subsystem that recognizes elementary units of speech, is crucial to the success of the following levels of speech recognition and understanding. In the present paper we aim towards *recognition of voiced speech phonemes,* based on a set of features extracted parametrically from the *bispectrum* of the speech signal.

The basic assumption underlying the proposed method is that voiced speech can be modeled as superposition of coupled harmonics, located at frequencies that are multiples of the pitch frequency, and modulated by a linear AR filter which models the vocal tract, [1]. The analysis proposed in this work employs the bispectrum of the voiced speech signal, [3]. The latter is proved here to yield non-zero estimates from finite-length speech records - a behavior earlier observed in [7]. The bispectrum serves as a basis for the reconstruction of the vocal tract transfer function on a specific set of frequency points. This involves inverting the vocal tract input-output relation in the bispectrum domain. A modification of the signal reconstruction algorithm of [5] is employed at this point. An AR model is next fitted on the vocal tract transfer function, using standard IIR filter design methods. AR filters are known to provide a good model for the vocal tract during voiced speech production. The AR parameters form the set of features upon which classification of voiced phonemes is performed. Any standard classifier can be used at this step, such as the LVQ neural network classifier.

Because of its third-order domain basis, the proposed method offers the advantages (i) of providing a more accurate model of the phoneme-dependent vocal tract transfer function, as it retains phase information, and (ii) of being robust to all symmetrically distributed additive noises. In addition, it is proved to be robust to harmonic components present in the additive noise, as far as their frequencies are not coupled with the pitch. This latter feature of the proposed method is of significant practical interest. Limited-scale recognition experiments using synthetic and real speech data give results that compare favorably with those of standard methods. In that sense, the proposed method provides a viable alternative, which in addition enhances our understanding of certain aspects of the bispectrum of voiced speech.

2 A MODEL FOR VOICED SPEECH SIGNALS

A simplified model for the physiological mechanism of speech production is shown in Figure 1, where the switch position is up [down] for voiced [unvoiced] speech, [1]. Speech signal is the output of a linear, AR-type filter, whose transfer function is

$$H(\omega) = 1/A(\omega) = 1/\sum_{i=0}^{p} a_i e^{j\omega i}, \qquad (1)$$

and which models the vocal tract along with the position of the lips and tongue. The filter is assumed practically time-invariant for the duration of the a certain phoneme. In the *voiced speech* case, the input $i(t)$ is an impulse train whose impulses are located at positions that are multiples of the fundamental *pitch period*. This signal is equivalent to a superposition of harmonics whose frequencies are multiples of the pitch frequency, by the

Poisson summation formula, [4]. This superposition of "equispaced" harmonics can be split into triplets of the general form

$$x(t) = A_1\, e^{j(\lambda_1 t+\phi_1)} + A_2\, e^{j(\lambda_2 t+\phi_2)} + A_3\, e^{j(\lambda_3 t+\phi_3)}, \quad (2)$$

where $\lambda_3 = \lambda_1 + \lambda_2$. To obtain the impulse train $i(t)$ from this expression, one should use $A_{1,2,3} = 1$ and $\phi_{1,2,3} = 0$ in each triplet. Consequently, frequency coupling is present in the input signal $i(t)$.

Because the vocal tract transfer function is linear, output contains harmonics of the same frequencies, modulated in amplitude and shifted in phase. Coupling is therefore reproduced in the output signal, which can be described as

$$\begin{aligned} s(t) \;&=\; \sum_{l=1}^{L} A_{1,l}(\lambda_{1,l})\, e^{j(\lambda_{1,l}t+\phi_{1,l})} && (3) \\ &+\; A_{2,l}(\lambda_{2,l})\, e^{j(\lambda_{2,l}t+\phi_{2,l})} + A_{3,l}(\lambda_{3,l})\, e^{j(\lambda_{3,l}t+\phi_{3,l})}, \end{aligned}$$

where $A_{i,l}(\lambda_{i,l}),\ i = 1, 2, 3$ are the amplitude modulation coefficients and $\phi_{i,l}(\lambda_{i,l}),\ i = 1, 2, 3$ are the phase delays caused by the AR filter. A synthetic example of the output (speech) signal of such a model is shown in Figure 2 (time and frequency domain). Frequency coupling is clearly present in the output.

The signals $y(t)$ considered here contain voiced speech contaminated by additive noise,

$$y(t) = s(t) + \sum_{m=1}^{M} A_m e^{j\lambda_m t} + v(t), \quad (4)$$

where the last two components account for the noise, including harmonics and random noise $v(t)$. The harmonics should not be coupled in frequency with the signal, i.e. their frequencies should not lie on multiples of the pitch, and the random noise $v(t)$ is assumed to be symmetrically distributed.

3 BISPECTRUM ESTIMATES OF VOICED SPEECH

For a signal $s(t)$ as in eq. (3) and under the alternative condition that phases $\phi_{i,l}$ are random variables uniformly distributed in $[-\pi, \pi]$, it is shown in [6] that the bispectrum $B_{3s}(\omega_1, \omega_2)$ is identically zero if only quadratic frequency coupling is present, while it is non-zero if both quadratic frequency and phase coupling ($\phi_{3,l} = \phi_{1,l} + \phi_{2,l}$) are present. In the latter case, the bispectrum contains impulses at the coordinates of the coupled frequencies, (λ_1, λ_2).

In the present work we show that phase coupling and, consequently, non-zero values of the bispectral estimates, can be produced through the estimation procedure employed (*bi-periodogram* with averaging over segments). The bispectrum is estimated as

$$\begin{aligned} \hat{B}_{3s}(\omega_1, \omega_2) \;&=\; \frac{1}{K}\sum_{k=1}^{K} \hat{B}_{3s,k}(\omega_1, \omega_2) && (5) \\ &=\; \frac{1}{K}\sum_{k=1}^{K} S_k(\omega_1)\; S_k(\omega_2)\; S_k^*(\omega_1+\omega_2), \end{aligned}$$

where $S_k(\omega)$ is the Fourier transform of the signal segment $s_k(t) \triangleq s(t+k),\ t = 0, 1, \ldots, N-1$ of length N, starting at time point k.

Due to eq. (3) the k-th segment $s_k(t)$ can be written as

$$\begin{aligned} s_k(t) \;&=\; \sum_{l=1}^{L} A_{1,l}(\lambda_{1,l})\, e^{j(\lambda_{1,l}t+\Phi_{1,l}^k)} && (6) \\ &+\; A_{2,l}(\lambda_{2,l})\, e^{j(\lambda_{2,l}t+\Phi_{2,l}^k)} + A_{3,l}(\lambda_{3,l})\, e^{j(\lambda_{3,l}t+\Phi_{3,l}^k)}, \end{aligned}$$

where $\Phi_{i,l}^k = [\phi_{i,l} + k\lambda_{i,l}] \mod 2\pi,\ i = 1, 2, 3$. If we use $\lambda_{3,l} = \lambda_{1,l} + \lambda_{2,l}$ in the equation above, we obtain

$$\Phi_{3,l}^k = \Phi_{1,l}^k + \Phi_{2,l}^k + C(\lambda_{1,l}, \lambda_{2,l}). \quad (7)$$

$C(\lambda_{1,l}, \lambda_{2,l}) = \phi_{3,l} - \phi_{1,l} - \phi_{2,l}$ is a constant w.r.t. k.

If the starting points k of the corresponding segments $s_k(t)$ are chosen at random, then the phases $\Phi_{i,l}^k$ are random variables, uniformly distributed in $[-\pi, \pi]$. This property is retained if the starting points are equispaced, provided that we (i) exclude the special case where the segment spacing is an exact multiple of some $T_{i,l} = 1/\lambda_{i,l}$ and (ii) let the number K of segments employed in eq. (5) become large.

The bispectrum of such signals takes on zero values except at the bi-frequency points $[\lambda_{1,l}, \lambda_{2,l}]$ where it peaks. The proof of this fact for *non-zero* values of the deterministic phase offset $C(\lambda_{1,l}, \lambda_{2,l})$ is an extension of the corresponding proof for $C = 0$ in [6].

As a result, the bispectrum estimate of the output signal $y(n)$, $\hat{B}_{3y}(\omega_1, \omega_2)$, obtained through the above estimation procedure, takes on non-zero values only on a grid of bi-frequency points that are multiples of the pitch. Here we have used the fact that $\hat{B}_{3y}(\omega_1, \omega_2) \equiv \hat{B}_{3s}(\omega_1, \omega_2)$, because under the assumptions made in the previous section the bispectra of the last two components in eq.(4) are (ideally) zero. The non-zero bispectral values, given by

$$\hat{B}_{3y}(\lambda_{1,l}, \lambda_{2,l}) \approx H(\lambda_{1,l})H(\lambda_{2,l})H^*(\lambda_{1,l} + \lambda_{2,l}) \quad (8)$$

(apart from a scalar ambiguity factor) leak into smoother peaks, because of the finite window imposed on the data by the estimation procedure. The pitch that will give us the grid of bi-frequency points, can be obtained either independently, or by picking and refining the peak values of $\hat{B}_{3s}(\lambda_{1,l}, \lambda_{2,l})$. This behavior is illustrated in Figure 4, which shows the 3D-plot and contours of the first quadrant of the bispectrum estimated from the voiced phoneme /ah/, shown in Figure 3. The pitch in this case is 0.09 rad.

4 ESTIMATION AND MODELING OF THE VOCAL TRACT TRANSFER FUNCTION

The transfer function $H(\omega)$, and consequently the AR coefficients $a_i, i = 1, \ldots, p$ that characterize the vocal tract, can be obtained if we solve eq. (8) for $H(\omega)$. This employs the solution of two overdetermined linear systems, one for the log-magnitude $\mu(\omega)$ and one for the phase $\phi(\omega)$ of $H(\omega)$. Both linear systems involve bispectral ordinates at bi-frequency plane positions $[\lambda_{1,l}, \lambda_{2,l}] = [k_1 f_p, k_2 f_p]$, where f_p denotes the pitch and k_1, k_2 are positive integers. The systems are formed by concatenation of equations obtained if we equate (i) phases and (ii) log-magnitudes of the two sides of eq. (8), along the lines of [5], namely

$$\phi_{3s}(k_1, k_2) = \phi(k_1) + \phi(k_2) - \phi(k_1 + k_2), \qquad (9)$$

$$\mu_{3s}(k_1, k_2) = \mu(k_1) + \mu(k_2) + \mu(k_1 + k_2), \qquad (10)$$

where $\hat{B}_{3s}(k_1, k_2) = \exp[\ \mu_{3s}(k_1, k_2)]\ \exp[j\phi_{3s}(k_1, k_2))$. The systems formed are overdetermined and are solved for $\mu(k)$ and $\phi(k)$ in the least squares sense. The estimated transfer function ordinates, $\hat{H}(k) = \exp[\mu(k)]\ \exp[\ j\phi(k)\ ]$ lie on frequency points kf_p that are multiples of the pitch.

The transfer function estimates characterizes the vocal tract and it could be used by itself as a feature for classification. However, because of the employed estimation method, the estimated ordinates $\hat{H}(k)$ correspond to pitch-dependent frequency points. To remove this dependency and obtain feature vectors of equal dimensions, we next model $\hat{H}(\omega)$ as an AR filter of a given order, say p, and use the AR coefficients as the feature set for classification. Standard IIR filter design methods can be used in this step.

5 CLASSIFICATION OF VOICED SPEECH PHONEMES

Following the procedure described above, classification can be based on the set of AR coefficients characterizing the vocal tract, or on a subset of them. Figure 5 shows the estimated transfer functions computed for five utterances of the voiced phonemes /aa/ (top) and /iy/ (bottom). The corresponding signals are obtained from DARPA/TIMIT database, and belong to different speakers, under different contexts. The corresponding AR($p = 7$) coefficients are shown in Figure 6, top and bottom, respectively. These figures illustrate the fact that the AR coefficients are discriminative features.

Any finite-dimension vector classification scheme can be employed next. Experiments using the Learning Vector Quantization (LVQ) classifier, [2], have given satisfactory classification scores on real speech data from TIMIT.

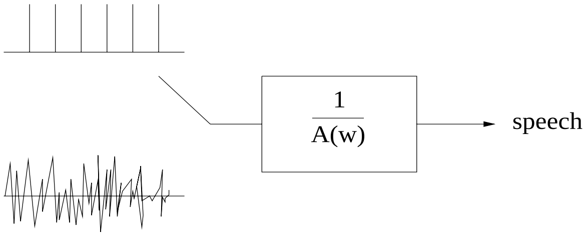

Figure 1: Speech production model

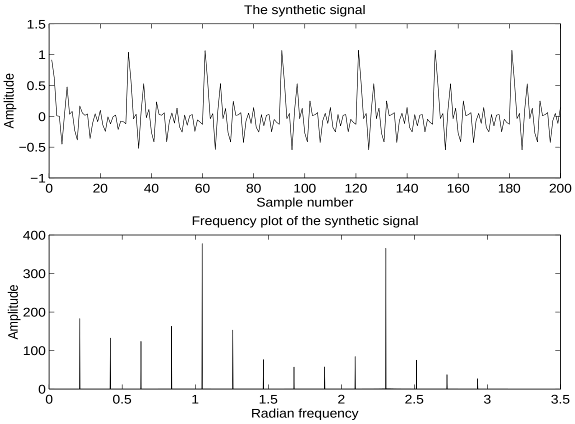

Figure 2: Synthetic speech, time (top) and frequency (bottom) domains.

References

[1] J.R.Deller, J.G.Proakis, J.H.L.Hansen, "Discrete-time processing of speech signals," *McMillan Publish. Co,* 1993.

[2] T.Kohonen, "Improved versions of LVQ," *Proc. IJCNN'90,* vol. 1, pp. 545-550, June 1990.

[3] C.L.Nikias, M.R.Raghuveer, "Bispectrum estimation: A digital signal processing framework," *Proc. of the IEEE,* vol. 75, no. 7, July 1987.

[4] A.Papoulis, "Probability, Random Variables and Stochastic Processes," *McGraw-Hill Int. Ed.,* 1991.

[5] M.Rangoussi, G.B.Giannakis, "FIR modeling using log-bispectra: Weighted least-squares algorithms and performance analysis," *IEEE Trans. on Circuits and Systems,* vol. 38, no. 3, March 1991.

[6] A.Swami, J.M.Mendel, "Cumulant-based approach to the harmonic retrieval problem," *IEEE Proc. ICASSP'88,* vol. 4, pp.2264-2267, Apr. 1988.

[7] G.Zhou, G.B.Giannakis, "Polyspectral analysis of mixed processes and coupled harmonics," *IEEE Trans. on Info Theory,* vol. 42, no.3, pp. 943-958, May 1996.

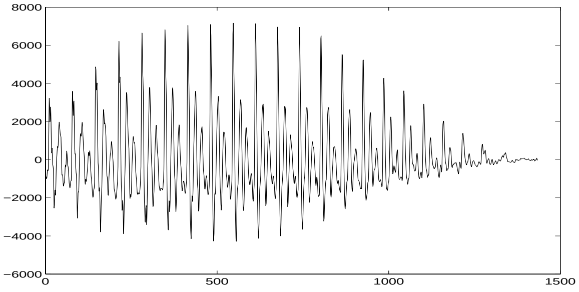

Figure 3: Voiced speech signal /ah/.

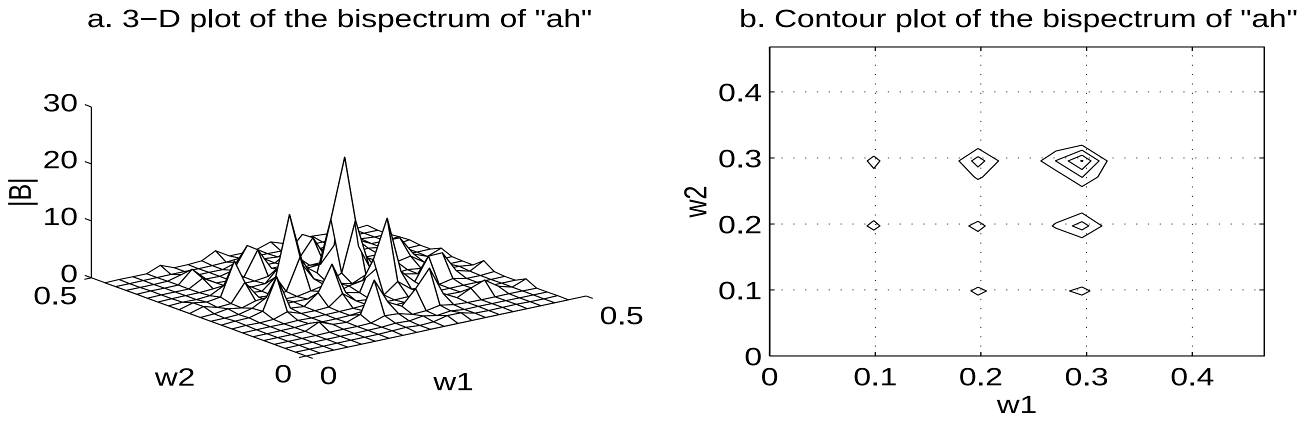

Figure 4: Bispectrum, 1st quadrant and contours.

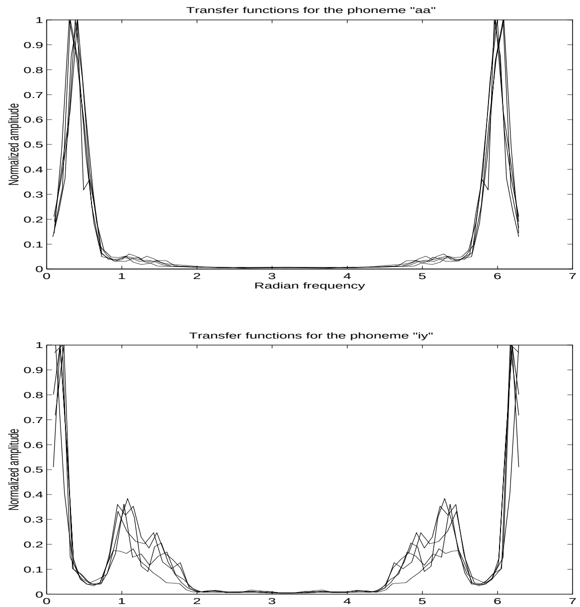

Figure 5: Reconstructed transfer functions of 5 utterances of the phonemes /aa/ (top) and /iy/ (bottom).

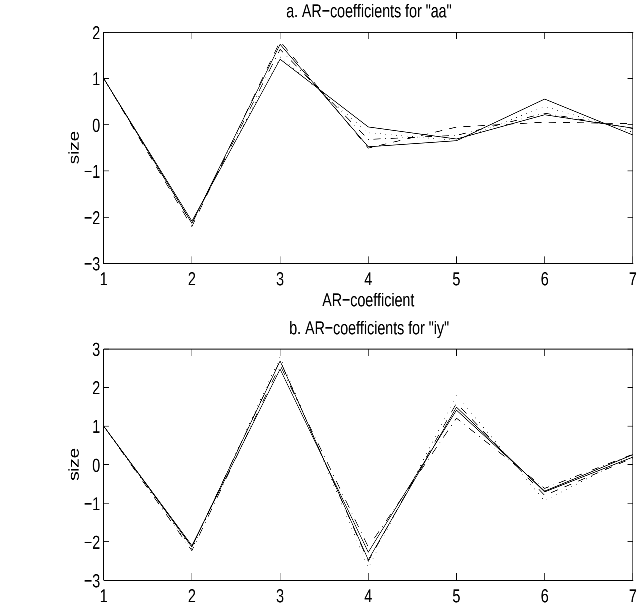

Figure 6: AR(7) coefficients for phonemes /aa/ (top) and /iy/ (bottom).