OPTIMAL MULTI-PITCH ESTIMATION USING THE EM ALGORITHM FOR CO-CHANNEL SPEECH SEPARATION

Dan Chazan[1], *Yoram Stettiner*[2] *and David Malah*[3]

[1] IBM Science and Technology Center - Technion City, Haifa 32000, Israel

[2] Dept. of Electrical Engineering, Technion - Israel Institute for Technology, Technion City, Haifa 32000, Israel
Also with Nexus Telecommunication Systems Ltd., Korazin 1, Givataim 53583, Israel

[3] Dept. of Electrical Engineering, Technion - Israel Institute for Technology, Technion City, Haifa 32000, Israel

ABSTRACT

The paper addresses the problem of optimally estimating (in the ML sense) the pitch of each of several speakers talking simultaneously. This information is needed in systems which perform co-channel speech separation. We propose a multi-pitch model which is used in conjunction with an EM-based iterative estimation scheme. In addition, the pitch period of each speaker is allowed to vary linearly in the analysis interval, thus offering improved co-channel speech separation. The proposed algorithm is shown to outperform standard pitch detection algorithms, in detecting the pitch of simulataneous speakers.

1. INTRODUCTION

Degradation of speech intelligibility due to noise, particularly when that noise is the voice of another speaker, is a recurrent problem in many contexts, for example automatic speech recognition systems and hearing impaired persons. Another recurrent problem is cross-talk over communication channels. Most of the intelligibility loss is due to the voiced portions of the interfering speech [1]. It has already been shown [1-5] that intelligibility may be improved by speaker separation or interference suppression algorithms.[1]

Several approaches to the speaker separation problem have been proposed [2,3 and references therein], all capitalizing on pitch differences between the speakers. The crucial and most difficult part is the accurate estimation of pitch for each speaker directly from the single input, without a-priori information. Another problem is the smearing of spectral harmonics due to varying pitch. Since relatively long frames are needed to effectively separate speakers with close spectral harmonics, even relatively slow variations in pitch cause widening of the harmonics, especially the higher ones. Most works, e.g. [2,3], try to cope with this problem by letting each spectral harmonic deviate from its ideal shape and frequency.

The approach taken in all of the previous works is either to assume that the pitch contours are known a-priori, or exploit one or even several of the many standard single-speaker pitch detection algorithms (PDAs) available [6]. These PDAs generate a list of pitch candidates, which are transformed into pitch contours by pitch tracking algorithms. However, since the underlying model of a standard PDA is that of a single speaker [6], this approach is not optimal when more than one pitch is present, e.g., when both speakers are voiced. The main difficulty is in estimating the pitch of the weaker speaker.

We propose an optimal, yet efficient, Maximum Likelihood (ML) procedure for the simultaneous estimation of the pitch for two or more speakers, based solely on the single co-channel speech signal. Unlike previous works, a multi-pitch model is assumed, and the approach is extended to estimate the pitch rate of change for each speaker. The signal is represented as the sum of quasi-periodic signals, where the pitch period is allowed to vary linearly within the analysis window. Efficient solution of the multidimensional estimation problem is facilitated by an iterative scheme (depicted in Fig. 3 for the two speakers case) based on the Estimate-Maximize (EM) algorithm [8,9]. The ML processor and the signal decomposition functions are derived and shown to be a time-varying comb filter, realized by a time-invariant comb filter, preceded by appropriate time warping and followed by its inverse. Estimation of the pitch period and its rate of change is done by adjusting the time-warping function to maximize the energy of the (inverse) time-warped comb output. In a special case of an EM separation scheme (Fig. 3), each of the composite

[1] We make the distinction between *speaker separation* algorithms, which can exploit the pitch of both the desired speaker and the interferer, e.g. [2,3], and *interference suppression* algorithms which exploit only the pitch of the interference, e.g. [4,5].

0-7803-0946-4/93 $3.00 © 1993 IEEE

comb filters operates on the residual signal of the other. Each speaker is time-warped *separately and independently* of the other speaker. The warped signals have a constant pitch and thus the harmonic smearing problem is avoided, allowing the use of longer analysis frames and enhancing the pitch detection. Finally, voiced/unvoiced decisions can be made by comparing the energies at the input and output of the composite comb filters. We will henceforth refer to the scheme of Fig. 3 as the Multi-Pitch Detection Algorithm (MPDA).

2. THE SINGLE PITCH MODEL

The deterministic model we use for voiced speech of a single speaker is that of a quasi periodic signal, i.e., a strictly periodic signal (inverse) time warped by an invertible function. The model is given by

$$s(t;\theta) = \sum_{k=1}^{L(\phi(t))} C_k\, e^{jk\phi(t)} \qquad t \in \left(-\frac{T}{2}, \frac{T}{2}\right) \qquad (1)$$

where $\{C_k\}$ are time-invariant Fourier coefficients, $\phi(t)$ is an invertible warping function, $L(t)$ is the number of pitch harmonics in the analysis bandwidth. The model parameters are denoted by θ and given by ϕ and $\{C_k\}_{k=1}^{L(t)}$. The derivative of the warping function, $\Omega(t)$, represents the instantaneous pitch frequency in radian/sec and thus it is constrained by the feasible range of human pitch.

Consider a given voiced speech signal $x(t) = s(t) + v(t)$ where $s(t)$ is the model [1] and $v(t)$is a noise process representing modeling errors and additive noise. We assume that $v(t)$is a zero mean additive white Gaussian noise (AWGN) process uncorrelated with $s(t)$. It is well known that Maximum Likelihood (ML) estimation, in the case of AWGN, is equivalent to Least Squares (LS) estimation. The LS problem is then,

$$\underset{s_0,\,\phi}{Min} \int_{-\infty}^{\infty} w(t)\left|x(t) - s_0(\phi(t))\right|^2 dt \qquad (2a)$$

subject to $\quad 0 < a \le \Omega(t) \le b \qquad (2b)$

$$s_0(u) = s_0(u + 2\pi) \quad u = \phi(t) \qquad (2c)$$

where w is a non-negative window function with support on $(-T/2,T/2)$ and s_0 is a *periodic* signal whose Fourier coefficients are $\{c_k\}$. Alternatively, (2a) can be posed in the warped time domain $u = \phi(t)$: Defining,

$$\tilde{w}(u,\,\phi) \triangleq w\left(\phi^{-1}(u)\right)\, \dot{\phi}^{-1}(u) \qquad (3a)$$

$$\tilde{x}(u,\,\phi) \triangleq x\left(\phi^{-1}(u)\right) \qquad (3b)$$

(2a) becomes

$$\underset{\phi,\,s_0}{Min} \int_{-\infty}^{\infty} \tilde{w}(u,\phi)\left|\tilde{x}(u,\phi) - s_0(u)\right|^2 du \qquad (4)$$

Although (2a) and (4) are equivalent LS problems, in the latter s_0 is independent of the warping function. Using results from [7], it can be shown that given ϕ, the s_0 that minimizes (4) subject to (2c) is given by

$$\hat{s}_0(u) = \frac{\sum_{k=-\infty}^{\infty} \tilde{w}(u+2\pi k)\,\tilde{x}(u+2\pi k)}{\sum_{k=-\infty}^{\infty} \tilde{w}(u+2\pi k)} \qquad (5)$$

This is the output of a comb filter tuned to a nominal pitch period of 2π and fed by the time-warped signal with coefficients weighted by the time-warped window function and the Jacobian of the warp.

Substituting (5) and (3) into (4), the minimization need be carried out on the only remaining parameter, the warping function. Let us now restrict ourselves to warping functions of the form

$$u = \phi(t) = \frac{1}{2}\alpha t^2 + 2\pi\beta t \qquad (6a)$$

$$\Omega(t) = \dot{\phi}(t) = \alpha t + 2\pi\beta > 0 \qquad (6b)$$

$$0 < a \le \Omega(t) \le b \qquad ; \quad t \in \left(-\frac{T}{2}, \frac{T}{2}\right) \qquad (6c)$$

Substituting (5) and (6) into (4), we obtain a simplified problem where the minimization should be made over a suitable domain of β and α, which are the average pitch and the pitch rate of change, respectively. In practice, we first set $\alpha = 0$, estimate a coarse β over a 50 to 500 Hz range, and then estimate α and β simultaneously using a gradient based search, where β is now limited to some neighborhood of the coarse estimate.

3. THE MULTI-PITCH MODEL

Co-channel speech is described by

$$y(t) = x_1(t) + x_2(t) = s_1(t;\underline{\theta}_1) + v_1(t) + s_2(t;\underline{\theta}_2) + v_2(t) \qquad (7)$$

where x_1 and x_2 are the target and jammer speakers respectively, s_1 and s_2 are the quasi-periodic models as above for the respective speakers with

$$\underline{\theta_1} \triangleq \left[\alpha_1 \quad \beta_1 \quad s_{0,1}\right] \qquad \underline{\theta_2} \triangleq \left[\alpha_2 \quad \beta_2 \quad s_{0,2}\right] \qquad (9)$$

and v_1, v_2 are assumed to be zero mean uncorrelated AWGN processes, each representing the modeling error of its respective speaker and a portion of the actual additive noise.

The multi-pitch estimation problem requires the simultaneous solution of (4) for the parameters α and β of N summed speakers. This is a non-linear optimization problem in 2N dimensions. Since under the AWGN assumption the LS problem is equivalent to ML estimation, we can apply here the results of [9], where the given signal is decomposed into its components and the parameters of each component are estimated separately. The algorithm iterates back and forth and uses the current estimate of the parameters to decompose the given signal, thus improving the next estimate. Under some regularity conditions [8], convergence to a stationary point of the likelihood function is guaranteed.

Applied to the multi-pitch detection problem, the signal components are estimates of the individual speakers' waveforms, so that the MPDA actually separates the speakers while estimating their pitch. The regularity conditions of [8] are met since with the model (1), all the signal components are continuous with respect to their parameters. In terms of complexity, the EM algorithm decomposes the 2N dimensional ML problem to N separate 2 dimensional ML problems (4) which are solved iteratively. In the terms of [9], the ML Processor solves (4) and the Signal Synthesis is performed according to (1). Together, it is equivalent to finding the warping function that maximize the energy at the output of a cascade of a time-warp, a comb filter tuned to 1 rad/sec, and an inverse time-warp.

In [9] the noise term is arbitrarily divided among the decomposed signal components at each iteration. In the special case where we assign the entire noise term to a single decomposed signal, and rotate this assignment every iteration, the parameters of each speaker are estimated from the residual signal of all the other speakers. The resulting MPDA scheme is shown in Fig. 3 for the case of two speakers and warping functions of the form (6).

4. SIMULATIONS AND RESULTS

Although the MPDA may in principle handle more than two simultaneous speakers, simulations have been conducted with real co-channel speech of two speakers only. A database was constructed from a subset of the DAM database, sampled at 8 kHz and band limited to 3.2 kHz. Speech from four speakers, two males and two females, each speaking 7 sentences, was compacted by removing silence intervals. The resulting speech duration was 50 seconds. Another speech file was created by reordering the speakers. A co-channel speech file was created by adding the first speech file to an attenuated version of the second, where the attenuation ranges from 0 to 20 dB. In the terminology of [1], the Target-to-Jammer Ratio (TJR) therefore ranges from 0 to -20 dB, where the first speech file is the jammer and the second is the target.

Reference pitch contours were made for each speech file before they were summed, using the Super Resolution Pitch Detector (SRPD) [10]. This PDA is highly accurate and is able to track abrupt pitch variations because it compares exactly two consecutive pitch periods. The SRPD provided voice/unvoiced/silence decisions as well.

The MPDA was compared to three other standard PDAs which are capable of detecting more than a single pitch. These PDAs were modified to provide two pitch candidates instead of just one, with the restriction that the second candidate cannot be harmonically related to the first (i.e., half, double ,2/3 etc.). This feature is inherent to the MPDA, since each pitch is estimated from a residual signal which has spectral nulls at the harmonics of the other pitch. The PDAs used were the well known Cepstrum (CEP) [6], the Harmonic Product Spectrum (HPS) [6] and the Pseudo Maximum Likelihood (PML) [6,7] PDAs. The test data was processed by the three PDAs and the MPDA. Trying two types of windows (Hamming and Blackman) and a range of window durations, we got the best results with a 60 ms Blackman window. The analysis bandwidth was 3.2 kHz. Each PDA supplied two pitch candidates. The MPDA almost always converged within 2-3 iterations. In the error analysis that followed, only frames where the SRPD labeled both the target and the jammer as voiced were considered, totaling 2409 frames.

Since for speaker separation purposes octave errors are of secondary importance, the two pitch candidates of each tested PDA were compared not only to the reference pitch values but also to their half and double values. In order to render the pitch errors invariant to octave errors, all pitch values were transformed to an octave scale, same as the one used in music. The procedure went as follows: First, all pitch values were converted to an octave scale by the operation $\log_2(x/440)$ where x is the pitch value in Hz. Then the triplet consisting of the first reference pitch joined by its half and double (obtained by adding 1 and -1 to the value in octaves) was compared to the two pitch candidates of the evaluated PDA, and the absolute difference between the closest pair was defined as the pitch error. The last step was then repeated for the second reference pitch.

The results are depicted in Fig.'s 1 and 2 and Tables 1 and 2 for the MPDA, CEP and HPS PDAs. The PML PDA results are not included as they are similar to those of the CEP PDA. Fig.'s 1 and 2 show the percentage of frames with pitch error (in octaves) exceeding the abscissa, for TJRs of 0 and -12 dB, respectively. The vertical arrow indicates a pitch error of 0.189 octave, equivalent to 1ms at an average pitch of 140 Hz, and classify the pitch errors to gross and fine. Tables 1 and 2 show the percentage of gross errors and the RMS of fine pitch errors (in octaves), calculated for TJRs of 0 and -12 dB, respectively. It is evident that the MPDA consistently outperforms the other PDAs. Its error distribution is better, both for fine and gross errors, and for both TJRs.

For TJR from 0 to -12 dB, when the pitch values of the speakers come within 5% of each other, the standard PDAs tend to detect them as one value while the MPDA can still distinguish between them to the point where they are only 1.5% apart, although with a somewhat slower convergence rate.

In another experiment, three analysis bandwidths of 1.6, 3.2 and 4.8 kHz were tried. It was found that increasing the bandwidth benefitted all the PDAs and especially the

MPDA, in terms of better detection of the weaker pitch and improved discrimination of close pitch values. The MPDA is able to make better use of the increased bandwidth primarily because of its time-warping capability which minimizes the smearing of the higher pitch harmonics.

The MPDA was integrated in a speaker separation system. It considerably improved its performance and extended its effective operation range in terms of TJR.

5. CONCLUSION

We presented a novel multi-pitch detection algorithm (MPDA) which outperforms standard PDAs in a multi-speaker environment. The MPDA has potential in improving the performance of speaker separation and interference suppression systems.

We believe that the next step should be the use of even longer analysis frames in conjunction with a long-term model for voiced speech which incorporates not only the pitch variation but also the other non-stationarities of the signal. This could improve the ability of the MPDA to distinguish between close pitch values, and the ability of the speaker separation system to separate close pitch harmonics. Preliminary results of this approach are reported in a companion paper in this conference.

REFERENCES

[1] Zissman M.A. et al., "Speech-state adaptive simulation of co-channel talker interference suppression", ICASSP-89, pp. 361-364, 1989.

[2] Parsons T., "Separation of speech from interfering speech by means of harmonic selection", JASA, vol. 60, pp. 911-918, 1976.

[3] Naylor J. and Porter J.,"An effective speech separation system which requires no a-priori information",55.S14.15,ICASSP-91,1991

[4] Hanson B. and Wong D., "The Harmonic Magnitude Suppression (HMS) technique for intelligibility enhancement in the presence of interfering speech", ICASSP-84, 18A.5.1-4, 1984.

[5] Naylor J. and Boll S., "Techniques for suppression of an interfering talker in co-channel speech", ICASSP-87, 6.12.1-4, pp. 205-208, 1987.

[6] Hess W.,"Pitch determination of speech signals", Springer-Verlag, Berlin,1983

[7] Friedman D.H., "Pseudo maximum likelihood pitch extraction", ASSP-25, No. 3, pp. 213-221, June 1977.

[8] Wu C.F.J., "On the convergence properties of the EM algorithm",Ann. Statist., vol. 11, No. 1,95-103, 1983.

[9] Feder M. and Weinstein E., "Parameter estimation of superimposed signals", ASSP-36, No. 4, April 1988.

[10] Meidan Y., Yair E. and Chazan D., "Super resolution pitch determination of speech signals",IEEE Trans. on Signal Processing, Vol. 39, No. 1, January 91

Fig. 1 - Pitch error distribution

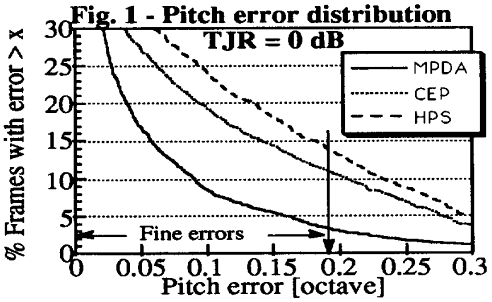

Fig. 2 - Pitch error distribution

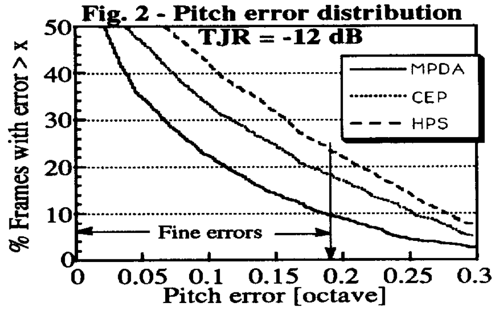

Table 1: TJR = 0 dB

PDA	% Gross errors	RMS error
MPDA	3.4	0.037
CEP	11.2	0.051
HPS	14.1	0.066

Table 2: TJR = -12 dB

PDA	% Gross errors	RMS error
MPDA	9.8	0.046
CEP	18.4	0.055
HPS	23.8	0.070

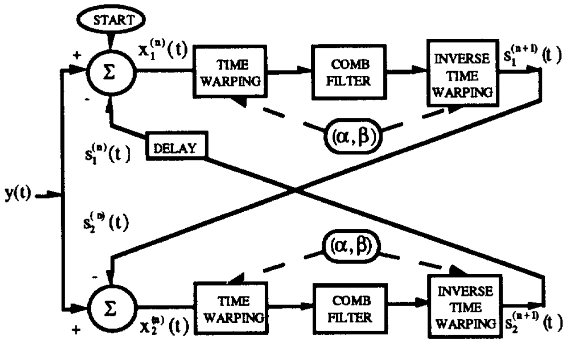

Fig. 3 - The MPDA EM based estimation scheme