# CO-CHANNEL SPEAKER SEPARATION

*David P. Morgan*, E. Bryan George, Leonard T. Lee*, and Stephen M. Kay†*

*Signal Processing Center of Technology, Lockheed Sanders Inc., Nashua, NH 03061-0868
†Department of Electrical Engineering, University of Rhode Island, Kingston, RI 02881

## ABSTRACT

This paper describes a system for the automatic separation of two-talker co-channel speech. This system is based on a frame-by-frame speaker separation algorithm that exploits a pitch estimate of the stronger talker derived from the co-channel signal. The concept underlying this approach is to recover the stronger talker's speech by enhancing harmonic frequencies and formants given a multi-resolution pitch estimate. The weaker talker's speech is obtained from the residual signal created when the harmonics and formants of the stronger talker are suppressed. A maximum likelihood speaker assignment algorithm is used to place the recovered frames from the *target* and *interfering* talkers in separate channels. The system has been tested at target-to-interferer ratios (TIRs) from -18 to 18 dB with human listening tests, and with machine-based tests employing a keyword spotting system on the *Switchboard Corpus* for target talkers at 6, 12, and 18 dB TIR.

## 1. INTRODUCTION

Co-channel speaker separation is employed when speech from two talkers has been summed into one signal and it is desirable to recover one or both of the speech signals from the composite signal.[1] Co-channel speech occurs in many common situations, such as when two AM signals containing speech are transmitted on the same frequency, or when two people are speaking simultaneously. The goal of co-channel speaker separation is to automatically process the co-channel signal and recover each talker's original speech. Minimizing artifacts in the processed speech is a key concern, especially if the recovered speech is passed to an automatic speech recognition system.

Efforts to develop algorithms addressing the co-channel speaker separation problem have spanned two decades [1], [2], [3], [4], and [5]. Initial work in co-channel speaker separation evolved from speech enhancement algorithms designed for separating voiced speech from background noise given a pitch estimate from the target talker
[1], [6]. Beginning with Hanson and Wong's *harmonic magnitude suppression* (HMS) technique [3], co-channel speaker separation algorithms have attempted to first estimate the pitch of at least one of the talkers, and then to exploit the pitch harmonics to separate the two talkers [2], [7].

This paper describes a *harmonic enhancement and suppression* (HES) based co-channel speaker separation system. This system uses a maximum likelihood (ML) pitch detector [8]. It provides an integer estimate of the pitch period of the stronger speech signal in a 40 ms analysis interval. Based on this initial integer estimate of the pitch period, a multi-resolution search is conducted to determine a fractional pitch period. The fractional pitch period is then used to construct two discrete-time filter pairs in the frequency domain. These filters are applied to the spectrum of the co-channel signal, $S_+(e^{j\omega})$, and used to separate the stronger and weaker talkers respectively. The recovered weaker signal is further processed by suppressing energy at frequencies corresponding to the 3 dB bandwidth of the stronger talker's formants. The recovered stronger and weaker signals $s_s[n]$ and $s_w[n]$ are then assigned to the target ($s_1[n]$) or interfering ($s_2[n]$) talker using a *maximum likelihood speaker assignment* (MLSA) algorithm, and resynthesized using overlap-add techniques.

## 2. APPROACH

The goal of this work was to develop a completely automatic co-channel speaker separation system which would operate without any *a priori* information, with a minimum delay in processing the signal, and would emphasize approaches with low computational complexity. Figure 1 is a block diagram of this system.

This system currently operates on 10 kHz sampled

---

E. Bryan George is now with the Systems and Information Science Laboratory, Texas Instruments, Inc., Dallas TX 75265

[1]This work was supported by the United States Air Force, Rome Labs, under Contract No. #F30602-91-C-0158.

0-7803-2431-5/95 $4.00 © 1995 IEEE

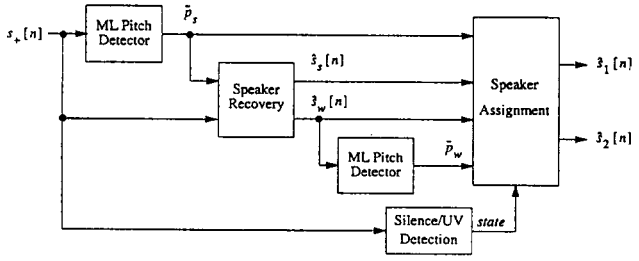


Figure 1: A block diagram of the HES co-channel speaker separation system.

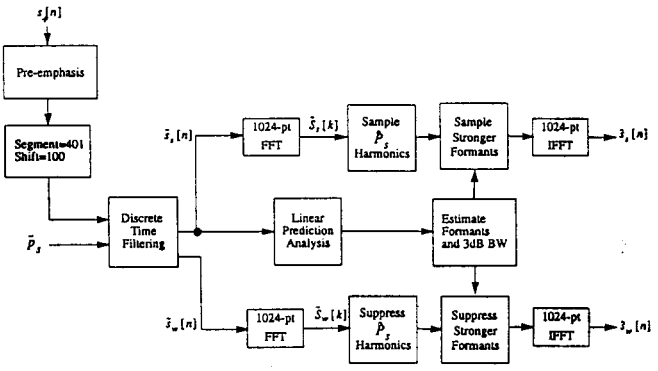


Figure 2: A block diagram of the HES Speaker recovery system.

data with a 401-point analysis window and a 10 ms frame rate. This window length was selected so that a pitch as low as 50 Hz could be identified. For each analysis frame, the co-channel signal $s_+[n]$ is analyzed to determine the pitch (fundamental frequency) of the stronger talker $\tilde{P}_s$. This pitch is used to drive the Speaker Recovery algorithm, which produces estimates of the stronger and weaker speech signals within that window. The recovered weaker signal is analyzed to determine the pitch of the weaker talker $\tilde{P}_w$. Silence/UV detection is then performed on this frame.

The MLSA algorithm examines these inputs, and evaluates the voicing probability resulting from hypothesized assignments in a dynamic programming fashion given the current pitch estimates and the recovered signals from four previous frames. This approach, which attempts to maximize the coherence of the reconstructed output signals, proved to be much more reliable for speaker assignment than attempting to track the recovered pitch. This algorithm produces two recovered signals, $\hat{s}_1[n]$ and $\hat{s}_2[n]$, one for the the target and the other for the interferer. For example, if $s_1$ is the target on the right channel (of a set of headphones), the MLSA algorithm will attempt to keep the target on this channel and any interfering talker on the left channel throughout the transmission.

One advantage of this approach is that it avoids the need to jointly estimate the pitch of both talkers. Estimating the pitch of the stronger talker is sufficient to achieve separation and make an estimate of the pitch of the weaker talker during subsequent processing. This approach also facilitates the recovery and tracking of both talkers as their respective signal strengths and voicing states change over time.

The Speaker Recovery algorithm operates on each analysis frame and attempts to recover the speech of both the stronger and weaker talkers using discrete time filter pairs based on the pitch estimate of the stronger talker. The strategy behind this algorithm is to recover the stronger talker by enhancing their formants and pitch harmonics. The recovered weaker talker is the residual signal obtained by suppressing the stronger talker's pitch harmonics and formants. Figure 2 is a block diagram which shows the steps in the Speaker Recovery algorithm.

One of the drawbacks of discrete-time filters is that they are extremely dependent upon an accurate estimate of $\tilde{p}_s$. We found that an integer estimate of the pitch period was inadequate, because the estimates of the pitch harmonics were off by more than one DFT bin at higher frequencies. Therefore, a high-resolution fractional pitch estimate was formulated to ensure that the location of higher-order harmonics was accurate to within one DFT bin. The fractional pitch period estimate is given by $\tilde{p}_s + i\tau$, and the transfer functions that were eventually implemented are:

$$\begin{aligned} H_+[z] &= (1+\alpha z^{-(\tilde{p}_s+i\tau)})/2, \\ H_-[z] &= (1-\alpha z^{-(\tilde{p}_s+i\tau)})(1-\alpha z^{(\tilde{p}_s+i\tau)}). \end{aligned} \quad (1)$$

The factor $\alpha = .99$ was introduced to avoid $H_+[z] = 0$ at $z = e^{jk2\pi\tilde{p}_s}$. The variable $\alpha$ was used in the same manner in $H_-[z]$ to dither the magnitude of the zeros at the pitch harmonics.

Our implementation uses the time-shift property of an $M$-point DFT in which

$$x(n-\tau) \overset{DFT}{\longleftrightarrow} X[k]e^{-\frac{j2\pi k\tau}{M}},$$

and $\tau$ is the time delay expressed in fractional samples. To use this property to delay a real signal by **non-integer** samples, it is important to correctly implement the wrap-around due to the periodicity of the DFT in the range $0 \le k \le M-1$. This can be accomplished for an even value of $M$ by multiplying $X[k]$ by

$$e^{-j\frac{2\pi}{M}k\tau} \quad \text{for} \quad 0 \le k \le \frac{M}{2}, \quad (2)$$

$$e^{-j\frac{2\pi}{M}(k-M)\tau} \quad \text{for} \quad \frac{M}{2}+1 \leq k \leq M-1.$$

In our high-resolution pitch estimate, the value of $\tau$ varies as a function of the estimated pitch period $\tilde{p}_s$. Given the true fundamental $\omega_0 = \frac{\pi}{h}$ for some $h$, we want to estimate $\omega_0$ such that the difference between the $h^{th}$ harmonics of $\omega_0$ and the estimate's fundamental $\tilde{\omega}_0$ is less than one-half bin of an $M$-point DFT:

$$\begin{aligned} \frac{-\pi}{M} &\leq h(\tilde{\omega}_0 - \omega_0) \leq \frac{\pi}{M} \\ \frac{-\pi}{M} &\leq \frac{\pi}{\omega_0}(\tilde{\omega}_0 - \omega_0) \leq \frac{\pi}{M} \\ 1 - \frac{1}{M} &\leq \frac{\tilde{\omega}_0}{\omega_0} \leq 1 + \frac{1}{M}. \end{aligned} \tag{3}$$

Letting $\eta = 1 - \frac{1}{M}$ and $\mu = 1 + \frac{1}{M}$ implies

$$log\eta \leq log\tilde{\omega}_0 - log\omega_0 \leq log\mu. \tag{4}$$

To ensure this accuracy when searching in the neighborhood of the stronger talker's pitch derived from the **integer** pitch estimate, $\tilde{P}_s$, it is necessary to specify a candidate grid $i$ so that the optimal **fractional** pitch period, $\hat{p}_s = \tilde{p}_s + i\tau$ can be determined. For a given $\tilde{P}_s$, $i$ is determined from

$$log\hat{P}_s = log\tilde{P}_s + i(log\mu - log\eta). \tag{5}$$

For large values of $M$, $log\eta \approx -\frac{1}{M}$ and $log\mu \approx \frac{1}{M}$:

$$\begin{aligned} log\hat{P}_s &\approx log\tilde{P}_s + \frac{2i}{M} \\ \hat{P}_s &\approx e^{\frac{2i}{M}}\tilde{P}_s \\ \frac{2\pi}{\hat{P}_s} &= \hat{p}_s \approx e^{-\frac{2i}{M}}\tilde{p}_s \end{aligned} \tag{6}$$

For small values of $\frac{2i}{M}$ (i.e., $< .01$), $e^{-\frac{2i}{M}} \approx (1 - \frac{2i}{M})$:

$$\hat{p}_s \approx (1 - \frac{2i}{M})\tilde{p}_s \quad \text{for } -R \leq i \leq R. \tag{7}$$

For $\tau = \frac{2\tilde{p}_s}{M}$, this is equivalent to $\hat{p}_s = \tilde{p}_s + i\tau$. Thus our multi-resolution estimate has $\tau$ proportional to $\tilde{p}_s$, and in our experiments we have selected $R = 4$.

The optimal fractional pitch period is determined by minimizing the energy passed by a zero-phase discrete-time notch filter $N[z]$. This notch filter is a function of $i\tau$ and extends one pitch period to the left and right of the current pitch period as follows:

$$N[z] = (1 - \alpha z^{-(\tilde{p}_s + i\tau)})(1 - \alpha z^{(\tilde{p}_s + i\tau)}) = H_-[z]. \tag{8}$$

The notch filter is evaluated at $z = e^{\frac{j2\pi k}{M}}$ to determine $N[k]$ and is multiplied by $S_m[k]$ in the frequency domain. $S_m[k]$ is the Fourier transform of the current N-point window starting at time sample $m$, which includes data points from a previous and past pitch period. $S_m[k]$ is defined as the M-point FFT of the signal:

$$\begin{aligned} s_m[n] = s_+[m], ...s_+[m + N + \lceil\tilde{p}_s + i\tau\rceil], 0, 0, 0, \\ ..., s_+[m - \lceil\tilde{p}_s + i\tau\rceil], ..., s_+[m-1]. \end{aligned}$$

The last $\lceil\tilde{p}_s + i\tau\rceil$ points of the M-point FFT contain samples from the previous pitch period.

The product $S_m[k]N[k]$ is computed for $-R \leq i \leq R$ as:

$$S_i[k] = S_m[k](1 - \alpha e^{-j\frac{2\pi}{M}k(\tilde{p}_s + i\tau)})(1 - \alpha e^{j\frac{2\pi}{M}k(\tilde{p}_s + i\tau)}) \tag{9}$$

in which the computation of $S_i[k]$ adheres to the wrap-around rule given in Equation 2 for $0 \leq k \leq M$. $S_i[k]$ can then be inverse transformed to obtain $s_i[k]$, and the norm of the first $N = 401$ points can be computed. The value of $i$ that minimizes the norm is selected.

## 3. TESTING

The HES algorithm was subjected to several types of tests: keyword spotting (KWS) tests on linearly added speech signals, listening tests on linearly added speech signals, and listening tests using speech transmitted over RF channels. For the RF experiments, transmitters were connected to antennas on top of our building which were transmitting AM speech signals on the same frequency to a receiving antenna approximately 50 meters away. Only the results on the keyword spotting tests are presented here; see [9] for more extensive test results.

Three standard speech databases were used for testing the HES algorithm: the Switchboard Credit Card corpus, the MIT-CBG database [10], and the TIMIT database. The MIT-CBG and TIMIT databases are two "read speech" databases which were selected because they are phonetically labeled, enabling us to determine performance as a function of three voicing states: voiced, unvoiced, and silence. The MIT-CBG database contains 3 male talkers and 630 total sentences (210 per talker). Of these 210 sentences per talker, 110 are phonetically transcribed.

Thirty-five telephone conversations were evaluated from the Credit Card corpus. Each conversation contains two talkers recorded separately (four wire recordings). The first 20 conversations were selected for training and the remaining 15 for testing, resulting in a total of 160.6 minutes of testing data. A KWS vocabulary consisting of seven multi-syllable keywords/keyphrases and their variants was selected from this Corpus. Speaker-independent tests were conducted using four female templates and four male templates per keyword.

An internal (and subsequently outdated) KWS system was used to produce objective co-channel results for linearly added co-channel and co-channel processed speech. This KWS system uses a feature representation based on a vector-quantized representation of 20 filterbank energies computed over a 300-3785 Hz bandwidth. A dynamic time warping (DTW) algorithm was used to compare each template to the input utterance.

The MLSA algorithm was not used in these tests. Instead, the assignment was based on pitch estimates computed prior to linearly adding the signals. Therefore, this test measured the performance of the Speaker Recovery algorithm, and relieved us from the burden of spotting against both recovered channels.

Random utterances from the three male talkers from the MIT-CBG database were used as the interfering speech. Tests were conducted to determine the performance of the KWS system at decreasing TIRs. Table 1 presents baseline results against the database with no interference, as well as both unprocessed and processed speech at 18, 12 and 6 dB TIRs (i.e., the interfering talker was the weaker of the two signals). Results are presented in terms of a *figure of merit* (FOM) for each keyword.

| | | Unprocessed - TIR | | | Processed - TIR | | |
|---|---|---|---|---|---|---|---|
| keyword | baseline | 18 dB | 12 dB | 6 dB | 18 dB | 12 dB | 6 dB |
| account | 11.56 | 14.22 | 1.78 | 0.0 | 7.11 | 11.11 | 0.0 |
| american_express | 18.33 | 19.33 | 12.67 | 4.83 | 22.00 | 13.33 | 18.67 |
| balance | 11.64 | 0.0 | 0.0 | 0.0 | 11.64 | 5.09 | 0.0 |
| credit_card | 21.32 | 9.50 | 10.94 | 3.71 | 16.29 | 16.74 | 3.12 |
| discover | 22.80 | 0.0 | 0.0 | 0.0 | 11.60 | 3.20 | 12.80 |
| dollar | 16.00 | 7.10 | 5.81 | 0.90 | 21.94 | 9.81 | 2.45 |
| visa | 10.71 | 7.14 | 2.71 | 0.14 | 3.29 | 1.00 | 3.71 |
| average FOM | 18.34 | 8.18 | 4.84 | 1.37 | 13.41 | 8.61 | 5.82 |

Table 1: KWS results for seven multi-syllable keywords in the Credit Card Corpus given in terms of figure of merit.

## 4. SUMMARY

The HES co-channel speaker separation system has been demonstrated to improve performance for a machine-based keyword spotting system at -6 and -18 dB TIR for linearly added co-channel speech. The current MLSA speaker assignment algorithm (which due to space constraints, was not fully described) requires additional work so that it can reliably keep a talker on a channel without swapping. This problem has been addressed by building a user interface for our system which displays the original and recovered speech signals, the ML pitch estimate, and the swap decisions, and which allows the user to change any swap decision and resynthesize the signal. In the future, a decision aid based upon speaker identification technology may also prove useful.

The HES algorithm has some inherent problems; when both talker's have the same instantaneous pitch, the algorithm will place both talkers on one channel and neither talker on the other channel. When there are more than two talkers in the co-channel signal, only the stronger talker can be separated, and the separation is predicated on that talker always being stronger and voiced. Although considerable work remains in developing co-channel algorithms, especially for the radio-mixed environment, this paper has presented the first complete and methodical evaluation of a co-channel speaker separation system.

## 5. REFERENCES


[1] V.C. Shields, *Separation of Added Speech Signals by Digital Comb Filtering.* S.M. Thesis, Dept. of Electrical Engineering, MIT, 1970.

[2] T.W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust. Soc. Am.*, vol. 60 (4), 1976, pp. 911–918.

[3] B.A. Hanson and D.Y. Wong, "The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence of interfering speech," in *Proc. ICASSP*, San Diego, CA, March 1984, pp. 18A.5.1–18A.5.4.

[4] J.A. Naylor and J. Porter, "An effective speech separation system which requires no a priori information," in *Proc. ICASSP*, Toronto, Canada, May 1991, pp. 937–940.

[5] M.A. Zissman and D.C. Seward IV, "Two-talker pitch tracking for co-channel interference suppression," MIT Lincoln Labs Technical Report 951, April 1992.

[6] R.H. Frazier, *An Adaptive Filtering Approach Toward Speech Enhancement.* S.M. Thesis, Dept. of Electrical Engineering, MIT, 1975.

[7] J.A. Naylor and S.F. Boll, "Techniques for suppression of an interfering talker in co-channel speech," in *Proc. ICASSP*, Dallas TX, April 1987, pp. 205–208.

[8] J.D. Wise, J.R. Caprio and T.W. Parks, "Maximum likelihood pitch estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24 (5), October 1976, pp. 418–423.

[9] D.P. Morgan, E.B. George, L.T. Lee and S.M. Kay, "Co-Channel Speech Improvements," Rome Labs Technical Report RL-TR-93, Griffis Air Force Base, NY, 13441.

[10] M.A. Picheny, N.I. Durlach and L.D. Braida, "Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech," *Journal of Speech and Hearing Research*, vol. 28, March 1985, pp. 96–103.