ESD-TR-91-251

Technical Report
951

Two-Talker Pitch Tracking for Co-channel Talker Interference Suppression

M.A. Zissman
D.C. Seward IV

30 April 1992

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LEXINGTON, MASSACHUSETTS

Prepared for the Department of the Air Force under Contract F19628-90-C-0002.

Approved for public release; distribution is unlimited.

This report is based on studies performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology. The work was sponsored by the Rome Laboratory, Department of the Air Force, under Contract F19628-90-C-0002.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

Hugh L. Southall

Hugh L. Southall, Lt. Col., USAF
Chief, ESD Lincoln Laboratory Project Office

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LINCOLN LABORATORY

TWO-TALKER PITCH TRACKING FOR CO-CHANNEL TALKER INTERFERENCE SUPPRESSION

M.A. ZISSMAN
D.C. SEWARD IV
Group 24

TECHNICAL REPORT 951

30 APRIL 1992

Approved for public release; distribuiton is unlimited.

LEXINGTON MASSACHUSETTS

ABSTRACT

Almost all co-channel talker interference suppression systems use the difference in the pitches of the target and jammer speakers to suppress the jammer and enhance the target. While joint pitch estimators outputting two pitch estimates as a function of time have been proposed, the task of proper assignment of pitch to speaker (two-talker pitch tracking) has proven difficult. This report describes several approaches to the two-talker pitch tracking problem including algorithms for pitch track interpolation, spectral envelope tracking, and spectral envelope classification. When evaluated on an all-voiced two-talker database, the best of these new tracking systems correctly assigned pitch 87% of the time given perfect joint pitch estimation.

ACKNOWLEDGMENTS

The authors wish to thank Bob McAulay and Tom Quatieri for useful discussions, Linda Kukolich for adapting the LNKnet software to meet our needs, Derek Graham for resurrecting Ron Danisewicz's software, and Jim Cupples and Carolyn Farnsworth of Rome Laboratory for helpful comments and suggestions.

TABLE OF CONTENTS

Abstract iii

Acknowledgments v

List of Illustrations ix

List of Tables xi

1. INTRODUCTION 1

2. PREVIOUS WORK 3

3. NEW APPROACHES 5

3.1 Identifying Runs 5

3.2 Pitch Track Interpolation 6

3.3 Spectral Envelope Tracking 11

3.4 Spectral Envelope Classification 14

4. EXPERIMENTS 19

4.1 Implementation 19

4.2 Database 19

4.3 Results 20

4.4 Discussion 23

5. CONCLUSIONS AND FUTURE WORK 27

REFERENCES 29

LIST OF ILLUSTRATIONS

Figure No.		Page
1	A block diagram that shows how the joint pitch estimation and joint pitch tracking modules fit into a generic CCTIS system.	2
2	The two-talker pitch tracking problem can be reduced to connecting adjacent (in time) runs.	5
3	An example run assignment for a large and a small value of p_t.	7
4	An example of pitch track interpolation with polynomial interpolation.	8
5	A means of resolving possible pitch track crosses.	11
6	The Quatieri/Danisewicz system.	13
7	A block diagram of the spectral envelope classification system.	16
8	Run analysis.	21
9	Pitch tracking performance.	24
10	Ability of the spectral envelope classification pitch tracker to classify correctly frames in runs, ignoring those not in runs.	25

LIST OF TABLES

Table No.		Page
1	Pitch Track Interpolation Results	22
2	Spectral Envelope Tracking Results	23
3	Spectral Envelope Classification Parameters	24

1. INTRODUCTION

Co-channel talker interference suppression (CCTIS) is defined as the enhancement of an input signal containing regions of simultaneous target and jammer speech, thereby producing an estimate of the isolated, single-speaker target speech signal. For example, CCTIS might be applied to speech received over a radio channel having two transmitters in the same band. Numerous digital signal processing techniques [1–14] have been applied to this problem over the past two decades. A review of many of these systems can be found in Zissman [15]. Each of the existing separation techniques uses the difference in pitch between the target and jammer as the primary means of separating the two speakers, requiring

- A pitch track from the target, or
- A pitch track from the jammer, or
- Pitch tracks from both the target and jammer,

where a pitch track is defined as an estimate of the pitch of the utterance as a function of time. (In this report, the term "pitch," which properly refers to a subjective, psychoacoustic parameter, is used in place of "fundamental frequency f_0," which is a measurable, physical parameter.) Realizable systems must have a means of estimating the pitch track(s) from the co-channel speech. This is a challenging problem when only one pitch track is required. In such systems a standard one-talker pitch estimation algorithm is modified to take as input co-channel speech; however, the pitch estimation and tracking problem is particularly difficult in the case of systems that require both target and jammer pitch tracks. Such systems must not only generate two pitch estimates for each frame (a time interval, typically from 5 to 50 ms, over which speech analysis is performed on what is assumed to be a quasi-stationary signal), but they must also assign these two pitch estimates to the target and the jammer properly for each frame. During each frame having misassigned pitch (the pitch assigned to the target really belongs to the jammer and vice versa), the reconstructed target speech would be expected to sound more like the jammer than the target. The problem of generating two pitch estimates for each frame is known as joint pitch estimation. The task of assigning the pitch estimates properly to the target and the jammer is known as joint pitch tracking. Figure 1 is a block diagram that shows how the joint pitch estimation and the joint pitch tracking modules fit into a generic CCTIS system.

There are some cases where joint pitch tracking is not required, for instance when it is known a priori that one talker always has a higher pitch than the other, as in the case of a male target and female jammer. Additionally, joint pitch tracking is not required for speaker separation systems needing only one of the two pitch tracks; however, for systems that require both target and jammer pitch tracks (see Parsons [5], Quatieri and Danisewicz [16], and Naylor and Porter [13]), joint pitch tracking is an absolute requirement for practical implementation. The fact that reliable joint pitch tracking has proven so difficult has prevented practical application of completely automated versions of these systems. The purpose of this report is to discuss the effectiveness of some new approaches to the joint pitch tracking problem.

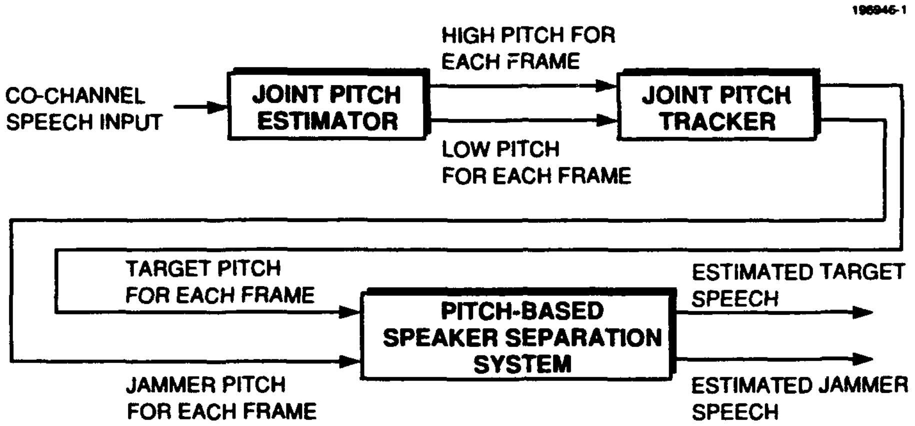


Figure 1. A joint pitch estimator is a system that takes (as input) co-channel speech and produces (as output) two pitch values as a function of time. A joint pitch tracker takes (as input) two pitch values as a function of time (the output of the joint pitch estimator) and produces a pitch track for the target and for the jammer.

To uncouple the task of developing a joint pitch tracker from developing a joint pitch estimator, the two pitch tracks that are output by a single-talker pitch estimation system and run *separately* on the target and jammer waveforms can be input to the joint pitch tracker. Performance of the pitch tracker given these near-perfect pitch values obtained with a single-talker pitch estimator sets an upper bound on the performance expected using the imperfect pitch estimates produced by any realizable joint pitch estimator operating only on the co-channel speech. This simplification has allowed development of a pitch tracker while sidestepping the equally important and difficult problem of obtaining accurate pitch estimates. For final evaluation, however, the output of a joint pitch estimator operating on co-channel speech has been used as input to the best of the joint pitch trackers described in this report.

Some of the two-talker pitch tracking approaches that were used prior to this report are reviewed in Chapter 2, while the three new algorithms that have been applied to pitch tracking (pitch track interpolation, spectral envelope tracking, and spectral envelope classification) are outlined in Chapter 3. Databases, experimental procedures, results, and discussion are provided in Chapter 4, and concluding remarks and suggestions for further work are made in Chapter 5.

2. PREVIOUS WORK

This section reviews the joint pitch tracking systems suggested by Parsons [5,17]. Naylor and Porter [13] have also addressed the pitch tracking problem, but their system works only when there is a significant difference in energy levels of the target and jammer. In these cases they expect to see significant differences in the heights of the spectral peaks due to the target and jammer talkers. In cases where the levels are equal, 0 dB target-to-jammer ratio (TJR), they apparently rely on intermittent human intervention [18]. The reader is reminded that many other co-channel systems use a prioriinformation to track pitch and are, therefore, unrealizable in practice. For example, the Stubbs/Summerfield [12] and the Quatieri/Danisewicz [16] systems must be told which talker, target or jammer, has the higher pitch in each frame.

Parsons proposed a two-step pitch tracking system: (1) on eacı. frame, assign the new pitch values to their proper tracks by matching them to previously predicted values and then (2) predict two pitch values for the next frame. During step (1), with two predictions and two pitches, there are four distances to be considered. The Parsons matching procedure computes these distances and decides either to assign pitch 1 to the target and pitch 2 to the jammer or vice versa. Although he claims that this would be a trivial problem in regions where both talkers are active (not silent), results are shown in Chapter 4 that suggest that this is a difficult problem even in the case of always active, all-voiced targets and jammers. During prediction, step (2), Parsons uses a three-point linear extrapolator to predict the next pitch value. Tracker performance is not reported either objectively in terms of number of frames correctly tracked or subjectively in terms of intelligibility of speech separated by a system using the pitch tracker.

In an effort to improve the prediction step, Parsons also tried linear prediction (LP) as opposed to extrapolation [17]. In this system the value of the pitch at frame n was estimated as a linear combination of the k pitch values for the k preceding frames according to

$$p(n) = \sum_{i=1}^{k} a_i p(n-i), \tag{1}$$

where $p(n)$ is the pitch at frame n, and the a_i are fixed, talker-independent coefficients computed in advance. By processing a large amount of single talker input speech from different talkers, Parsons generated sets of coefficients a_i for various values of k. The pitch tracker had an acquisition and a tracking mode. The order of the LP, k, varied as a function of mode and distance since the last mode switch, but in no case was k ever greater than 5. Parsons did observe occasional inversions of pitch tracks, which could only be corrected by user intervention. This pitch tracking system was never evaluated formally.

3. NEW APPROACHES

Each novel solution to the two-speaker pitch tracking problem begins by identifying regions of time called runs, during which the difference between the estimated pitches is consistently above a predetermined threshold. This section discusses run identification in detail and describes a method to connect them by pitch track interpolation, extending the system suggested originally by Parsons. The notion of using spectral envelope information to aid in the run connection problem is introduced, and by creating spectral envelope libraries from single-talker target and jammer training speech, the pitch tracking problem can be converted to a template matching problem.

3.1 Identifying Runs

By loosely defining a run as a contiguous region of time during which it is hypothesized that no pitch track crossings have occurred, the two-talker pitch tracking problem can be reduced to connecting adjacent (in time) runs, as shown in Figure 2. More precisely, a run is the longest

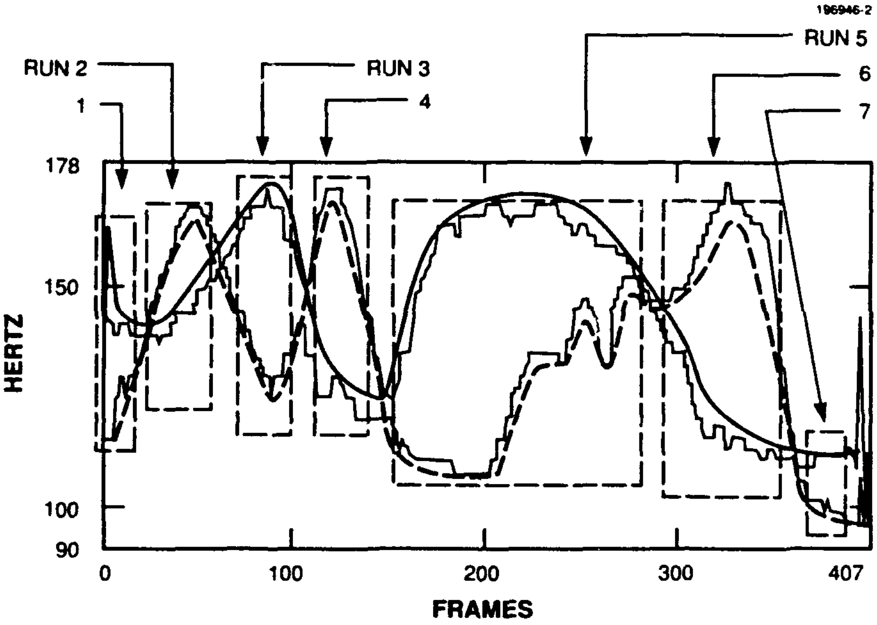


Figure 2. Two pitch tracks for a co-channel sentence pair look somewhat jagged and piecewise constant. Seven runs are enclosed in the dashed boxes. The heavy solid and heavy dashed curves show how the pitch track segments in each run should be connected to the segments in adjacent runs to form the true target and jammer pitch tracks.

possible contiguous region of time, having the property that the two pitch values for all frames in the run are separated by more than some pitch separation threshold. For most values of pitch separation thresholds, one would expect that a co-channel utterance comprising a male and a female would have a single run. On the other hand, for two male talkers of roughly comparable average pitch, one would expect multiple runs in the utterance. In this report the term "pitch separation threshold" p_t will be used only as the criterion by which runs are defined, that is

p_t = pitch separation threshold for runs = minimum pitch difference in frame $i+1$, allowing continuation of a run existing in frame i to $i+1$.

The following algorithm identifies the runs within a co-channel utterance. For each frame t the run analysis algorithm takes as input a pair of pitch values $p_h(t)$ and $p_l(t)$, of which $p_h(t)$ is the higher frequency of the two, and $p_l(t)$ is the lower. The first frame having $p_h(t) - p_l(t) > p_t$ is declared the first frame of the first run. All contiguous subsequent frames having $p_h(t) - p_l(t) > p_t$ are also part of this first run. As soon as a frame is located for which $p_h(t) - p_l(t) <= p_t$, the first run is terminated, and the search begins for the first frame of the second run, and so on. Eventually, the entire utterance is segmented into runs with at least one frame (and usually more) separating each pair of adjacent runs. Because runs are separated by regions where the difference between target and jammer pitch is less than p_t, not every frame in a co-channel utterance belongs to a run. Large values of p_t tend to exclude more frames from membership in runs and tend to ensure that each run is, in fact, free of pitch crossings. Small values of p_t tend to include a greater percentage of frames in runs and greater average run lengths; however, the chance of a cross occurring within a run increases. One should choose a value of p_t large enough to ensure that runs are really cross free but small enough to ensure that as many frames as possible are included. Figure 3 shows an example run assignment for a large and a small value of p_t.

Associated with each run are two pitch track segments, one for the high pitch speaker and one for the low. The task of pitch tracking has now been reduced to determining, for each run, whether the high-pitch track belongs to the target and the low-pitch track belongs to the jammer (the so-called target-higher assignment rule), or whether the high-pitch track belongs to the jammer and the low-pitch track belongs to the target (the so-called jammer-higher assignment rule). The remainder of this section explores three different methods for hypothesizing assignment rules for each run in the utterance.

3.2 Pitch Track Interpolation

In the pitch track interpolation approach to two-talker pitch tracking, two pairs of possible pitch trajectories are interpolated for each pair of adjacent runs: one pair of trajectories for the hypothesis that no cross occurred and one pair of trajectories for the hypothesis that a cross did occur. The pair of interpolated trajectories that best fit the data (in a mean square error sense) are selected, and the corresponding hypothesis is deemed correct. An example of the pitch track interpolation with polynomial interpolation is shown in Figure 4.

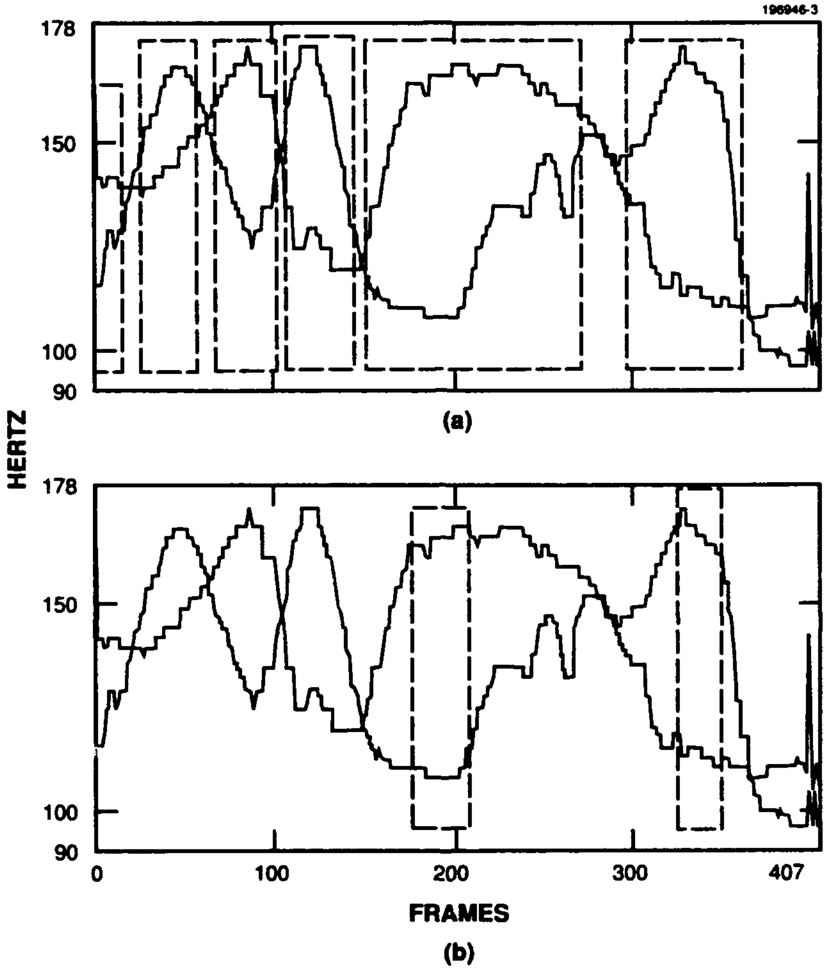


Figure 3. Run locations for (a) p_t = 7.3 Hz and (b) p_t = 48.8 Hz. Note that p_t is defined as the minimum pitch difference in frame i + 1, allowing continuation of a run existing in frame i to frame i + 1. Runs are enclosed in dashed boxes.

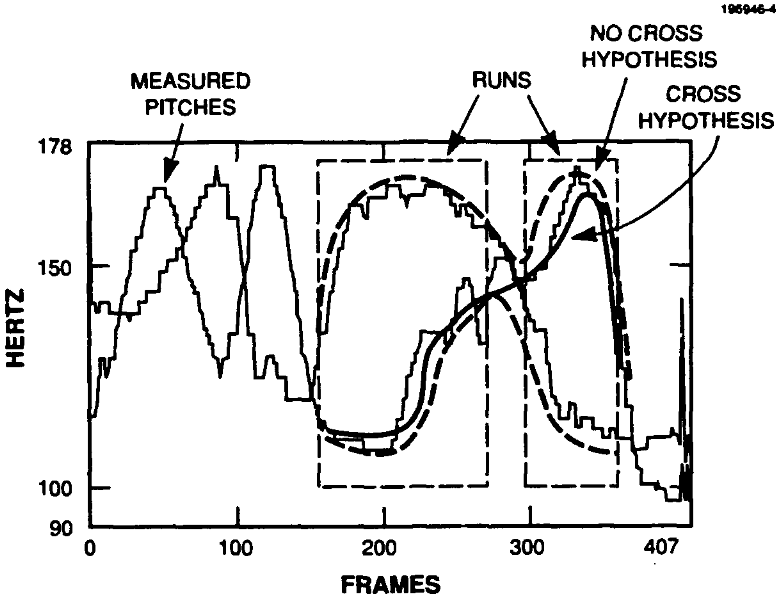


Figure 4. The piecewise constant lines show the measured pitches. The dashed boxes show two adjacent runs. The heavy solid lines show the pitch tracks under the cross hypothesis. The heavy dashed lines show the pitch tracks under the no cross hypothesis.

3.2.1 The Algorithm

The algorithm takes as input the output of run analysis (a list of runs) where each run i has associated with it a high-pitch value $p_h(t)$ and a low-pitch value $p_l(t)$ for all frames $t_b^i <= t <= t_e^i$, where t_b^i is the first frame of run i and t_e^i is the final frame. For each adjacent pair of runs i and $i+1$, the system uses the method of linear least square estimation to fit two pairs of polynomials [19]:

- **Straight.** Polynomial $q_{h,h}^{i,i+1}(t)$ is fit to $p_h(t)$ for the union of frames $t_b^i <= t <= t_e^i$ and $t_b^{i+1} <= t <= t_e^{i+1}$, and polynomial $q_{l,l}^{i,i+1}(t)$ is fit to $p_l(t)$ for the the same union of frames.
- **Cross.** Polynomial $q_{h,l}^{i,i+1}(t)$ is fit to $p_h(t)$ for $t_b^i <= t <= t_e^i$ and $p_l(t)$ for $t_b^{i+1} <= t <= t_e^{i+1}$, and polynomial $q_{l,h}^{i,i+1}(t)$ is fit to $p_l(t)$ for $t_b^i <= t <= t_e^i$ $p_h(t)$ for $t_b^{i+1} <= t <= t_e^{i+1}$.

Then, for all frames in the two runs, the straight error $E_s^{i,i+1}$ and the cross error $E_c^{i,i+1}$ are calculated as follows:

$$\begin{aligned}
E_s^{i,i+1} &= \sum_{t=t_b^i}^{t_e^i} ((q_{h,h}^{i,i+1}(t) - p_h(t))^2 + (q_{l,l}^{i,i+1}(t) - p_l(t))^2) + \\
&\quad \sum_{t=t_b^{i+1}}^{t_e^{i+1}} ((q_{h,h}^{i,i+1}(t) - p_h(t))^2 + (q_{l,l}^{i,i+1}(t) - p_l(t))^2) \\
E_c^{i,i+1} &= \sum_{t=t_b^i}^{t_e^i} ((q_{h,l}^{i,i+1}(t) - p_h(t))^2 + (q_{l,h}^{i,i+1}(t) - p_l(t))^2) + \\
&\quad \sum_{t=t_b^{i+1}}^{t_e^{i+1}} ((q_{l,h}^{i,i+1}(t) - p_h(t))^2 + (q_{h,l}^{i,i+1}(t) - p_l(t))^2). \qquad (2)
\end{aligned}$$

If $E_c^{i,i+1} < E_s^{i,i+1}$, then a pitch track crossing is hypothesized between runs i and $i+1$. Otherwise it is hypothesized that no pitch track crossing occurred between runs i and $i+1$.

If it is known a priori which speaker (target or jammer) has the high-pitch track in run 0, then the hypotheses of where crosses occur can be used to assign recursively the labels "target" or "jammer" to each pitch track in each run of the utterance. If no a priori information is available, only labels such as "speaker-1" and "speaker-2" can be assigned to the pitch tracks; hence, the association of speaker-1 and speaker-2 with target and jammer must be resolved by the user (see Section 3.2.4).

3.2.2 Handling Frames Between Runs

Until now the data offered by the frames between runs have been ignored. So too, no indication of how pitch values should be assigned during interrun frames has been offered. Perhaps the simplest approach is to assign the pitches in these frames arbitrarily. Generally, one would expect that half the frames would have correctly classified pitch in this case.

Alternatively, for each pair of runs i and $i+1$, if cross was hypothesized, then the time of intersection $t_x^{i,i+1}$, where $q_{h,l}^{i,i+1}(t) = q_{l,h}^{i,i+1}(t)$ could be located. The frames in the interval $t_e^i < t < t_x^{i,i+1}$ would follow the assignment rule of run i, while the frames in the interval $t_x^{i,i+1} <= t < t_b^{i+1}$ would follow the assignment rule of run $i+1$. On the other hand, if straight had been hypothesized, then there is no cross, so the assignment rules of run i and $i+1$ are identical and could be imposed on the intervening region as well. This method also suggests modified error criteria $\hat{E}_s^{i,i+1}$ and

$\hat{E}_c^{i,i+1}$:

$$\begin{aligned}
\hat{E}_s^{i,i+1} &= E_s^{i,i+1} + \sum_{t=t_e^i}^{t_b^{i+1}} ((q_{h,h}^{i,i+1}(t) - p_h(t))^2 + (q_{l,l}^{i,i+1}(t) - p_l(t))^2) \\
\hat{E}_c^{i,i+1} &= E_c^{i,i+1} + \sum_{t=t_e^i}^{t_x^{i,i+1}} ((q_{h,l}^{i,i+1}(t) - p_h(t))^2 + (q_{l,h}^{i,i+1}(t) - p_l(t))^2) + \\
&\sum_{t=t_x^{i,i+1}}^{t_b^{i+1}} ((q_{l,h}^{i,i+1}(t) - p_h(t))^2 + (q_{h,l}^{i,i+1}(t) - p_l(t))^2) \qquad (3)
\end{aligned}$$

to replace the error criteria defined in Equation (2). These new criteria use all the pitch data available in the interval $t_b^i <= t <= t_e^{i+1}$, which could be an advantage, depending on the relative accuracy of the interrun versus intrarun pitch data.

3.2.3 Comparison to the Parsons System

Pitch track interpolation is similar to the Parsons approaches in that pitch data from previous frames are used to predict pitch values for the current frame; however, the method described here interpolates, using both past and future information to predict current pitch values, whereas the Parsons system is strictly causal, using only past values to predict the current value. Also the Parsons system does not use the two-stage method of demarking and then connecting runs. Finally, a Parsons prediction is either linear extrapolation (best fit to the previous three points) or LP, using fixed coefficients. The tracking method introduced here uses polynomial interpolation in an effort to better fit the pitch tracks.

3.2.4 Problems

One problem with the pitch track interpolation method of joint pitch tracking is that even if the system performs perfectly, it has no notion of which pitch track belongs to the target and which belongs to the jammer. This lack of knowledge is probably not an issue if the separation system is to be used with significant user intervention, but it prevents its use in a truly automatic mode.

A second important drawback of the pitch track interpolation approach to two-talker pitch tracking is the fact that a single tracking error can cause all subsequent tracking to fail. Consider a co-channel utterance with 10 true runs roughly equally spaced through the utterance. Assume that the system is told that the target has the higher pitch and the jammer has the lower pitch in run 1. If the tracker were 90% effective at connecting runs and if the first 9 were correctly connected, but runs 9 and 10 were misconnected, then target and jammer pitch assignment would be correct for about 90% of the frames. If, however, runs 1 and 2 were misconnected with all subsequent runs properly connected, then 90% of the frames would have misidentified target and jammer pitches. Thus, because assignment of run i is dependent on the assignment of all runs prior to i, the system is not robust against a few tracking errors.

3.3 Spectral Envelope Tracking

A second approach to joint pitch tracking exploits the ability of theQuatieri/Danisewicz (QD) speaker separation system [16] to take as input two pitch estimates for a given frame (without assignment of pitch to speaker) with a frame of co-channel speech and produce a pair of estimated single-speaker harmonic magnitude spectra. Of course without knowing which pitch belongs to which speaker, the system cannot assign the magnitude spectra to the target or the jammer; however, using the pair of harmonic magnitude spectra for each frame to estimate a pair of spectral envelopes, the notion of comparing them from the last frame of one run with those from the first frame of the next as a means of resolving possible pitch track crosses (see Figure 5) has been studied. If a cross did indeed occur, it was expected that the envelope of the high-pitch speaker at the end of the first run would match the envelope of the low-pitch speaker at the beginning of the next run and vice versa. A detailed description of the algorithm and a discussion of its merits and drawbacks follows.

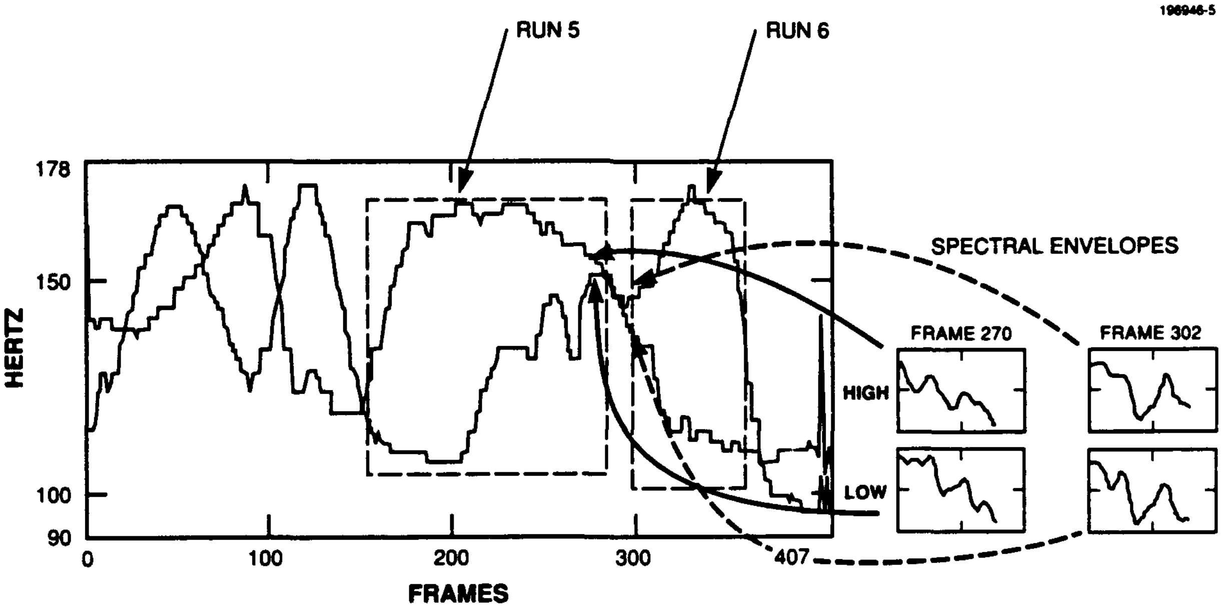


Figure 5. Spectral envelope estimates are shown for each speaker for the last frame of the first run and the first frame of the subsequent run. For this example the data suggest a cross has occurred, as the envelope of the high-pitch speaker at the end of run 5 is similar to the envelope of the low-pitch speaker at the beginning of run 6 and vice versa.

3.3.1 Algorithm

The spectral envelope tracking method of joint pitch tracking exploits the ability of the QD CCTIS system to take as input two pitch values with a frame of co-channel speech and produce two estimated harmonic spectra (one corresponding to each pitch value). For each speaker the QD system outputs a set of harmonically related sinewave frequencies, magnitudes, and phases. The sinewave frequencies are merely the harmonics of the input pitch. The magnitudes and phases are calculated using the method of linear least squares estimation, which finds the two sets of magnitudes and phases (one set each for the high- and the low-pitch speaker, each at harmonics of the appropriate pitch frequency) that best fit the co-channel input speech. For each speaker the harmonic magnitude line spectrum is converted to an envelope magnitude spectrum according to the following procedure [20]:

1. The envelope function $env(f)$ is generated as a piecewise constant interpolation of the harmonic line frequency spectrum
$$env(f) = mag(s_{f_i}), \text{ for } f_i - f_0/2 <= f <= f_i + f_0/2, \tag{4}$$
where f_0 is the estimated pitch for the speaker, $f_i = i * f_0$ is the ith harmonic of the pitch frequency, $mag(s_{f_i})$ is the magnitude of the sinewave having frequency f_i as output by the DQ system.

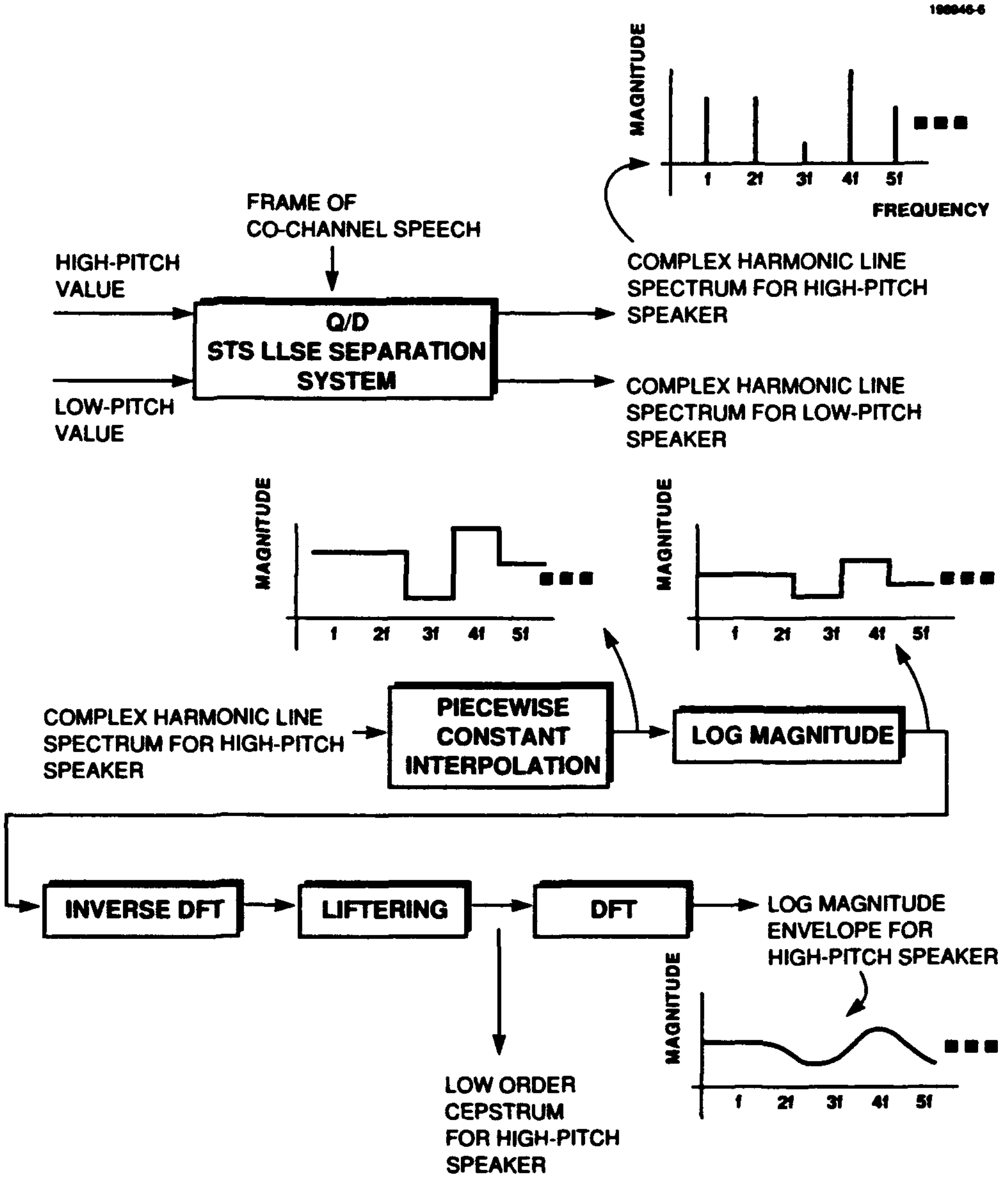
2. An inverse discrete Fourier transform (IDFT) is performed on the $log(env(f))$. The resulting cepstrum is "liftered" by zeroing its high-order coefficients, which has the effect of smoothing the spectrum. (According to the standard speech production model of a pulse train at the pitch frequency exciting a time-varying linear filter, this low pass liftering has the effect of removing the speech excitation information, leaving an estimate of the filter transfer function, which is an estimate of the vocal tract transfer function.)

3. Optionally, a discrete Fourier transform (DFT) is applied to transform the cepstrum back to a log magnitude.

These steps are summarized in Figure 6.

Given either the liftered cepstra or the corresponding envelope spectra, joint pitch tracking is reduced to comparing the two final cepstra (or spectra) of run i (one for the higher pitch speaker and one for the lower pitch speaker) with the first cepstra (or spectra) of run $i+1$.[1] For the cepstral case, defining $\vec{c}_h^i(t_e^i)$ as the cepstrum of the higher pitch speaker in the last frame of run i, $\vec{c}_l^i(t_e^i)$ as

[1] There are actually two separate tracking approaches, *spectral envelope* and *liftered cepstra*. As they differ only according to which vector ($\vec{s}$ or $\vec{c}$) is used in the distance metric defined in Equation (5), the term "spectral envelope tracking" will be employed exclusively. Results, however, will be reported in terms of the vector ($\vec{s}$ or $\vec{c}$) employed.


MAGNITUDE
1 2f 3f 4f 5f
FREQUENCY
FRAME OF CO-CHANNEL SPEECH
HIGH-PITCH VALUE
LOW-PITCH VALUE
Q/D STS LLSE SEPARATION SYSTEM
COMPLEX HARMONIC LINE SPECTRUM FOR HIGH-PITCH SPEAKER
COMPLEX HARMONIC LINE SPECTRUM FOR LOW-PITCH SPEAKER
MAGNITUDE
1 2f 3f 4f 5f
MAGNITUDE
1 2f 3f 4f 5f
COMPLEX HARMONIC LINE SPECTRUM FOR HIGH-PITCH SPEAKER
PIECEWISE CONSTANT INTERPOLATION
LOG MAGNITUDE
INVERSE DFT
LIFTERING
DFT
LOG MAGNITUDE ENVELOPE FOR HIGH-PITCH SPEAKER
MAGNITUDE
1 2f 3f 4f 5f
LOW ORDER CEPSTRUM FOR HIGH-PITCH SPEAKER

(Process Is Identical for Low-Pitch Speaker)

Figure 6. The Quatieri/Danisewicz system takes as input two pitch values and the co-channel speech for a given frame of speech and produces two sets of harmonically related sinewave frequencies, magnitudes, and phases. These harmonic line spectra are converted to spectral envelopes by means of cepstral liftering. Either the liftered cepstra or the corresponding spectra can be used as input to the tracker.

the cepstrumof the lower pitch speaker in the last frame of run i, $\vec{c}_h^{i+1}(t_b^{i+1})$ as the cepstrum of the higher pitch speaker in the first frame of run $i+1$, $\vec{c}_l^{i+1}(t_b^{i+1})$ as the cepstrum of the lower pitch speaker in the first frame of run $i+1$, a cross is hypothesized if $\epsilon_c^{i,i+1} < \epsilon_s^{i,i+1}$, where

$$\begin{aligned} \epsilon_s^{i,i+1} &= \|\vec{c}_h^i(t_e^i) - \vec{c}_h^{i+1}(t_b^{i+1})\| + \|\vec{c}_l^i(t_e^i) - \vec{c}_l^{i+1}(t_b^{i+1})\| \\ \epsilon_c^{i,i+1} &= \|\vec{c}_h^i(t_e^i) - \vec{c}_l^{i+1}(t_b^{i+1})\| + \|\vec{c}_l^i(t_e^i) - \vec{c}_h^{i+1}(t_b^{i+1})\|, \end{aligned} \tag{5}$$

and $\|\vec{x}\|$ is the magnitude of $\vec{x}$.

3.3.2 Handling Frames Between Runs

Just as in the case of pitch track interpolation, the simplest approach for handling frames between runs is to assign the pitches in these frames arbitrarily. Other approaches, including assigning frames based on their cepstral or spectral differences with $\vec{c}_h^i(t_e^i)$, $\vec{c}_l^i(t_e^i)$, $\vec{c}_h^{i+1}(t_b^{i+1})$, $\vec{c}_l^{i+1}(t_b^{i+1})$ were postulated but not evaluated.

3.3.3 Discussion

The spectral envelope tracking joint pitch estimator shares the main weaknesses of the pitch track interpolator: (1) it has no way of knowing which track belongs to the target and which to the jammer and (2) run assignment is not independent from one run to the next. It does, however, demonstrate the use of nonpitch information for pitch tracking—an idea that led to spectral envelope classification described in Section 3.4.

3.4 Spectral Envelope Classification

Using either of the two techniques for performing two-speaker pitch tracking results in a nonrobust system in that a single error (hypothesize cross when no-cross occurred or hypothesize no-cross when cross occurred) can invalidate all subsequent cross resolution efforts. The third and final pitch tracking approach, spectral envelope classification, performs speaker identification as first introduced for co-channel interference in Zissman, Weinstein, and Braida [21] on the cepstral vectors generated from the output of the QD system. This system requires training on isolated speech from the target and jammer speakers, making it inappropriate when isolated single-speaker target and jammer utterances are not available. Details of the algorithm with some discussion follow.

3.4.1 Algorithm

Figure 7 is a block diagram of the spectral envelope classification system. (Inasmuch as the liftered cepstra are actually being classified, this system might be better described as the liftered-cepstra classification system; however, as there is a one-to-one correspondence between cepstra and spectra, and due to the strong, intuitive connection between spectral envelopes and the model of speech production, "spectral envelope classification" will be employed in this report.) During training a cepstral exemplar library is created for both target and jammer. The target library contains every cepstrum observed in the target training speech, while the jammer library contains every cepstrum observed in the jammer training speech. These cepstra are calculated by performing single-speaker pitch analysis followed by harmonic sinusoidal analysis on the single-speaker speech. The resulting harmonic frequencies, magnitudes, and phases are processed as described in Section 3.3.1. During tracking, for each frame of a run the system compares the two measured liftered cepstra (generated as described in Section 3.3.1) with the cepstra stored in the two libraries. Given the ith vector $\vec{C}_i^T$ in the target library, the ith vector $\vec{C}_i^J$ in the jammer library, the observed high-pitch cepstrum vector $\vec{c}_h(t)$ at frame t, and the observed low-pitch cepstrum vector $\vec{c}_l(t)$ at frame t, the classifier uses nearest neighbor classification as follows:

$$\begin{aligned} \varepsilon_{Th}(t) &= \arg\min_i \|\vec{C}_i^T - \vec{c}_h(t)\| + \arg\min_i \|\vec{C}_i^J - \vec{c}_l(t)\| \\ \varepsilon_{Jh}(t) &= \arg\min_i \|\vec{C}_i^J - \vec{c}_h(t)\| + \arg\min_i \|\vec{C}_i^T - \vec{c}_l(t)\|. \end{aligned} \tag{6}$$

If $\varepsilon_{Th}(t) < \varepsilon_{Jh}(t)$, then it is hypothesized that the target was higher for frame t. Otherwise the jammer-higher rule is hypothesized. After each frame in the run is classified target-higher or jammer-higher, the entire run is classified as target-higher or jammer-higher according to the number of individual frames so classified, with the majority ruling.

3.4.2 Handling Frames Between Runs

As described in the discussion of the spectral envelope tracker, the simplest approach for handling frames between runs is to assign the pitches in these frames arbitrarily. That approach was employed here as well.

3.4.3 Discussion

The spectral envelope classification system requires training on both target and jammer single speaker speech. Given such training data, it is capable of labeling each run as target-higher or jammer-higher, thereby producing hypothesized target and jammer pitch tracks contrasted with the pitch track interpolation and spectral envelope tracking systems that could only label the tracks as speaker-1 and speaker-2. In addition, as each run is labeled independently of all others, the effect of a single run mislabeling does not extend outside the run. Thus, to use the example of Section 3.2.4, given a co-channel utterance with 10 true runs roughly equally spaced through the utterance and given a 90% effective tracker (9 runs correctly classified, 1 incorrectly), 90% of the frames would be correctly labeled independent of which run was mislabeled. This more graceful failure characteristic of the spectral envelope classifier is seen as a significant advantage.

As the number of available single-speaker training vectors was predicted to be small, the nearest neighbor was the only classifier employed in this work. Small amounts of training data make the impact of individual storage of each training vector and subsequent comparisons against

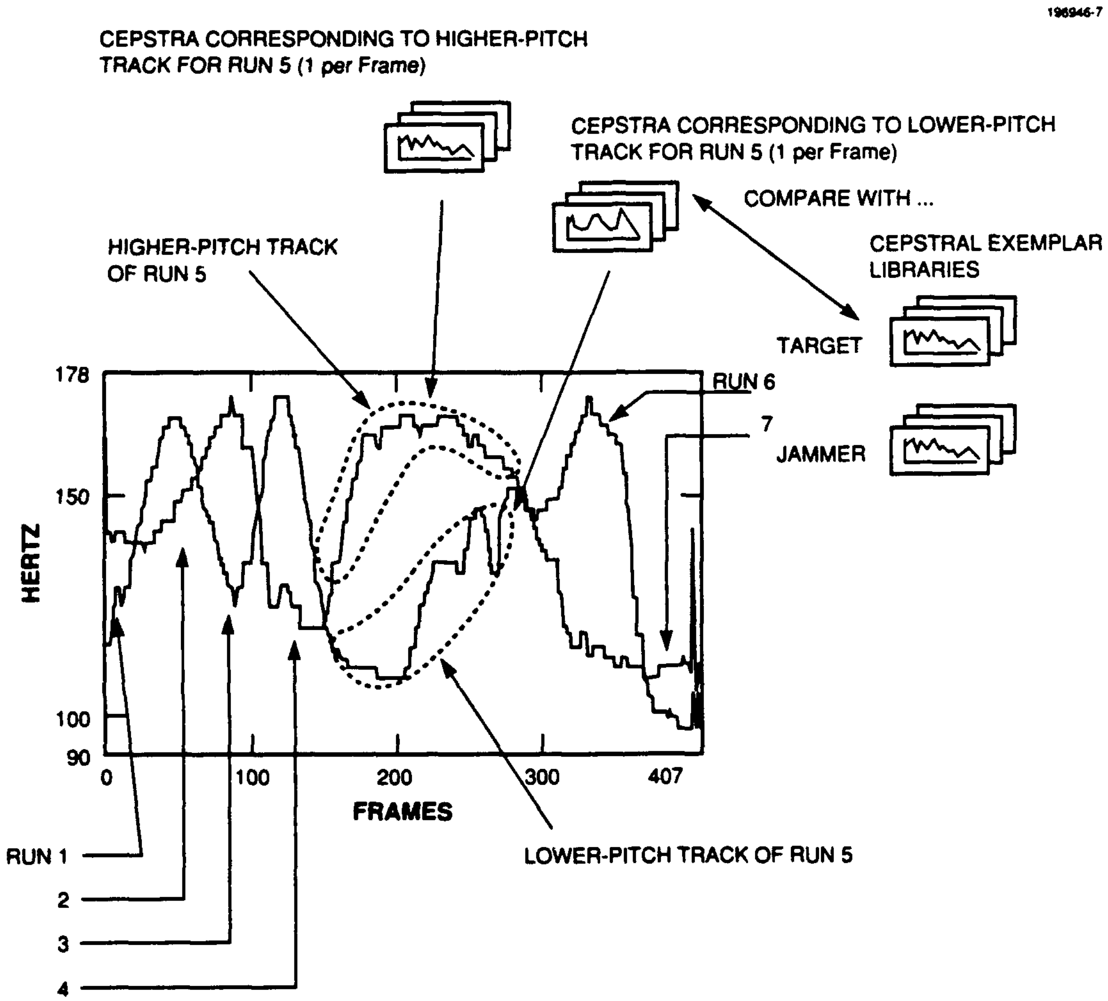


Figure 7. A block diagram of the spectral envelope classification system. The dashed ovals show the components of a single run. Two cepstra are calculated for each frame of the run, one for the higher pitch speaker and one for the lower. These cepstra are comparedagainst libraries of target and jammer speech. Eventually, the tracks in each run are assigned either target-higher or jammer-higher.

each training vector mandated by the nearest neighbor classifier manageable. Other classifiers, e.g., Gaussian mixture or neural network-based approaches, could be employed with larger training sets.

4. EXPERIMENTS

This section describes the experiments that were run to evaluate the three pitch trackers described earlier.

4.1 Implementation

All software was written in C for a Sun Microsystems SPARCstation 2 running SunOS 4.1.1. In addition, the spectral envelope classifier used the nearest neighbor classifier algorithm embedded in the LNKnet classifier software package developed at Lincoln Laboratory by Richard Lippmann and Linda Kukolich.

4.2 Database

The Stubbs/Summerfield [12] all-voiced sentence database was used to evaluate the pitch trackers. This database comprises 9 sentences from talker q and 36 from talker r, from which 27 q sentences were dropped, leaving a subset of the database, containing 9 talker q and 9 talker r sentences. (The talker r sentences used were numbers 2, 7, 11, 13, 15, 17, 19, 31, and 36.) For evaluating the pitch track interpolation and spectral envelope tracking systems, 81 pairs of co-channel sentences were generated for testing. Truncation of the longer (either target or jammer) utterance was used to prevent processing of single-speaker speech.

The spectral envelope classification system required reserving some of the sentences for training. Rather than split the 9 sentences per talker into test and training sets, the 81 pairs of sentences were processed individually. For a run on sentence r_i and q_j, all r_k with $k \neq i$ were used to form the r library and all q_k with $k \neq j$ were used to form the q library. This method of conserving training and test data, called jackknifing or cross-validation, ensures that test and training data remain disjoint for each individual test run.

This database is imperfect in many aspects, one of which is that it contains the speech for only two speakers. While the pitch contours for these speakers did cross several times in each co-channel sentence pair, a richer database, both in number of speakers and number of sentences per speaker is desirable; however, given the constraint that the trackers could only handle voiced speech and given the status and availability of the database as a CCTIS "standard" (used in a very loose sense; it appears that this may be the only database used by more than one CCTIS site), it was used nonetheless. In addition, while training and test data were always disjoint, the database was development-only—there were no previously unseen data available for a one time, final evaluation.

Some of the work done previously at the Laboratory in speech-state-adaptive simulation of co-channel suppression [22] and speaker activity labeling [21] had used a three-speaker database made available by the Communication Biophysics Group at MIT's Research Laboratory of Electronics.

This so-called MIT-CBG database was not used in the present study, as its utterances contain voiced and unvoiced speech.

4.3 Results

This section describes the performance of the three pitch trackers, preceded by some discussion of run characteristics. Pitch tracking results arequoted in terms of the number of frames in the sentence pair for which pitch assignment (target-higher or jammer-higher) was correct. Truth (target-higher or jammer-higher) was determined from single-speaker cepstral pitch analysis [23] of each single-speaker utterance in isolation.

4.3.1 Run Analysis

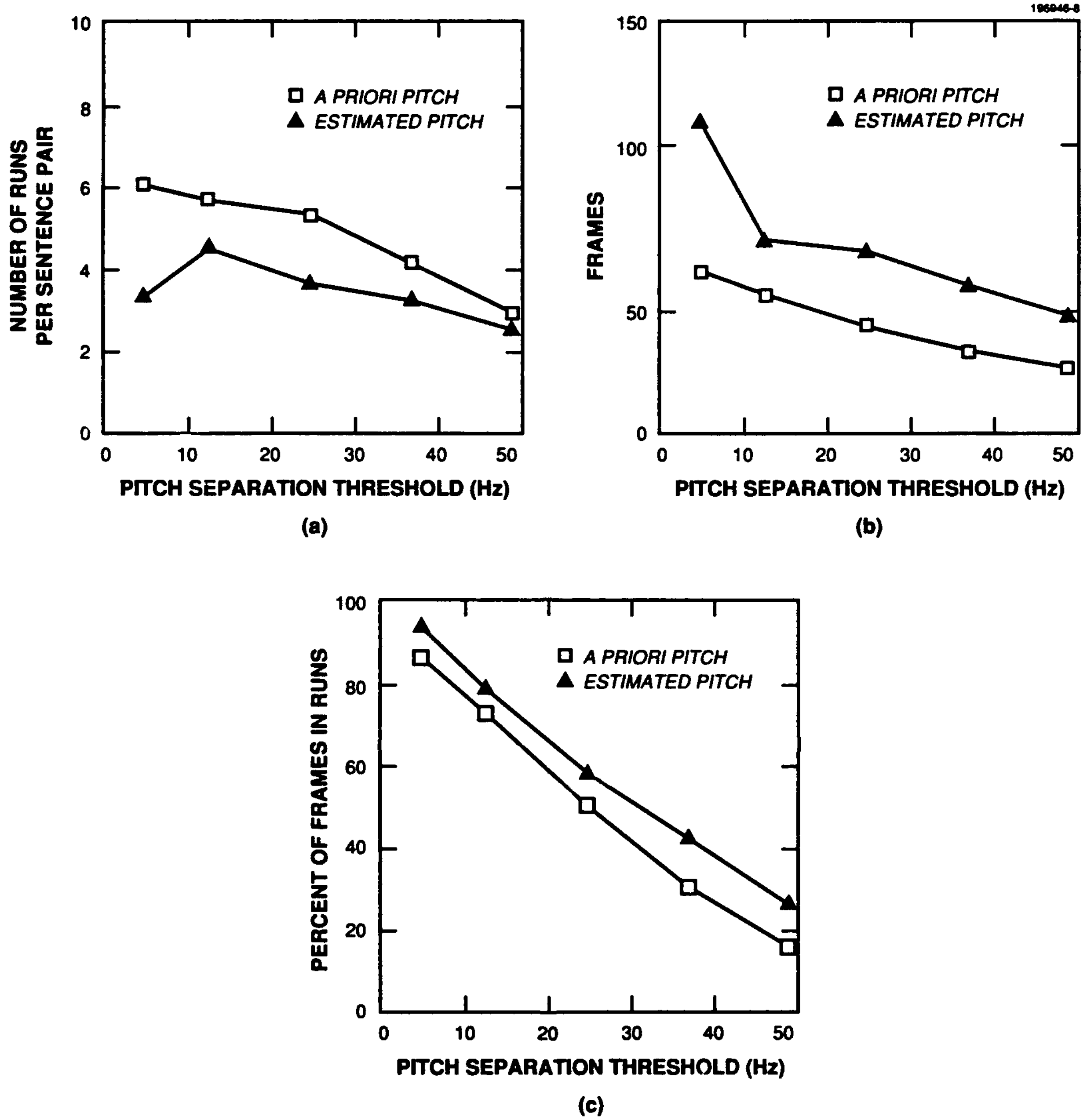
Figure 8 shows how average run length, average number of runs per sentence pair, and average percentage of runs in frames varies with p_t for the all-voiced database. As expected, the number of runs per sentence pair decreases generally as p_t increases, as there are fewer frames with pitch differences above threshold. The increase at low values of p_t for the estimated pitch case is due to the noisy pitch values produced by the joint pitch estimator when the actual pitches are very close. The average length of the runs decreases as p_t increases because as p_t increases fewer and fewer frames have a pitch difference above threshold. Finally, the number of frames in a sentence pair that are part of a run decreases linearly with p_t.

4.3.2 Pitch Track Interpolation

The results of the pitch track interpolation experiments that are shown in Table 1 were generated using the two tracks produced by the interpolator to produce a rule (speaker-1-higher or speaker-2-higher) for each frame in the utterance. All processing used the modified error criteria $\hat{E}_s^{i,i+1}$ and $\hat{E}_c^{i,i+1}$ that was defined in Equation (3). All input pitch values were those produced by single-speaker cepstral pitch estimation, that is, no tests of the pitch track interpolator were run using the outputs of joint pitch estimator. Thus, results present upper-bound performance obtainable only with a perfect joint pitch estimator. Further, as the pitch track interpolator cannot assign speaker-1 and speaker-2 to target and jammer, results were calculated for both hypotheses (speaker-1 is target, speaker-2 is target), and the hypothesis that resulted in the best score was chosen. The performances reported here are upper bounds on the performance of a standalone system. This "cheating" is analogous to playing a reconstructed target and jammer, instructing the listener to choose which signal better enhances the target, and asking the listener to perform target transcription on the chosen utterance for formal intelligibility evaluation of the system.

4.3.3 Spectral Envelope Tracking

The results of the spectral envelope tracking experiments that are shown in Table 2 were generated using the sets of spectral envelopes (or liftered cepstra) at the endpoints of each run to



(a)

(b)

(c)

Figure 8. Run analysis, showing (a) the average number of runs in each sentence pair as a function of p_t, (b) how run length varies with p_t, and (c) the average percentage of frames in the entire sentence pair that were contained in runs runs as a function of p_t. All graphs show results for a priori and jointly estimated pitch input.

TABLE 1

Pitch Track Interpolation Results

Pitch Separation Threshold (p_t) (Hz)	Order of Interpolating Polynomial	Performance (Fraction of Frames Correct)
7	2	0.668
12	2	0.675
7	3	0.699
12	3	0.682
7	4	0.707
12	4	0.669

produce a rule (speaker-1-higher or speaker-2-higher) for each frame in the utterance. All input pitch values were those produced by single-speaker cepstral pitch estimation, that is, no tests of the spectral envelope tracker were run using the outputs of the joint pitch estimator. Thus, results present upper-bound performance obtainable only with a perfect joint pitch estimator. Once again, as the envelope tracker cannot assign speaker-1 and speaker-2 to target and jammer, results were calculated for both hypotheses, and the hypothesis that resulted in the best score was chosen.

4.3.4 Combined Interpolation and Envelope Tracking

Combined pitch track interpolation and spectral envelope tracking was investigated as a means of improving performance. Given the pitch track interpolation modified error criteria $\hat{E}_s^{i,i+1}$ and $\hat{E}_c^{i,i+1}$ [see Equation (3)] and the spectral envelope tracking error criteria $\epsilon_s^{i,i+1}$ and $\epsilon_c^{i,i+1}$ [(see Equation (5)], a weighted sum of the two errors was formed:

$$\begin{aligned} e_s^{i,i+1} &= (1-\alpha)\hat{E}_s^{i,i+1} + \alpha\epsilon_s^{i,i+1} \\ e_c^{i,i+1} &= (1-\alpha)\hat{E}_c^{i,i+1} + \alpha\epsilon_c^{i,i+1}. \end{aligned} \tag{7}$$

Best performance was obtained with $\alpha = 1$, that is by using only the spectral envelope information and ignoring that of the pitch track interpolation.

4.3.5 Spectral Envelope Classification

Figure 9 shows the pitch tracking performance of the spectral envelope classification tested under the conditions listed in Table 3. Generally, processing a priori pitch values obtained from

TABLE 2

Spectral Envelope Tracking Results

Pitch Separation Threshold (p_t) (Hz)	Order of Cepstral Window	Vector Type Cepstral (C) or Log Magnitude (LM)	Performance (Fraction of Frames Correct)
7	3	C	0.749
12	3	LM	0.741
7	3	C	0.724
12	3	LM	0.726
7	5	C	0.744
12	5	LM	0.735
7	5	C	0.717
12	5	LM	0.703
7	17	C	0.744
12	17	LM	0.747
7	17	C	0.719
12	17	LM	0.718

single-speaker cepstral estimates running on the single-speaker speech resulted in higher performance than processing jointly estimated pitch values obtained from the co-channel speech. Figure 10 shows how the ability to classify frames in runs, as opposed to all frames of a sentence pair, varied as a function of p_t. The graph shows that as p_t increased, classification performance decreased, which is explained by the drop in average run length as p_t increases — fewer runs in frames mean fewer observations available on which to make each run classification.

4.4 Discussion

The results show that performance of the spectral envelope classification method of pitch tracking exceeded that of pitch track interpolation, spectral envelope tracking, and their combination. These results were sustained even when the latter two methods were allowed to operate at their upper-bound — when the track that matched the target better was labeled by the grading system as the target track.

The best spectral envelope classification system operated at 87% frames correctly classified when supplied with two pitch tracks generated using a pitch estimator operating on the single

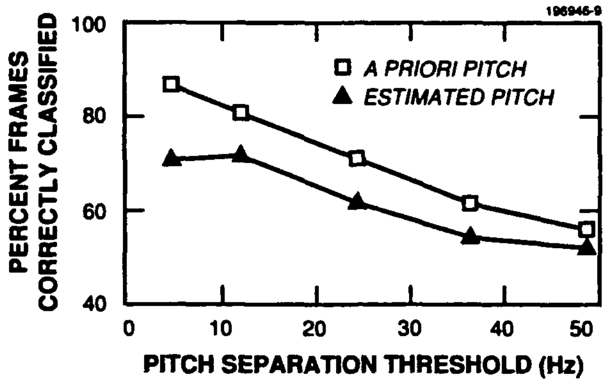


Figure 9. Performance of the spectral envelopeclassification pitch tracker. Results are shown as a function of pitch source (a priori or jointly estimated) and pitch separation threshold (p_t).

TABLE 3

Spectral Envelope Classification Parameters

Parameter	Value
Single-speaker pitch estimator	Cepstral
Co-channel joint pitch estimator	Grid/gradient search method [24]
FFT size for STS analysis	4096
STS peaks forced harmonic	Yes
Maximum number of peaks	100
IFFT size for cepstral analysis	4096
Feature vector composition	Cepstral coefficients 1 through 19
Zeroth cepstral coefficient used	No
Classification algorithm	Nearest neighbor

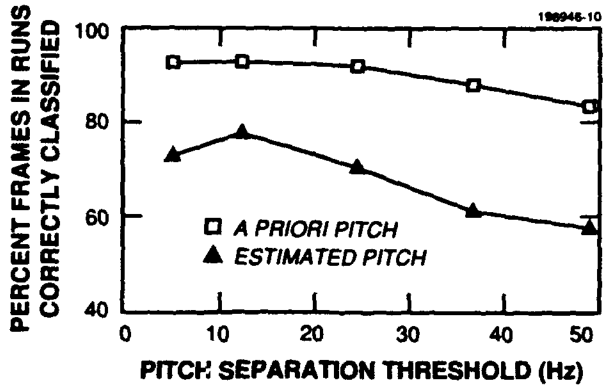


Figure 10. Ability of the spectral envelope classification pitch tracker to classify correctly frames in runs, ignoring those not in runs.

speaker speech. When input was supplied from a joint pitch estimator, performance dropped to 72%, indicating that the joint pitch estimates were quite noisy.

5. CONCLUSIONS AND FUTURE WORK

Three different approaches to the co-channel talker pitch tracking problem have been proposed. Each approach identifies and then connects runs of frames where the pitches are well separated. While the first two methods, pitch track interpolation and spectral envelope tracking, require no training, they have the disadvantage that a single run connection error can contaminate all subsequent tracking efforts. The third method, spectral envelope classification, is more robust to errors, but requires training on the target and the jammer of interest.

As performance of the spectral envelope classification system was relatively high on the objective pitch tracking experiments that were performed as part of this study, future attention might focus on making this system less dependent on training data. For example, it might be possible to train the system only on target single-speaker speech, so that in each frame the system would decide target-higher or jammer-higher based on which spectral envelope was most similar to entries in the target library. With some human intervention it might be possible to train target and jammer libraries on the fly from the co-channel speech. The human might label a few pitch tracks as belonging to the target or jammer; the system would then create library entries from the spectra corresponding to these pitch tracks and classify subsequent spectra using these rather small, bootstrapped libraries. Finally, it might be possible to combine either or both of the other tracking techniques with the classification technique to arrive at a better overall pitch tracker.

As the results showed a significant difference between pitch tracking results on a priori versus *jointly estimated pitch* data, efforts might be focused on better joint pitch estimators. While the work in this report was limited to 0 dB TJR, the DQ separation system has been shown informally to work over a wide range of TJRs, though the grid/gradient joint pitch estimator has been tested successfully so far only at 0 dB. Future research might study the performance of the three pitch trackers in the non-0-dB cases. The systems described in this report have no means to handle unvoiced speech and silence, i.e., the cases where only one (or possibly none) pitch estimate is available from the joint pitch estimator. Extension to the unvoiced speech and silence cases would be a logical next step.

The true test of whether the pitch tracker is successful is its effect on the intelligibility and quality of the resulting enhanced target speech. The Stubbs/Summerfield database may not be appropriate for formal intelligibility tests because it contains only two speakers, both of whom speak "British" English, with only nine sentences for one of them. Because studies have shown that digital speech processing affectsintelligibility differently for native and nonnative speakers of a language [25], testing in the United States would be difficult. Perhaps a large, all-voiced "American" English database could be collected for joint evaluation of the DQ system with joint pitch tracking. Even more useful would be a corpus designed specifically for development and evaluation of CCTIS systems with a subset of all-voiced sentences. Such a corpus would have dozens of speech utterances for each speaker, allowing proper training of both speaker-dependent separation systems and human subjects used in intelligibility tests.

REFERENCES

1. V.C. Shields, Jr., "Separation of added speech signals by digital comb filtering," Department of Electrical Engineering, Massachusetts Institute of Technology (September 1970).

2. J.K. Everton, Sr., "The separation of the voice signals of simultaneous speakers," Department of Computer Science, University of Utah (June 1975).

3. R.H. Frazier, S. Samsam, L.D. Braida, and A.V. Oppenheim, "Enhancement of speech by adaptive filtering," *ICASSP '76 Proc.*, 251–253 (April 1976).

4. R.J. Dick, "Co-channel interference separation," Technical Rep. RADC-TR-80-365, Griffiss Air Force Base, N.Y.,: Rome Air Development Center (December 1980).

5. T.W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust. Soc. Am.*, 911–918 (October 1976).

6. Y.M. Perlmutter, L.D. Braida, R.H. Frazier, and A.V. Oppenheim, "Evaluation of a speech enhancement system," *ICASSP '77 Proc.*, 212–215 (May 1977).

7. B.A. Hanson and D.Y. Wong, "The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence of interfering speech," *ICASSP '84 Proc.*, 18A.5.1–18A.5.4 (March 1984).

8. M. Weintraub, "A computational model for separating two simultaneous talkers," *ICASSP '86 Proc.*, 81–84 (April 1986).

9. D.G. Childers and C.K. Lee, "Co-channel speech separation," *ICASSP '87 Proc.*, 181–184 (April 1987).

10. J.A. Naylor and S.F. Boll, "Techniques for suppression of an interfering talker in co-channel speech," *ICASSP '87 Proc.*, 205–208 (April 87).

11. T.F. Quatieri and R.G. Danisewicz, "An approach to co-channel talker interference suppression using a sinusoidal model for speech," *IEEE Trans. Acoust. Speech Signal Process.*, 56–69 (January 1990).

12. R.J. Stubbs and Q. Summerfield, "Algorithms for separating the speech of interfering talkers; evaluations with voiced sentences, and normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, 359–372 (January 1990).

13. J. Naylor and J. Porter. "An effective speech separation system which requires no a priori information," *ICASSP '91 Proc.*, 937–940 (May 1991).

14. L. Lee and D. Morgan, "An algorithm for co-channel speaker separation with applications for speech enhancement and suppression," Technical Rep. SPCOT-91-001, Nashua, N.H.: Lockheed Sanders, Inc. (May 1991).

REFERENCES
(Continued)

15. M.A. Zissman, "Cochannel talker interference suppression," Technical Rep. 895, Lexington, Mass.: MIT Lincoln Laboratory (July 1991).

16. T.F. Quatieri and R.G. Danisewicz, "An approach to co-channel talker interference suppression using a sinusoidal model for speech," *ICASSP '88 Proc.*, 565–568 (April 1988).

17. T.W. Parsons, "Study and development of speech-separation techniques," Technical Rep. RADC-TR-78-105, Griffiss Air Force Base, N.Y.: Rome Air Development Center (May 1978).

18. J. Naylor, Private communication (April 1991).

19. W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling, *Numerical Recipes in C*, Cambridge: Cambridge University Press (1988).

20. R.J. McAulay and T.F. Quatieri, "Low rate speech coding based on the sinusoidal model," in *Advances in Speech Signal Processing*, S. Furui and M.M. Sondhi, eds, New York: Marcel Dekker, Inc. (1992).

21. M.A. Zissman, C.J. Weinstein, and L.D. Braida, "Automatic talker activity labeling for co-channel talker interference suppression," *ICASSP '90Proc.* 813–816 (April 1990).

22. M.A. Zissman, C.J. Weinstein, L.D. Braida, R.M. Uchanski, and W.M Rabinowitz, "Speech-state-adaptive simulation of co-channel talker interference suppression," *ICASSP '89 Proc.* 361–364 (May 1989).

23. A.M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Am.*, 293–309 (February 1967).

24. R.G. Danisewicz and T.F. Quatieri, "An Approach to co-channel talker interference suppression using a sinusoidal model for speech," Technical Rep. 794, Lexington, Mass.: MIT Lincoln Laboratory (February 1988).

25. M.A. Mack and J. Tierney, "The intelligibility of natural and vocoded semantically anomalous sentences: A comparative analysis of English monolinguals and German-English bilinguals," Technical Rep. 972, Lexington, Mass.: MIT Lincoln Laboratory (December 1987).

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (*Leave blank*)

2. REPORT DATE
30 April 1992

3. REPORT TYPE AND DATES COVERED
Technical Report

4. TITLE AND SUBTITLE

Two-Talker Pitch Tracking for Co-channel Talker Interference Suppression

5. FUNDING NUMBERS

C — F19628-90-C-0002
PE — 33401F, 64771F, 62702F

6. AUTHOR(S)

Marc A. Zissman and Dewitt C. Seward IV

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Lincoln Laboratory, MIT
P.O. Box 73
Lexington, MA 02173-9108

8. PERFORMING ORGANIZATION REPORT NUMBER

TR-951

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Air Force
RL/IRAA
Griffiss AFB, NY 13441

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

ESD-TR-91-251

11. SUPPLEMENTARY NOTES

None

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (*Maximum 200 words*)

Almost all co-channel talker interference suppression systems use the difference in the pitches of the target and jammer speakers to suppress the jammer and enhance the target. While joint pitch estimators outputting two pitch estimates as a function of time have been proposed, the task of the proper assignment of pitch to speaker (two-talker pitch tracking) has proven difficult. This report describes several approaches to the two-talker pitch tracking problem, including algorithms for pitch track interpolation, spectral envelope tracking, and spectral envelope classification. When evaluated on an all-voiced two-talker database, the best of these new tracking systems correctly assigned pitch 87% of the time when given perfect joint pitch estimation.

14. SUBJECT TERMS

co-channel talker interference suppression
speaker separation
sinusoidal transform
pitch tracking

15. NUMBER OF PAGES
44

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
Unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified

20. LIMITATION OF ABSTRACT
SAR