Evaluation of a Method for Separating Digitized Duet Signals*

ROBERT C. MAHER

Department of Electrical Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588-0511, USA

A new digital signal-processing method is presented for separating two monophonic musical voices in a digital recording of a duet. The problem involves time-variant spectral analysis, duet frequency tracking, and composite signal separation. Analysis is performed using a quasi-harmonic sinusoidal representation based on short-time Fourier transform techniques. The performance of this approach is evaluated using real and artificial test signals. Applications include background noise reduction in live recordings, signal restoration, musicology, musique concrète, and digital editing, splicing, or other manipulations.

0 INTRODUCTION

Separation of superimposed signals is a problem of interest in audio engineering. For example, it would often be useful to identify and remove undesired interference (such as audience or traffic noise) present during a live recording. Other examples include separation and replacement of errors in a recorded musical performance, separation of two simultaneous talkers in a single communications channel, or even adjustment of the level imbalance occurring when one musician in an ensemble briefly turns away from the microphone.

Considered in this paper is a digital signal-processing approach to one aspect of the ensemble signal separation problem: separation of musical duet recordings. The primary goal of this project was to develop and evaluate an automatic signal separation system based primarily on physical measurements rather than psychoacoustic models of human behavior.

In order to separate the desired and undesired signals we must resort to prior knowledge of some aspect of the superimposed signals, whereby a set of separation criteria may be identified. If two interfering signals occupy nonoverlapping frequency bands, for example, the separation problem can be solved by using frequency-selective filters. In other cases the competing signals may be described in a statistical sense, allowing separation using correlation or a nonlinear detection method. However, most superimposed signals, such as two musical instruments playing simultaneously, do not allow for such elementary decomposition methods, and other strategies applicable for signal separation must be discovered.

In the case of ensemble music, sounds emanating from different musical instruments are combined in an acoustic signal, which may be recorded via a transducer of some kind. Despite the typical complexity of the recorded ensemble signal, a human listener can usually identify the instruments playing at a given point in time. Further, a listener with some musical training or experience can often reliably transcribe each musical voice in terms of standard musical pitch and rhythm. Unfortunately the methods and strategies used by human observers are not introspectable and thus cannot serve easily as models for automatic musical transcription or signal separation systems.

In order to put the current work into perspective, the

* Manuscript received 1990 January 29; revised 1990 June 21.

narrative portion of this paper begins with a review of the approach and methods used in this investigation. The separation procedure is described next, followed by a critical evaluation of the results and a concluding section concerning the successes, failures, and future prospects of this research.

1 REVIEW OF APPROACH AND METHODS

Previous work related to the signal separation problem considered here has been primarily in two areas, 1) separating the speech of two people talking simultaneously in a monaural channel (also called cochannel speech separation) [1]–[6] and 2) segmentation and/or transcription of musical signals [7]–[10]. The goal of the speech separation task is to improve the intelligibility of one talker by selectively reducing the speech of the other talker, while for music separation the goal is to extract the signal of a single instrumental line (or to produce printed musical notation) directly from a recording.

1.1 Separation of Speech and Music

The cochannel speech and musical signal separation problems share some common approaches. Both tasks have typically been formulated in terms of the time-variant spectrum of the individual signal sources. This is appropriate because one possible basis for separating additively combined signals is to distinguish the frequency content of the individual signals, assuming a linear system.

1.1.1 Cochannel Speech Separation

The cochannel speech separation work reported to date typically relies on some assumptions about the spectral properties of speech. For voiced speech (such as vowel sounds), the short-time magnitude spectrum contains a series of nearly harmonic peaks corresponding to the fundamental frequency and overtones of the speech signal. With two talkers, the composite spectrum contains the overlapping series of peaks for *both* voices. The common approach has been somehow to identify which peaks go with which talker and to isolate them. The separation itself has been attempted using comb filters to pass only the spectral energy belonging to one of the talkers (or notch filters to reject one of the talkers), identification and separation of spectral features (peaks) belonging to one of the talkers, and even extraction of speech parameters appropriate for use in regenerating the desired speech using a synthesis algorithm. No separation process specifically for unvoiced speech (such as fricatives or noise) has been reported in the literature.

1.1.2 Segregation of Voices in Ensemble Music Recordings

Identification of pitches, rhythms, and timbres from a musical recording is not a trivial task in general. The difficulties are formidable. The ensemble voices may occur simultaneously or separately (and no voices may occur during shared rests); level imbalances between voices may be present; noise of various kinds may hinder the detection process, and so forth. In the frequency domain the partials (overtones) of the various voices will often overlap, preventing simple identification of the spectra of each voice. Further, the fundamental basis of most music is *time*, so some means to segment the recording into time intervals and to correlate the parameters from instant to instant must be concocted.

1.2 Research Limitations on the Scope of the Separation Problem

Because the musical signal separation problem is so complex, the initial need for this investigation was to simplify the conditions. Thus the range of input possibilities was limited by the following restrictions:

1) The recordings to be processed may contain only two separate, monophonic voices (musical duets).

2) Each voice of the duet must be harmonic, or nearly so, and contain a sufficient number of partials so that a meaningful fundamental frequency can be determined.

3) The range of fundamental frequencies for each voice must be restricted to nonoverlapping ranges, that is, the lowest musical pitch of the upper voice must be greater than the highest pitch of the lower voice. Note that a duet that does not meet this requirement in toto may still be processed if it can be divided manually into segments obeying this restriction.

4) Reverberation, echoes, and other correlated noise sources are discouraged since, in effect, they represent additional "background voices" in the recording and violate the duet assumption.

Despite these seemingly severe restrictions, the remaining difficulties are still nontrivial: how to separate the partials of the two voices when the spectra overlap; how to determine whether zero, one, or both voices are present at a given point in time; how to track each voice reliably when one is louder than the other; and so on. Moreover, success with a particular duet does not automatically guarantee success on every other duet example. In fact, projects of this sort can rapidly fall into the trap of ad hoc, special-purpose techniques to solve a particular problem, only to find another problem created.

The system developed during this research project was not necessarily intended for real-time operation. Thus the algorithms were implemented in software on a general-purpose computer. This approach has the advantages of extensive software support, relative ease of testing, and rapid debugging cycles.

1.3 Fundamental Research Questions

The major goal of this investigation was to demonstrate the feasibility of automatic composite signal decomposition using a time–frequency analysis procedure. This problem can be stated as two fundamental questions.

1) How may we automatically obtain accurate estimates of the time-variant fundamental frequency of

each musical voice from a digital recording of a duet?

2) Given time-variant fundamental frequency estimates of each voice in a duet, how may we identify and separate the interfering partials (overtones) of each voice?

Question 1) treats the problem of estimating the time-variant frequencies of each partial for each voice. Assuming nearly harmonic input signals, specification of a fundamental frequency identifies the partial component frequencies of that voice. Conflicting (coincident) partial frequencies between the two voices can then be identified by comparing the predicted harmonic series of the two duet voices.

Question 2) involves the fundamental constraints on simultaneous time and frequency resolution. The desire for high-resolution frequency domain information requires observation of the input signal over a long time span. However, long observation spans often result in an unacceptable loss of time resolution by averaging out any spectral changes during the observation interval. Thus the analysis system must somehow cope with this inherent uncertainty in determining the best time-versus-frequency representation for the input duet signal.

Note that question 2) can be treated separately from question 1) if the time-variant fundamental frequency pair for the duet can be obtained by some manual means. For example, a duet synthesized with known fundamental frequencies (a priori frequency information) can be used to evaluate a preliminary separation algorithm. Thus the two fundamental questions can be treated initially as separate problems if desired.

1.4 Research Question 1: Duet Frequency Tracking

The duet separation methods considered in this paper require good estimates of the fundamental frequency of each voice at all times. This information could come from an accurate musical score, some manual means of tabulation, or an automatic frequency tracking system. However, even if a musical score is available, musicians seldom play music with an exact, one-to-one correspondence with the printed information. Manual methods can be quite reliable, but are extremely tedious and time consuming. Thus automatic methods are of primary interest in this paper.

1.4.1 Common Methods for Pitch Detection

Fundamental frequency tracking is often called pitch detection or pitch extraction. Numerous reports describing algorithms for monophonic pitch detection have been published, including the cepstrum method [11], autocorrelation [12], the period histogram and other harmonic-based methods [13], [14], the optimum comb and average magnitude difference function (AMDF) [15], [16], and methods based on linear prediction [17], [18]. Also, time-domain methods to determine pitch periods by zero crossings, peak detection, or clipped waveform analysis have been developed. Unfortunately no single method for pitch detection has been found to be reliable for arbitrary input signals.

1.4.2 Application of Monophonic Methods to Duets

For the duet separation task, the difficulties of monophonic pitch detection are compounded by the presence of two competing signal sources. There is no certainty that a monophonic pitch detection scheme can handle multiple simultaneous signals. For example, the autocorrelation and optimum comb methods are used to identify periodicites in the input signal by searching for a delay lag T_0 that maximizes the integrated product (autocorrelation) or minimizes the summed absolute value of the difference (optimum comb and AMDF). The fundamental frequency estimate is given by $f_0 = 1/T_0$. However, identification of the extremum corresponding to the "best" T_0 is not trivial because the search functions contain many subextrema, that is, the functions are not unimodal. Also, delay lags of an integral number of waveform periods will show similar extrema in the autocorrelation or AMDF, leading to possible octave errors.

The problem of octave errors is a common obstacle to many pitch detection algorithms, including harmonic-based methods. The difficulties are particularly noticeable for instruments with strong resonances such that certain upper partials (or ranges of partials) contain much more energy than the lower partials. In situations where the search range is known to be limited to less than an octave (often the case with speech) the octave error problem can be reduced. Musical melodies, on the other hand, often span a larger fundamental frequency range. Moreover, when two sources are present in the input signal, interactions between the numerous pairs of partials cause additional difficulties, which make most monophonic pitch detection methods impractical for direct application to the duet case. For this reason, a new scheme for duet frequency tracking was developed for this project, as described in Sec. 2.

1.5 Research Question 2: Time–Frequency Analysis

The second research question treats the general problem of identification and separation of the spectral components in a musical duet. Specifically, some useful representation of the duet signal simultaneously in the frequency and time domains must be obtained. Useful parametric models of musical instruments are usually not known, so any parametric spectral analysis method will require estimation of an unwieldy number of parameters. Thus the approach for this investigation was to use a standard nonparametric spectral estimation method, the short-time Fourier transform (STFT).

1.5.1 Review of the Short-Time Fourier Transform (STFT) Analysis

The STFT has been used widely in the analysis of time-varying signals, such as speech and music [19]–[24]. The STFT takes a one-dimensional time-domain signal (amplitude versus time) and produces a two-dimensional representation (amplitude versus frequency

versus time). This can be expressed for time-sampled signals in discrete form [25]:

$$X(n, k) = \sum_{m=-\infty}^{\infty} w(n - m)x(m)e^{-j2\pi mk/L} , \quad (1)$$

in which: $x(m)$ is a signal defined for any sample time m, $w(m)$ is a low-pass impulse response (window) function defined for any m, L is the number of equally spaced frequency samples between 0 Hz and the sample rate (or 0 to 2π normalized radian frequency), and $X(n, k)$ is the discrete STFT of $x(m)$ at every sample time n at normalized radian frequency $2\pi k/L$.

This equation is called the STFT analysis equation because it describes the STFT in terms of the input signal $x(m)$. With time-variant input the STFT can be thought of as providing a series of "snap shots" of the signal spectrum obtained over some chosen time interval.

The infinite sum in the STFT analysis equation is actually finite in practice because the window function $w(m)$ is typically chosen to be real, with even symmetry about the origin (noncausal, zero phase), and nonzero only for a finite range of points centered about the origin (see Harris [26] for a description of various window functions).

For computational efficiency it is often useful to express the STFT analysis equation in the form of the discrete Fourier transform (DFT) so that a fast Fourier transform (FFT) algorithm can be used to perform the summation calculations. Using a change of variables $m = n = pL + r$ and exchanging the order of summation, the analysis equation can be written in terms of blocks L samples long,

$$X(n, k) = e^{-j2\pi nk/L} \sum_{r=0}^{L-1} \sum_{p=-\infty}^{\infty} w(-pL - r)x(n + pL + r)e^{-j2\pi pk} e^{-j2\pi rk/L} . \quad (2)$$

Defining

$$\tilde{x}(n) = \sum_{p=-\infty}^{\infty} w(-pL - r)x(n + pL + r) e^{-j2\pi pk} \quad (3a)$$

or since $e^{-j2\pi pk} = 1$ (for p and k integers), we have

$$\tilde{x}(n) = \sum_{\text{all } p} w(-pL - r)x(n + pL + r) . \quad (3b)$$

Then Eq. (2) becomes

$$X(n, k) = e^{-j2\pi nk/L} \sum_{r=0}^{L-1} \tilde{x}(n) e^{-j2\pi rk/L} \quad (4)$$

which can be recognized as the DFT of $\tilde{x}(n)$ multiplied by a linear phase shift term.

If we choose the window function $w(q)$ to be zero for $q \geq L/2$ and $q < -L/2$, and noting that $0 \leq r < L$, the expression for $\tilde{x}(n)$ is nonzero only under the following conditions on p and r:

$$p = 0 \quad \text{and} \quad \{0 \leq r \leq L/2\}$$

or

$$p = -1 \quad \text{and} \quad \{L/2 < r < L\}$$

giving

$$\tilde{x}(n) = \begin{cases} w(-r)x(n + r) , & \text{for } 0 \leq r \leq L/2 \\ w(L - r)x(n - L + r) , & \text{for } L/2 < r < L . \end{cases} \quad (5)$$

Thus we can compute $X(n, k)$ by generating the intermediate signal $\tilde{x}(n)$, performing the DFT (using an FFT algorithm, if desired), and compensating for the linear phase term $e^{-j2\pi nk/L}$. This process is depicted in Fig. 1.

In considering Eq. (4) we see that the formal definition of the STFT requires a series of overlapping DFTs to represent $x(n)$ at every time n. This overlap may seem unnecessary, considering that the original signal can be reconstructed exactly from the inverse transforms of concatenated nonoverlapping segments, that is, the discrete Fourier transform is perfectly invertible. This observation would be useful and reasonable if the only

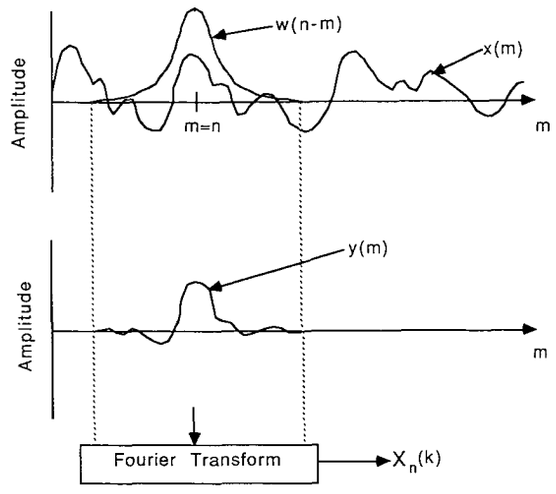


Fig. 1. The Fourier transform viewpoint of STFT. A segment of the digitzed signal $x(m)$ is multiplied by the reversed and shifted window function $w(n - m)$. The resulting signal $y(m)$ is processed by the discrete Fourier transform.

interest was in obtaining an identity analysis/synthesis procedure. However, for the duet separation problem (and for other tasks) it may be useful to interpret and modify the frequency-domain representation of the signal, which generally requires knowledge of the signal for every frequency index k at every time n. Fortunately, in practice the STFT can be performed every R samples of the input signal instead of every sample because the output sequence for a particular frequency index k is band-limited by the low-pass window function used in the analysis. This allows the STFT to be resampled at a lower rate than the signal sample rate. The frame spacing R can be called the analysis hop and must correspond to a rate $1/R$ at least twice the bandwidth of the analysis window in order to meet the Nyquist criterion, in this case applied to resampling the STFT. Choosing a smaller analysis hop size has the desirable effect of improving the time resolution of the sampled STFT, while choosing a larger hop reduces the frame rate, and thus the storage requirements and computation load. This resolution/computation tradeoff must be addressed to fit the needs of a given situation.

The STFT analysis equation, including an analysis hop, is given by

$$X(sR, k) = e^{-j2\pi sRk/L} \text{ DFT}\{\tilde{x}(sR)\} \tag{6}$$

where s is an integer.

1.5.2 Review of the Short-Time Fourier Transform (STFT) Synthesis

The synthesis equation corresponding to the STFT analysis equation, Eq. (4), can be expressed as an overlap-add (OLA) procedure,

$$\hat{x}(n) = \sum_{\text{all } m} \sum_{k=0}^{L-1} X(m, k)\, e^{+j2\pi nk/L} \tag{7}$$

and for the analysis hop case of Eq. (6),

$$\hat{x}(n) = \sum_{\text{all } s} \sum_{k=0}^{L-1} X(sR, k)\, e^{+j2\pi nk/L} \; . \tag{8}$$

The summation over k in Eq. (8) is almost the inverse DFT of $X(sR, k)$, namely,

$$\sum_{k=0}^{L-1} X(sR, k) = Lx(n)w(sR - n) \; . \tag{9}$$

Combining Eqs. (8) and (9),

$$\hat{x}(n) = \sum_{\text{all } s} Lx(n)w(sR - n) \tag{10a}$$

or

$$\hat{x}(n) = x(n)L \sum_{\text{all } s} w(sR - n) \; . \tag{10b}$$

The summation in Eq. (10b) is the sum at time n of copies of the window function $w(n)$ reversed and shifted by multiples of the hop size R. If the summation is constant and exactly equal to L for all s and n, the OLA process can exactly invert the STFT. Fortunately it can be shown [27] that any low-pass window function $w(n)$ which is band-limited to frequency $B = 1/2R$ satisfies the equation

$$\sum_{\text{all } s} w(sR - n) = \frac{1}{R} W(0) = \text{constant} \tag{11}$$

where $W(\cdot)$ is the Fourier transform of $w(\cdot)$. Of course, any time-limited window cannot be completely band-limited, so the hop size R must often be chosen based on some performance criterion. The scaling factor L can be included implicitly by scaling the time-domain window function prior to analysis, if desired.

The identity property of the STFT (the original signal can be resynthesized perfectly) implies that the analysis data contain all the information present in the original signal. This attribute is important because it theoretically allows processing to occur in either the time or the frequency domain; the most convenient representation can be chosen.

1.5.3 A Variation of the STFT Concept: MQ Analysis

McAulay and Quatieri [28] proposed an analysis/synthesis procedure for speech based on a sinusoidal representation. Their approach was to model speech waveforms as a sum of possibly inharmonic, time-varying, sinusoidal components. The basic McAulay and Quatieri (MQ) signal model assumes a priori that each segment of the input $x(n)$ consists of a finite number of sinusoidal components J. Each component may have arbitrary amplitude a_k, angular frequency ω_k, and phase p_k. Thus this model indicates

$$x(n) = \sum_{k=1}^{J} a_k \cos(\omega_k n + p_k) \; . \tag{12}$$

Like the STFT, the MQ process assumes that the parameters of $x(n)$ may be time-variant, so the amplitude, frequency, and phase parameters must be updated frequently to remain a valid representation of the input signal.

According to the original MQ analysis algorithm, the input signal is segmented into blocks of length N (possibly overlapping). Each block is windowed with an appropriate low-pass window (as in the STFT analyzer), and its discrete Fourier transform is computed via an FFT algorithm. For each DFT the magnitude spectrum is calculated, and all peaks in the spectrum are identified simply by searching for groups of three adjacent spectral samples where the magnitude's slope changes from positive to negative. McAulay and Quatieri assume that each peak may be attributed to the presence of an underlying sinusoidal component during the current segment of the input signal. Once all the peaks are determined, the complex (real, imaginary)

spectrum is used to identify the phase information for each peak. Finally, the amplitude, frequency, and phase parameters for each peak are stored in a data structure. The number of peaks chosen in each data frame can be limited by 1) choosing a fixed number of peaks or 2) imposing some amplitude threshold. The MQ procedure is depicted in Fig. 2.

One difficulty in the peak identification procedure is due to the limited density of frequency points resulting from the DFT. Indeed, the actual frequency of an underlying sinusoidal component may lie *between* the frequency samples of the DFT. This limitation can be reduced by increasing the density of the DFT frequency samples by means of a longer zero-padded DFT, and by using an interpolation method on the magnitude spectrum itself [29].

The analysis and peak identification process is repeated for successive frames of the input signal. As mentioned previously, the spacing R between the starting points of adjacent frames (the hop size) is chosen as a tradeoff between the computation expense associated with a small hop and the loss of time resolution due to a large hop. The hop size can also be selected according to the frequency domain sampling criteria described previously for the STFT analyzer. If desired, the hop size and analysis block length can be changed adaptively to follow changes detected in the input signal.

In most practical cases the signal spectrum presented to the MQ analyzer varies considerably with time so that the number of detected components and their frequencies will, in fact, change from frame to frame. For this reason a matching procedure is performed to connect components (peaks) from frame $[i]$ with corresponding ones from frame $[i + 1]$, thereby tracking

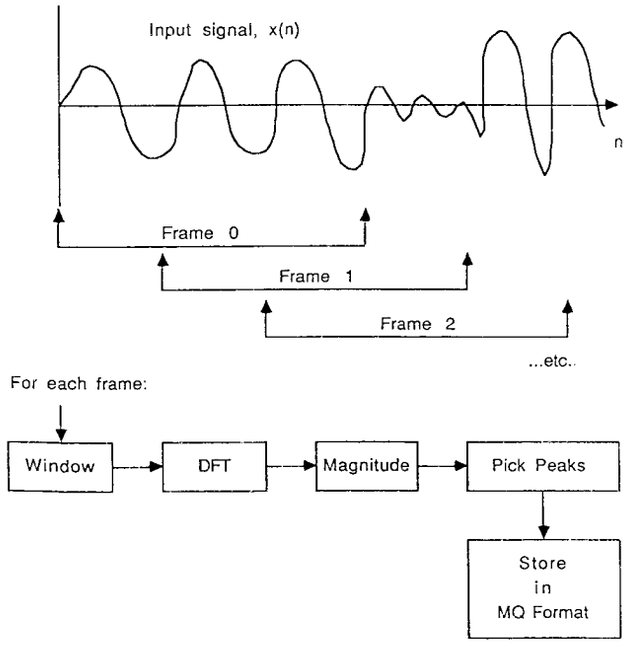


Fig. 2. McAulay–Quatieri (MQ) analysis procedure. The input signal $x(n)$ is segmented into overlapping "frames," which are windowed and the DFT is computed. The peaks in the magnitude spectrum of each frame are identified and stored.

the time-varying sinusoidal components.

Once the peak matching process on the current frame is complete, the procedure is repeated for the subsequent data frames. The final output database consists of chains of peaks, or tracks, which trace the behavior of the underlying sinusoidal components comprising the signal. An example is shown in Fig. 3.

1.5.4 MQ Synthesis

Synthesis from the MQ model can be accomplished via a simple additive procedure, where each of the J components is regenerated by a sinusoidal oscillator with amplitude, frequency, and phase modulation applied according to the analysis model parameters. However, extra care is required to "unwrap" the phase parameter due to the inherent range ambiguity of the principal value through the Fourier transform [arg (y) = arg $(y + 2\pi q)$].

Each frequency track is used to control a sinusoidal oscillator whose amplitude, frequency, and phase are modulated in such a way that they exactly match the measured values from the analysis at the frame boundary times and change smoothly between frames. Linear interpolation has been found to be adequate for the amplitude values but the frequency and phase values require more careful treatment [24], [28], [29].

The continuous-time frequency and phase functions are incapable of fully independent variation because they are related by the time derivative. The problem becomes one of choosing a phase interpolation function whose slope between a pair of linked peaks (instantaneous frequency) matches the measured frequency at the frame boundaries and whose phase corresponds to the measured phase value *unwrapped* to provide a smooth frequency function. A cubic phase function has been found to provide satisfactory results.

The MQ process can be extended to allow a wide range of signal analysis interpretations [30]. For example, the peak-matching and smooth-phase interpolation methods are useful for splicing and editing sound segments without clicks or pops [31]. Also, the frequency tracks obtained from the analysis step can be scaled for shifting pitch without changing the evolution of the sound with respect to time. Similarly, time compression and expansion without pitch change, filtering, smoothing, and other manipulations can be accomplished within the MQ model [32].

1.5.5 MQ Analysis/Synthesis Results

In many informal experiments accompanying this project, the MQ process was applied to isolated musical tones, speech, singing, and polyphonic music. With careful listening, the synthesis output was sometimes found to be perceptually distinguishable from the original signal. In particular, the character of noiselike components of the input signal was often altered in the synthetic sound, presumably due to an inadequate characterization of the noisy material by the sum-of-sinusoids MQ model [33]. On the other hand, this attribute of the MQ procedure shows some potential for

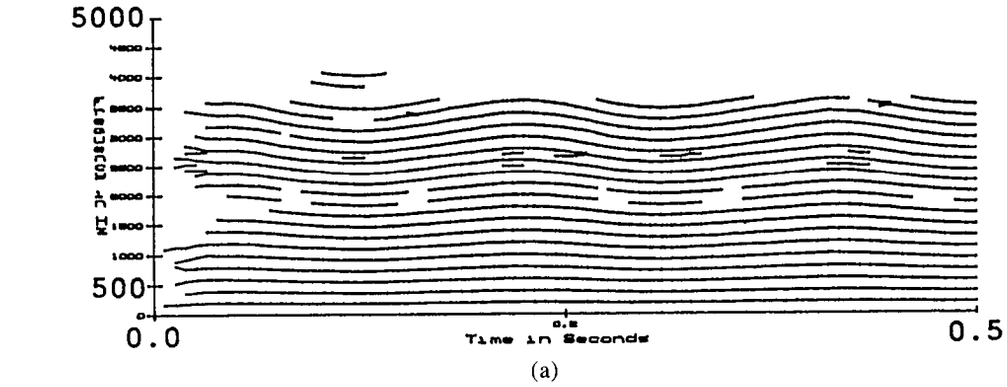
noise reduction of recordings, particularly if a careful choice of thresholds is made for the peak-picking step of the analysis. For the most part, the synthesis was not found to be "better" or "worse" than the original, only distinguishable. This informal result is encouraging because it indicates that the MQ model retains the essence of the original recording and, therefore, we may conclude that the information necessary for separating duets is present in the MQ analysis data. Further, the convenient data representation of the MQ procedure retains many of the useful features of the STFT.

2 DESCRIPTION OF SEPARATION PROCEDURE

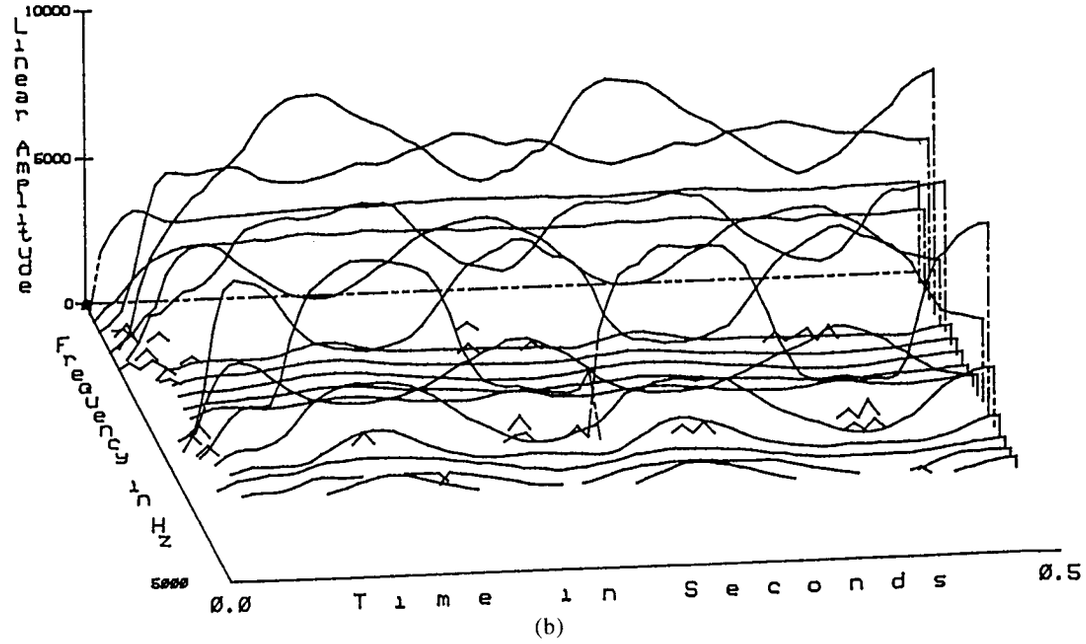
The duet separation approach considered in this paper requires estimates of the fundamental frequency of each voice. The pair of fundamental frequencies are used to determine the spectral energy distribution of the duet: if each voice is harmonic (one of the research assumptions), the composite magnitude spectrum should contain peaks at each harmonic frequency. Because of the inherent frequency selectivity limitations in the analysis, harmonics of the two voices which are spaced more closely than the resolution of the analyzer will "collide" and appear as a single, smeared peak in the magnitude spectrum. However, the colliding harmonic frequencies can be predicted from the fundamental frequency pair, thereby providing a means to deduce the contribution of each component to the composite spectrum. Thus accurate estimates of the fundamental frequencies in each frame are vital to the success of this procedure.

2.1 A New Duet Frequency-Tracking Approach: Two-Way Mismatch

The interpolated FFT method used in the MQ analysis provides samples of the signal spectrum at points equally spaced in linear frequency. In other words, the uncertainty associated with identifying the frequency of a particular spectral peak is constant at all frequencies. For example, a fixed resolution of, say, 2 Hz represents a significant fractional error at lower frequencies. The greater fractional resolution at higher frequencies implies that the frequencies of the upper partials of a signal can be used to improve the estimate of the lower partials and fundamental frequencies of the harmonic



(a)



(b)

Fig. 3. MQ signal representation of a tenor voice with vibrato. (a) Frequency versus time. (b) Same data, but frequency versus amplitude versus time.

signal.

For example, consider a sequence of spectral components: {110, 220, 330, 440} Hz (harmonics of a 110-Hz fundamental). However, due to the 2-Hz resolution limit the measured sequence could actually be {108, 222, 328, 442} Hz. Simply choosing 108 Hz as the fundamental results in a predicted harmonic sequence of {108, 216, 324, 432}, corresponding to a difference sequence between the prediction using 108 Hz and the measured sequence of {0, 6, 4, 10} Hz or, in relative terms, {0, 2.7, 1.2, 2.3}%. Note that if the "correct" fundamental frequency, 110 Hz, is chosen, the resulting error sequences are {2, 2, 2, 2}, or {1.9, 0.9, 0.6, 0.45}%. Thus the actual harmonic sequence should be the one with the *smallest total mismatch error* when compared to the measured sequence.

The frequency estimation task in the context of this paper is to choose a *pair* of fundamental frequencies which *together* minimize the mismatch between the predicted partial frequencies (harmonics of the two fundamentals) and the observed partial frequencies (from the MQ analyzer). The mismatch error is calculated as the sum of weighted, squared normalized differences between each predicted partial frequency and the nearest measured partial frequency *and* between each measured partial frequency and the nearest predicted partial frequency. Note that the mismatch error (predicted to measured) may be different from the reverse mismatch error (measured to predicted). The frequency differences are normalized by dividing the frequency difference by the predicted partial frequency. An empirically derived amplitude weighting function is applied to emphasize the contribution of stronger partials, whose frequency estimates are presumably more accurate. The mismatch weighting rules (from best match to worst match) can be summarized:

1) Missing a *large* amplitude partial by a *small* frequency difference

2) Missing a *small* amplitude partial by a *small* frequency difference

3) Missing a *small* amplitude partial by a *large* frequency difference

4) Missing a *large* amplitude partial by a *large* frequency difference.

This "two-way mismatch" calculation has two advantages: 1) it favors frequency choices which correctly predict the measured components and 2) it penalizes frequency choices which predict components that are not found in the set of measured partials. An example of the two-way mismatch calculation for a particular pair of estimated frequencies is given in Fig. 4.

The two-way mismatch procedure is provided with two nonoverlapping frequency ranges in which to concentrate its search. This frequency range information is supplied by the user from prior knowledge of the expected input signal. The task of the frequency tracker is to identify the pair of estimated fundamental frequencies which result in the minimum two-way mismatch error. This must be performed as an iterative search procedure because the global minimum of the mismatch error function will not, in general, be the only local minimum.

In the current implementation the search procedure first repetitively calculates the error values for a frequency pair with the upper voice frequency fixed, and the lower voice frequency increasing from the minimum of the user-supplied low-frequency range to the maximum in semitone increments. When a local minimum is detected using the semitone increments, the region of the minimum is processed iteratively with a decreasing frequency increment to refine the true estimate of the minimum location. Note that the initial semitone steps do not constitute the final resolution of the procedure: the iterative search improves the estimate, thereby allowing for frequency tracking of vibrato, glissandi, nonstandard tunings, and so on.

Once the entire low-frequency range is processed and the overall minimum is obtained, the search procedure continues for a frequency pair with the lower voice frequency set to the "best" frequency obtained in the first step, and the upper voice frequency increasing in semitone steps across the high-frequency range. The resulting global minimum pair for the entire frame is saved.

In practice this global search is only performed several times per second of the input signal. On the frames between the global search frames, the search is restricted to a semitone range ($\pm 2.5\%$ of each fundamental frequency). If the global search turns up a better pair of frequencies outside the semitone range, the tracking process must back up to the preceding global search frame and check the intervening frames to isolate the frame at which the change occurred. By choosing the period between global searches to be less than an arbitrary minimum note duration, no frequency transitions (for example, note changes) will be missed. This approach uses the assumption that the frequency pair will

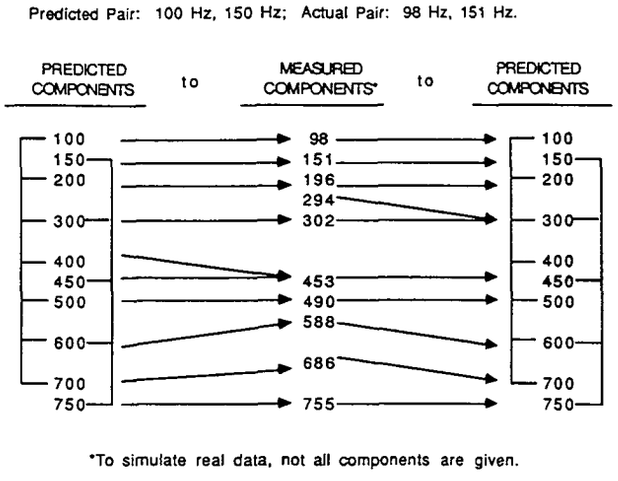


Fig. 4. Example of two-way mismatch (TWM) error calculation. Center column contains the peaks obtained from MQ analysis of one frame of input signal. Arrows indicate two-way nearest-neighbor error calculation for predicted set of duet harmonics at one step in iteration. Note that two-way errors (predicted to measured and measured to predicted) are not the same.

often remain roughly constant for many analysis frames during each musical note. The computation savings of this method is, of course, governed by the degree to which this assumption holds for a particular duet.

If the analysis data were noise free and the partials of the two voices did not interact, we might only attempt to maximize the number of correct predictions (the coverage) of the measured frequencies, then choose the highest octave for a given level of coverage. However, any spurious frequencies due to noise or interference between the duet voices could cause this estimate to be extremely sensitive to small variations. The two-way mismatch approach lessens the impact of these errors by including "bad" predictions as an unfavorable parameter in the error calculation. Of course, the presence of background sounds, such as the resonation of undamped strings in a guitar or harp recording, could seriously degrade the usefulness of the two-way mismatch method. This situation has been avoided (conveniently) in this study by the a priori assumptions made in Sec. 1.2.

2.2 MQ Formulation for Separation Task

At first glance the MQ sum-of-sinusoids model appears to be an ideal representation for the duet separation problem because the sinusoidal components of each voice should appear as independent, harmonic tracks in the MQ analysis data. By this naive reasoning, the two voices could be separated and resynthesized by identifying the pair of fundamental frequencies and simply choosing which frequency tracks went with which voice. Unfortunately the limited frequency resolution of the MQ analyzer causes any closely spaced components to "collide" and appear as a single, broad peak instead of two separate peaks.

In order to resolve spectral collisions between the partials of different voices we must ascertain the contribution of each voice to the composite information observed in the short-time spectrum. Given that predictions of the two fundamental frequencies are known (using the two-way mismatch procedure), it is possible to identify conflicting partials by comparing the predicted harmonic series of the two voices. Partials with spacing greater than the resolution limit of the short-time analysis can simply be segregated into groups belonging to one or the other voice of the duet. Partials with spacing less than the resolution of the analyzer will appear as corrupted data due to the crosstalk between the two partials.

Three approaches for separating closely spaced components are considered.

1) A set of linear equations can be specified and solved for the contribution of each windowed sinusoid to the observed complex spectrum in the vicinity of the conflict.

2) The amplitude modulation (beats) and frequency modulation functions due to the closely spaced partials may be used to calculate the amplitudes of the colliding partials (assuming the amplitudes and frequencies remain relatively constant for a period of time sufficient to estimate the various parameters involved).

3) If an accurate signal model is known for each voice, colliding partials can be handled by synthesizing artificial amplitude and frequency functions to replace the corrupted data. If an adequate model is not known, some form of interpolation or spectral templates may be used. This process should ideally be inaudible, that is, we wish to retain the timbre and performance character of the performance.

2.2.1 Separation Strategy I: Linear Equations Solution

If the input signal is perfectly harmonic during one analysis frame, its short-time spectrum is a frequency-domain convolution of the analysis window spectrum and a series of weighted impulse functions at frequencies corresponding to the harmonic partials of the signal. This convolution produces a short-time spectrum consisting of overlapped replicas of the window function spectrum centered at each harmonic frequency and scaled by the amplitude of the harmonic component at that frequency. The amplitude and phase of the short-time spectrum at a particular frequency is the complex sum of all the overlapped window spectrum contributions at that frequency. This "crosstalk" can be minimized (but not eliminated) by appropriate choice of the analysis window function.

The spectral width of the main lobe of the analysis window spectrum is usually chosen so that no spectral overlap occurs between adjacent harmonics for the lowest fundamental frequency of interest (closest harmonic spacing). However, the spectral collisions typically present in an analyzed duet involve two components (one from each voice) which may be spaced more closely than the no-overlap condition. Assuming a good analysis window with very low sidelobe levels, it is reasonable to consider only the contributions from the two colliding components and neglect the contributions of the other spectral components. Since the shape of the analysis window spectrum, the spacing of the two colliding frequency components, and the spectral values at the two frequencies are known—and neglecting all contributions except the two colliding components—a pair of linear equations can be determined for the two unknown quantities: the actual amplitude of the two spectral components without overlap.

Denoting the two colliding frequencies by ω_1 and ω_2, the known composite complex short-time spectra by $G(\omega_1)$ and $G(\omega_2)$, the analysis window spectral magnitude by $W(\omega_1 - \omega_2)$ [normalized, $W(0) = 1$], and the desired complex component spectra by $G_1(\omega_1)$ and $G_2(\omega_2)$, we have

$$\begin{aligned} G_1(\omega_1) &= \alpha_1 - W(\omega - \omega_1) \\ G_2(\omega) &= \alpha_2 - W(\omega - \omega_2) \\ G(\omega_1) &= G_1(\omega_1) + W(\omega_1 - \omega_2)G_2(\omega_2) \\ G(\omega_2) &= G_2(\omega_2) + W(\omega_1 - \omega_2)G_1(\omega_1) \end{aligned} \tag{13}$$

where α_1 and α_2 are the underlying amplitudes of the colliding partials.

The unknown complex quantities $G_1(\omega_1) = \alpha_1$ and $G_2(\omega_2) = \alpha_2$ can be solved from this pair of equations [Eqs. (13) are complex, but the real and imaginary parts may be computed separately]. Thus estimates of the amplitude and phase of any pair of closely spaced partials can be obtained. A schematic representation of this process is depicted in Fig. 5.

Unfortunately the simple linear equations approach suffers from several deficiencies. First the "known" quantities (the window spectral shape and the two colliding frequencies) are really *not* known with great accuracy in practice because 1) the input signal is seldom truly periodic, 2) the pair of linear equations becomes singular (both equations become the same) as the frequency between the components gets small, that is, as $W(\omega_1 - \omega_2)$ approaches unity, and 3) collisions between two components with very different magnitudes can amplify the effect of parameter errors and provide unsatisfactory results for the lower amplitude component. In summary, the quality of the linear equations solution depends on the degree to which the signal assumptions are met.

As noted, the linear equations solution becomes singular as the frequency spacing between colliding components decreases. The frequency spacing of the colliding components (based on the two-way mismatch fundamental frequency tracking data) can be observed to determine whether the linear equations approach is applicable. If not, another separation strategy must be applied.

It should be noted that Danisewicz and Quatieri [34] independently proposed a similar linear equations solution method (for cochannel speech) which includes the effects of all shifted window transforms, not just the nearest two, as described here. They also provide an interesting interpretation of the frequency-domain linear equation solution as an equivalent time-domain least-squares viewpoint.

2.2.2 Separation Strategy II: Analysis of Beating Components

When the frequency separation of two partials becomes smaller than the resolution of the STFT analysis, the MQ analyzer does not see two distinct peaks, but a single peak with amplitude and frequency modulation: the two components exhibit "beating." In other words, the two colliding components appear as a single component with sinusoidal amplitude modulation occurring at a best frequency equal to the frequency difference between the colliding components.

If the frequencies (ω_1, ω_2) and amplitudes (A_1, A_2) of the colliding partials remain constant for one or more beat periods, the composite amplitude and frequency functions can be used to estimate the amplitudes of the colliding components, as depicted in Fig. 6. The maximum of the amplitude beat is $A_1 + A_2$, the minimum is $|A_1 - A_2|$. If the amplitude minimum occurs when the composite frequency is a minimum, the amplitude of the lower frequency partial is $\max(A_1, A_2)$, while if the amplitude maximum occurs when the frequency is minimum, the lower frequency partial's amplitude is $\min(A_1, A_2)$.

Successful use of component beating functions to solve the partial collision problem requires that the duet voices contain no significant amplitude and frequency fluctuations of their own. Long observation times may also be required to identify the best amplitude maximum and minimum since the beat period is inversely related to the frequency spacing of the colliding partials. This probably eliminates the use of this strategy for notes with duration less than the beat period and for notes with significant vibrato, tremolo, or other amplitude–frequency modulation.

2.2.3 Separation Strategy III: Use of Models and Spectral Templates

Separation strategies I and II require the colliding partials to behave in a "nice" manner, meaning that their parameters remain relatively constant during the entire time interval of interest. Strict assumptions of this sort do not always hold true in the real world, however, and errors in the separation procedure may be audible in the final result. Thus an alternative strategy incorporating a priori knowledge of the characteristic behavior of the competing input signals must be developed. These signal "models" can be matched to the composite duet signal in order to refine the separation results of strategies I and II, or to generate reasonable predictions of signal behavior to replace the uncertain data present during spectral collisions.

Although signal modeling initially appears to be a promising method for improved separation of cochannel signals, it requires solutions to numerous system design and implementation problems. Musical instruments are generally extremely complex acoustomechanical or electroacoustomechanical systems which defy simple characterization by a small number of parameters. Some preliminary work reported by Wold and Despain [35] indicates that as many as 300 separate states must be estimated to describe a single clarinet tone for iden-

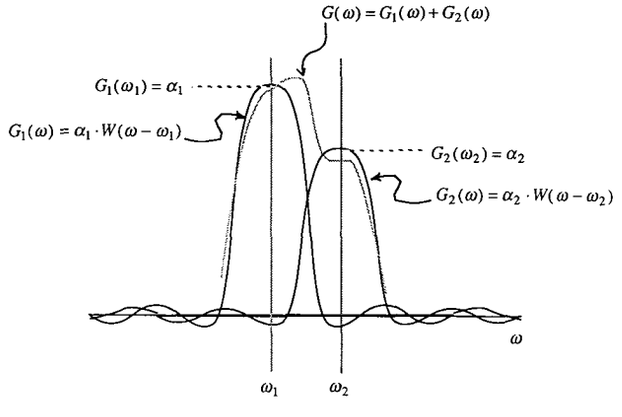
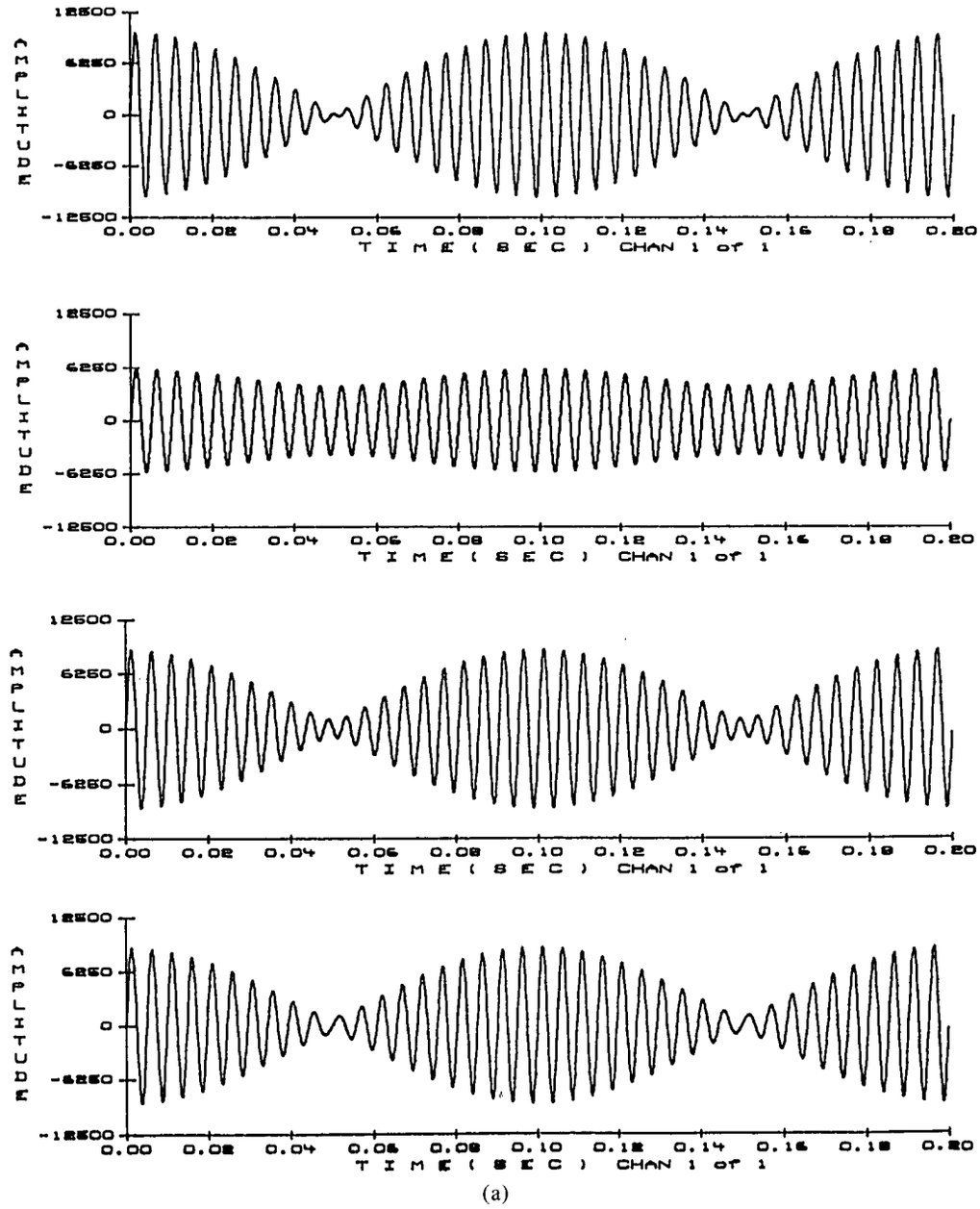


Fig. 5. Overlapped spectra for two closely spaced partials (real part). Solution by linear equations method identifies amplitudes of two partials, given measured composite spectrum (dotted line) and pair of partial frequencies.

tification and separation. Even if a satisfactory model of the sound production properties of an instrument is available, most recordings are made in reverberant surroundings or with some form of electronic signal processing and equalization which alters the temporal and spectral properties of the recorded signal. Further, a model capable of synthesizing a sound simply perceived to be a natural timbre may not be capable of matching the performance characteristics of a particular human musician. For these reasons the task of exactly modeling a recorded musical instrument sound is quite difficult.

The preliminary modeling approach used in this investigation does not approach the level of sophistication indicated by the preceding discussion. Instead, a simpler approach was evaluated to determine whether a straightforward modeling concept might provide satisfactory results. The modeling method chosen was to determine a set of spectral templates for the voices of interest. Each template describes a characteristic spectral envelope, that is, the relative amplitude of each partial in the spectrum of a constant musical note played by the instrument. The templates are normalized so that the amplitude of the strongest partial is unity. If several of the partials of the musical voice are uncollided, the template can be matched to the known partials and the collided partials can be estimated using the template as a look-up table.

The template algorithm can be described as follows (see Fig. 7). First the measured amplitude Q_i of each uncorrupted partial number i is obtained from the short-time spectrum. Next the total squared error between



(a)

Fig. 6. (a) Time-domain beating waveforms for two closely spaced sinusoids given by $x(t) = A_1 \cos(\omega_1 t) + A_2 \cos(\omega_2 t)$. (b) Instantaneous frequency functions for beating waveforms of 6(a).

the Q_i and the corresponding scaled template values, the GT_i, is defined according to

$$E_{\text{total}} = \sum_{\substack{i=1 \\ i \neq k}}^{J} (Q_i - GT_i)^2 \tag{14}$$

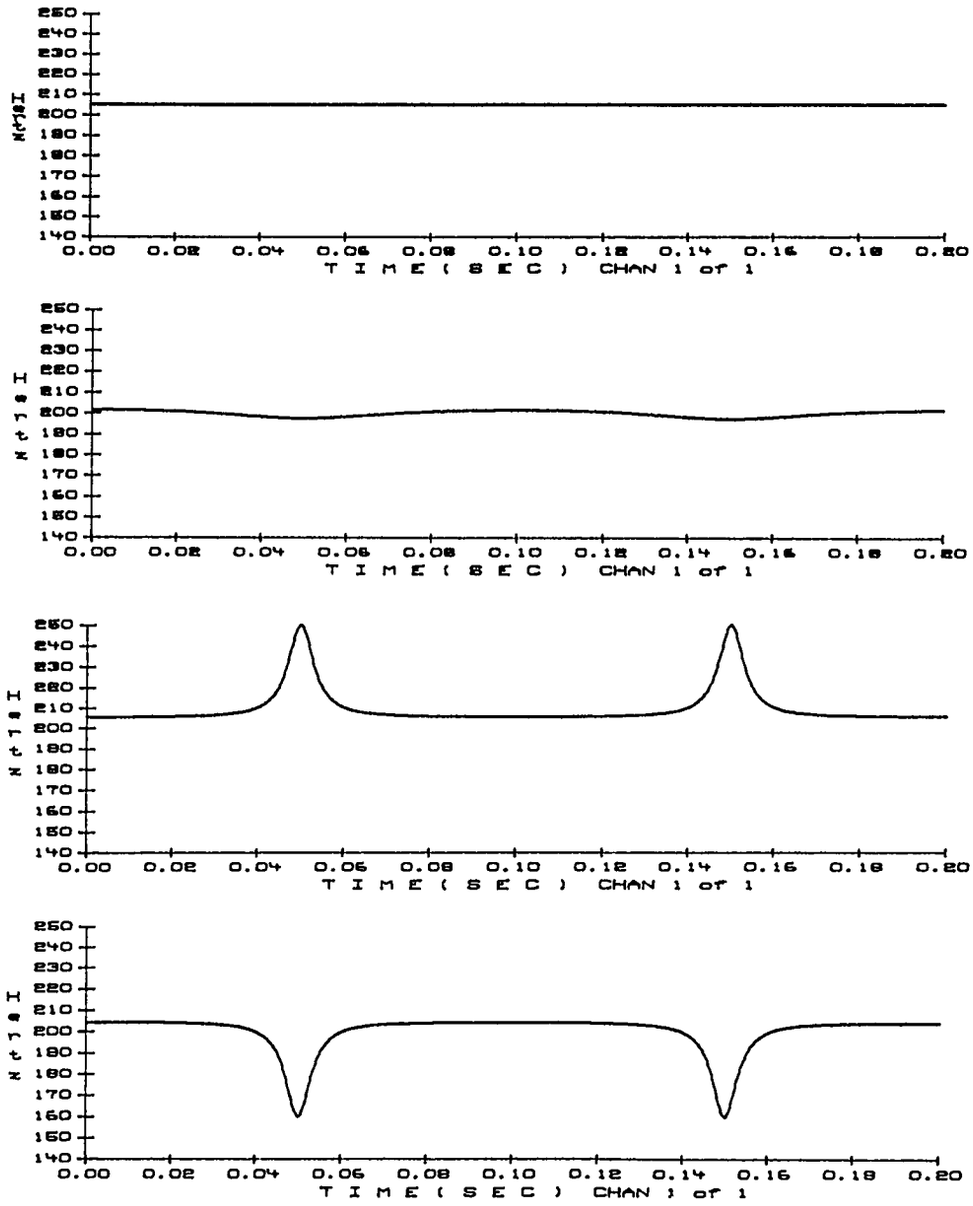
where partial number $i = k$ is corrupted by a spectral collision. This error expression is minimized with respect to the template amplitude scaling parameter G, giving

$$G = \left\{ \sum_{\substack{i=1 \\ i \neq k}}^{J} Q_i T_i \right\} \Big/ \left\{ \sum_{\substack{i=1 \\ i \neq k}}^{J} T_i^2 \right\} . \tag{15}$$

Finally, the estimated amplitude Q_k of the corrupted partial k is given by

$$Q_k = GT_k . \tag{16}$$

Hidden in this simple description is the complex task of choosing the appropriate template to use in a given situation—and even whether a template-based approach is reasonable in general. The use of several isolated "target" notes from an instrument in order to derive templates belies the true nature of the analysis task. Most real musical instruments exhibit both intentional and unintentional variation from note to note and from style to style during a performance. Indeed, these nuances convey the skill and emotion of the performer, representing the essence of a good musical performance.



(b)

Fig. 6. (*Continued*)

Retaining the expressive quality of the two competing voices is a goal of the signal separation problem, but a small set of templates is unlikely to capture these subtleties.

For these and other reasons, the experiments using the template strategy were limited to a small number of templates obtained from a studio recording of a soprano singer. Unfortunately the results obtained in those experiments were not sufficiently encouraging to merit further investigation. Thus the use to templates was reluctantly abandoned for the examples considered in Sec. 3.

Instead we investigated another simple spectral repair strategy applicable when all the spectral components except one are uncollided in a particular frequency range. If the spectral envelope of the voice is relatively smooth, or if the fundamental frequency is low, the corrupted partial's amplitude can be estimated by interpolating from the surrounding uncorrupted partials, as sketched in Fig. 8.

2.2.4 Multistrategy Approach

In the current implementation the choice of the appropriate separation strategy for a pair of colliding partials is as follows (using a Kaiser STFT analysis window with 6-dB bandwidth of 40 Hz).

1) The two fundamental frequencies of the duet (obtained using the two-way mismatch procedure) are used to generate the harmonic series of the two voices. The spacing between all adjacent partials is calculated.

2) If a partial is at least 50 Hz away from every other partial, the component is considered uncorrupted and no collision repair occurs.

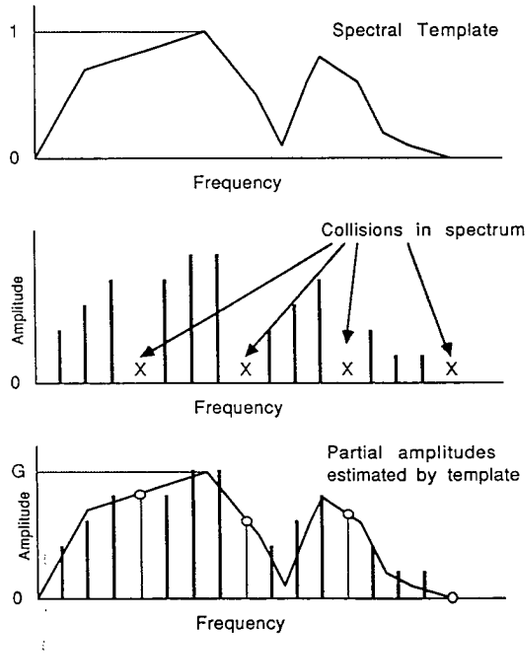


Fig. 7. Example of possible use of spectral templates to resolve collisions. Normalized template is scaled to minimize error between template function and measured (uncollided) partials. Collided partials are filled in by template value at each harmonic frequency. This approach proved ineffective due to the difficulty in constructing and selecting template functions.

3) If two partials are separated by less than 50 Hz but more than 25 Hz, the linear equations solution (strategy I) is applied.

4) If two partials are separated by less than 25 Hz, the beating analysis (strategy II) is attempted. However, if the collision is found to be less than two or three beat periods ($T < 2/|f_1 - f_2|$), estimates of the beating parameters are not reliable. In this case the spectral interpolation solutions (strategy III) are applied.

In the case where the fundamental frequency of one voice is an integer multiple (such as an octave) of the other, *all* harmonics of the upper voice coincide with partials of the lower voice. Extraction of the upper voice may become impossible when this occurs. The lower voice, on the other hand, may have at least some of its partials uncorrupted because the partials of the upper voice will be spaced by two or more times the harmonic spacing of the lower voice. In this situation the lower voice spectrum may allow reconstruction using the separation strategies considered in the preceding. Note that the upper voice might be recoverable if the spectral envelope of the lower voice's partials happens to be confined to a frequency range below the partials of the upper voice so that no (or few) collisions occur.

The case of unison duets, although implicitly excluded from this study by the nonoverlapping range assumption of Sec. 1.2, would be the most difficult to handle using the current approach. Separation of unisons might only be possible through the use of differing vibrato characteristics or other subtle pitch differences between the two voices.

2.2.5 Further Considerations

The primary difficult with a multistrategy approach is handling transitions between strategies as the cochannel input signals vary. Because the duet voices are generally independent, a note from one voice may start or stop while a note from the other voice is sustained. Thus the spectral collision situation may change rapidly as the voices enter and exit. A similar problem occurs during glissando or vibrato: spectral collisions may vary even during an ostensibly sustained note.

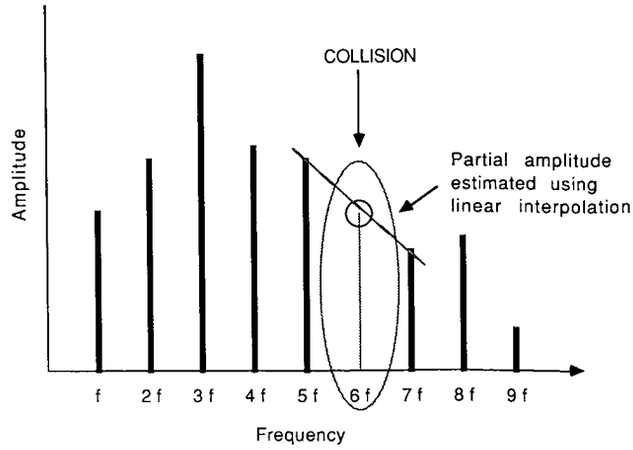


Fig. 8. Example of use of spectral interpolation to resolve collisions.

Simply switching from one separation strategy to another as the spectral collisions vary may result in audible discontinuities in the output signals due to estimation differences among the separation methods. The solution to this problem involves a layer of continuity comparisons between the results of the different separation strategies, particularly when a change occurs in the collision status.

3 RESULTS AND EVALUATION

3.1 Testing and Evaluation Outline

The testing approach for this project involves both acoustically generated (real) signals obtained from musical recordings and artificial signals generated by software. The real signals provide examples of practical interest, while the synthetic signals help to define performance limits using known parameters. Most of the processed duet segments were 20 s or less in duration to conserve disk storage and processing time. Four representative results (two artificial inputs, two real inputs) selected from the many test inputs are described next.

3.1.1 Artificial Input Examples

Synthetic example duet 1 (Fig. 9) contains one voice with a constant fundamental frequency of 800 Hz for a duration of 1 s and another voice with a linear fundamental frequency ramp from 1200 to 880 Hz over a duration of 1 s. The constant-frequency voice contains six harmonic partials with equal amplitudes, while the changing-frequency voice contains six partials with somewhat arbitrary amplitude weightings {1, 0.5, 0.33, 0.25, 0.2, and 0.266}. Both voices have equal peak waveform amplitudes. The signal was designed to test the collision detection/correction ability of the separation procedure and the behavior of the frequency tracking process for piecewise constant and rapidly changing fundamental frequency pairs.

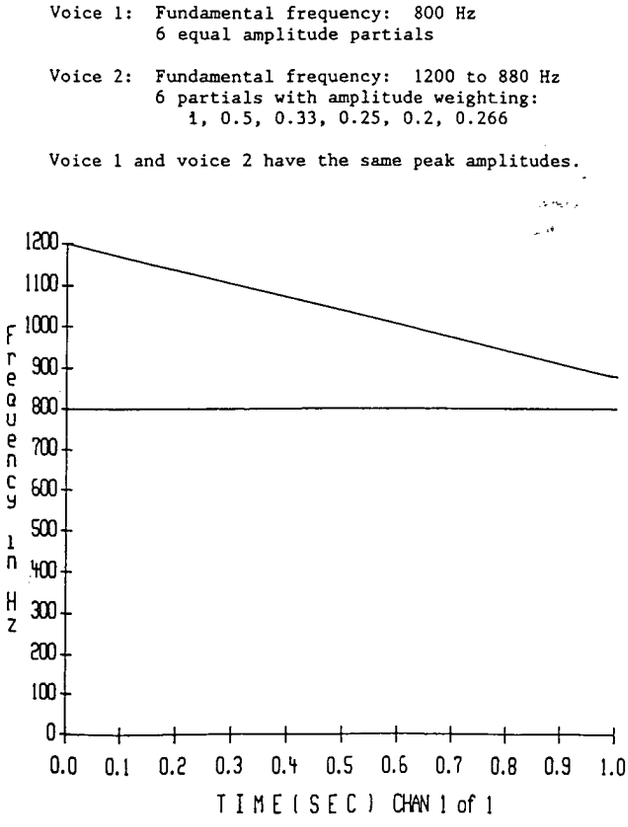


Fig. 9. Artificial duet test example 1. Synthesis frequencies and partial amplitudes.

Synthetic example duet 2 (Fig. 10) contains one voice generated using phase modulation and the other voice generated with the same amplitude for all partials. Phase modulation was chosen as a difficult case since phase-modulated tones can contain partials with distinctive, independent amplitude envelopes. Note that boundaries of the two voices occur at different times to allow evaluation of the transition capability of the frequency tracking and separation procedures.

3.1.2 Acoustic Input Examples

The first recorded duet (Fig. 11) is a short segment of "Duo Number 1 for Clarinet and Bassoon" by Ludwig van Beethoven, obtained from an analog LP record album. The example was chosen to test the system in the presence of typical reverberation, surface noise, and other distortion.

The second recorded duet (Fig. 12) is a segment of a tuba and trumpet duet from an analog LP recording of "Sonatina for Trumpet and Tuba" by Anthony Iannaccone. This example was chosen to check the tracking and separation procedures for widely separated voices and in the presence of background noise and reverberation.

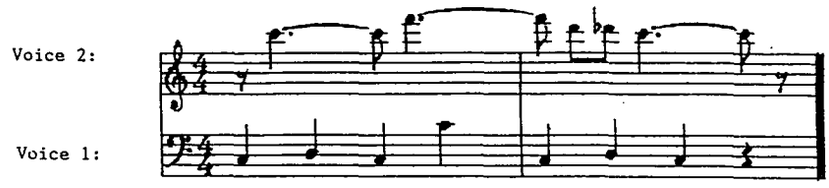
In the introduction to this paper, one of the stated assumptions was that only nonreverberant recordings should be processed. However, due to the pervasive presence of reverberation in recorded music, it was important to determine whether this restriction could be violated without compromising the overall quality of the separation procedure.

3.2 Evaluation of Two-Way Mismatch (TWM) Duet Fundamental Frequency Tracking Using Synthesized Duets

The raw MQ analysis and TWM frequency tracking results for artificial example 1 are shown in Fig. 13. The partials of the two voices are clearly visible in the output of the MQ analyzer, and the TWM algorithm has no difficulty in following the gross characteristics of the two fundamental frequencies. However, a specific examination of the fundamental frequency trace for the upper voice identifies occasional short-term errors (all less than 1%). Comparing the frequency trace with the MQ analysis output reveals that the frequency errors occur at points where partial collisions take place, disrupting the harmonic series of the voice. The TWM algorithm has some immunity to the collision problem due to its "best match" criterion, but the amplitude and frequency fluctuations present during a partial collision still cause some uncertainty.

The MQ and TWM results for artificial example 2

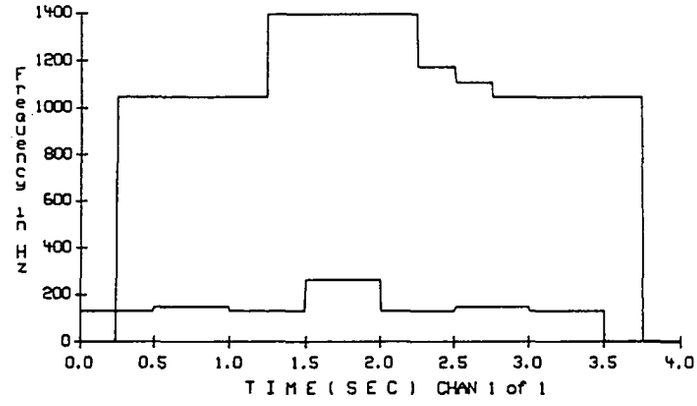
are shown in Fig. 14. The TWM tracking results match the true values quite well during portions of the duet where both voices are present, but the tracker output fluctuates when only one voice is present. This result is less surprising when one considers the duet assumptions built into the TWM algorithm, that is, the algorithm "looks" for two sets of harmonic peaks in every frame of the MQ analysis data. During solo passages (or between staccato notes), only one set of harmonic peaks is found in the MQ output and the TWM algorithm must switch to a solo tracking strategy. The result is a tendency to skew slightly at the transistions between



(a)

Voice 1: Phase modulation synthesis,
carrier/modulator ratio = 1:1, index = 4
Peak amplitude = 15000

Voice 2: 7 equal amplitude partials
Peak amplitude = 10000



(b)

Fig. 10. Artificial duet. (a) Test example 2, musical score. (b) Test example 2, frequency specification.

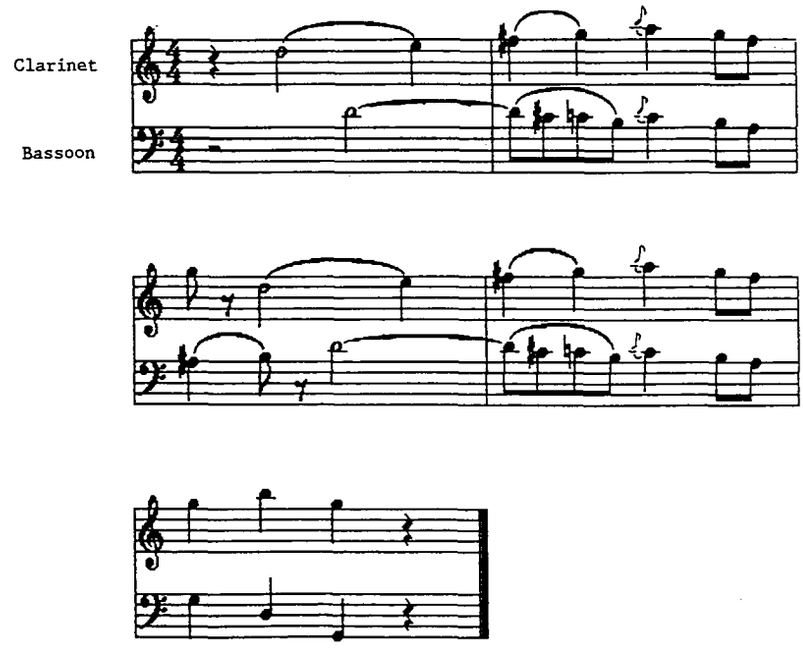


Fig. 11. Real duet test example 1: Ludwig van Beethoven, "Duo Number 1 for Clarinet and Basson."

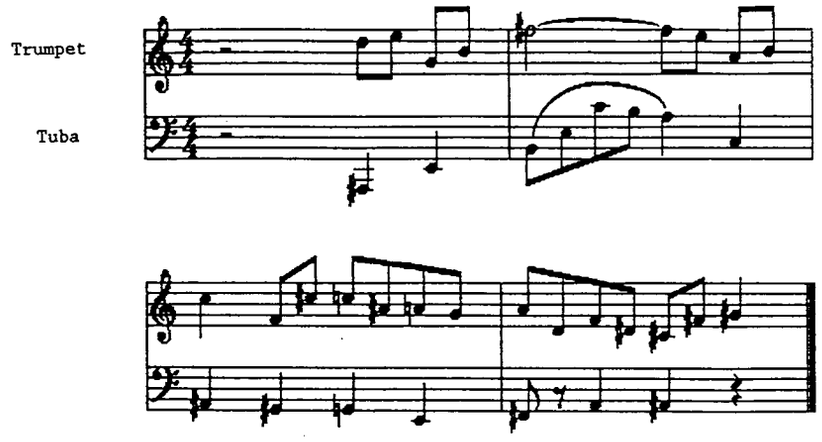
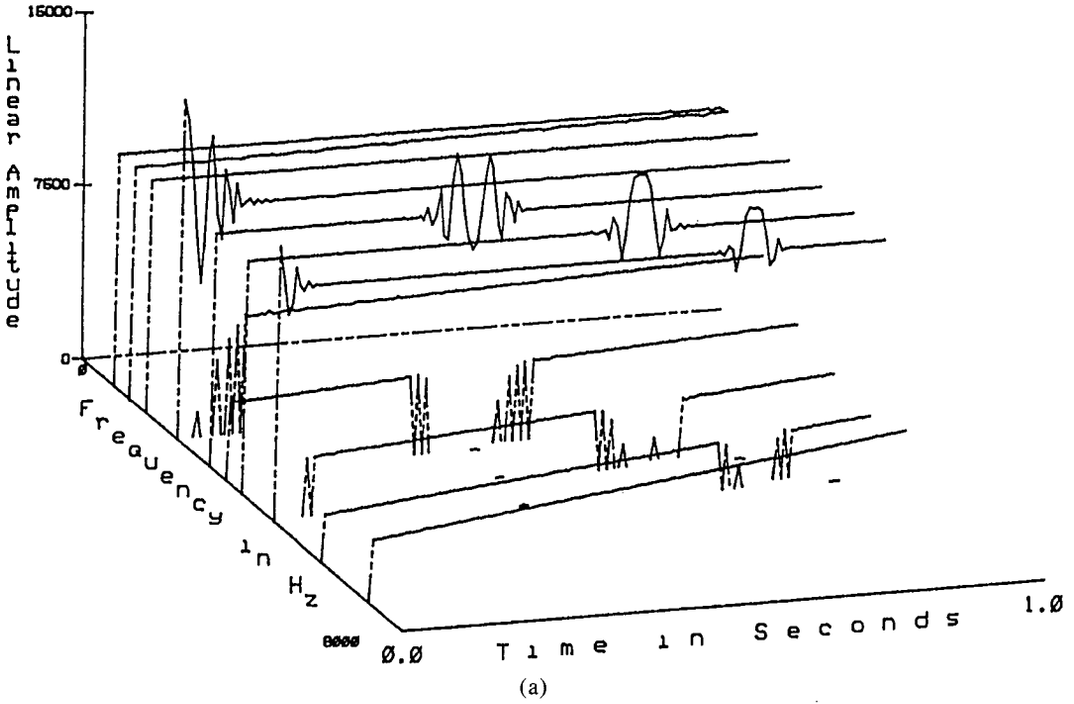
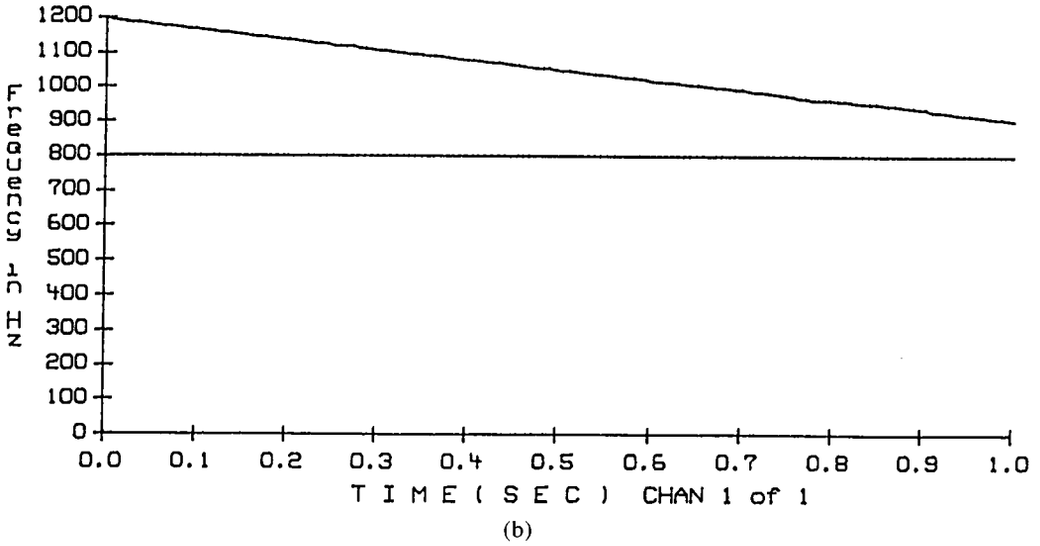


Fig. 12. Real duet test example 2 (hand transcription): Anthony Iannaccone, "Sonatina for Trumpet and Tuba."



(a)



(b)

Fig. 13. Artificial duet example 1. (a) Raw MQ analysis. (b) TWM duet frequency tracking results.

duet and solo modes. Several methods to overcome this deficiency are under development.

The following conclusions may be drawn concerning the performance of the TWM frequency tracking algorithm based on these and other examples.

1) The TWM procedure works very well for duet signals with constant or slowly changing fundamental frequencies and similar partial amplitudes.

2) The tracking process has difficulty interpreting the input signal when staccato notes occur in one or both voices or during other transitions between the duet and solo paradigms.

3) The frequency tracking process may be impaired if the list of component frequencies supplied by the MQ analyzer is seriously corrupted by noise, a level mismatch between the two voices, or some other degradation.

3.3 Evaluation of Voice Separation with Known Fundamental Frequencies

The separation procedure was first supplied with the known fundamental frequency data for the synthetic duets (instead of the TWM output) in order to determine the best-case performance of the isolated system.

The results for artificial test example 1 are shown in Fig. 15. Note that even with the a priori frequency information the separation process is not perfect. The amplitude discrepancy (amplitude "bumps") between the extracted voices of Fig. 15 and the constant partial amplitudes of the original voices can be traced to one of the underlying assumptions of the separation process, namely, that the peaks in the short-time spectrum are simply shifted and scaled copies of the Fourier transform of the analysis window function. This assumption is

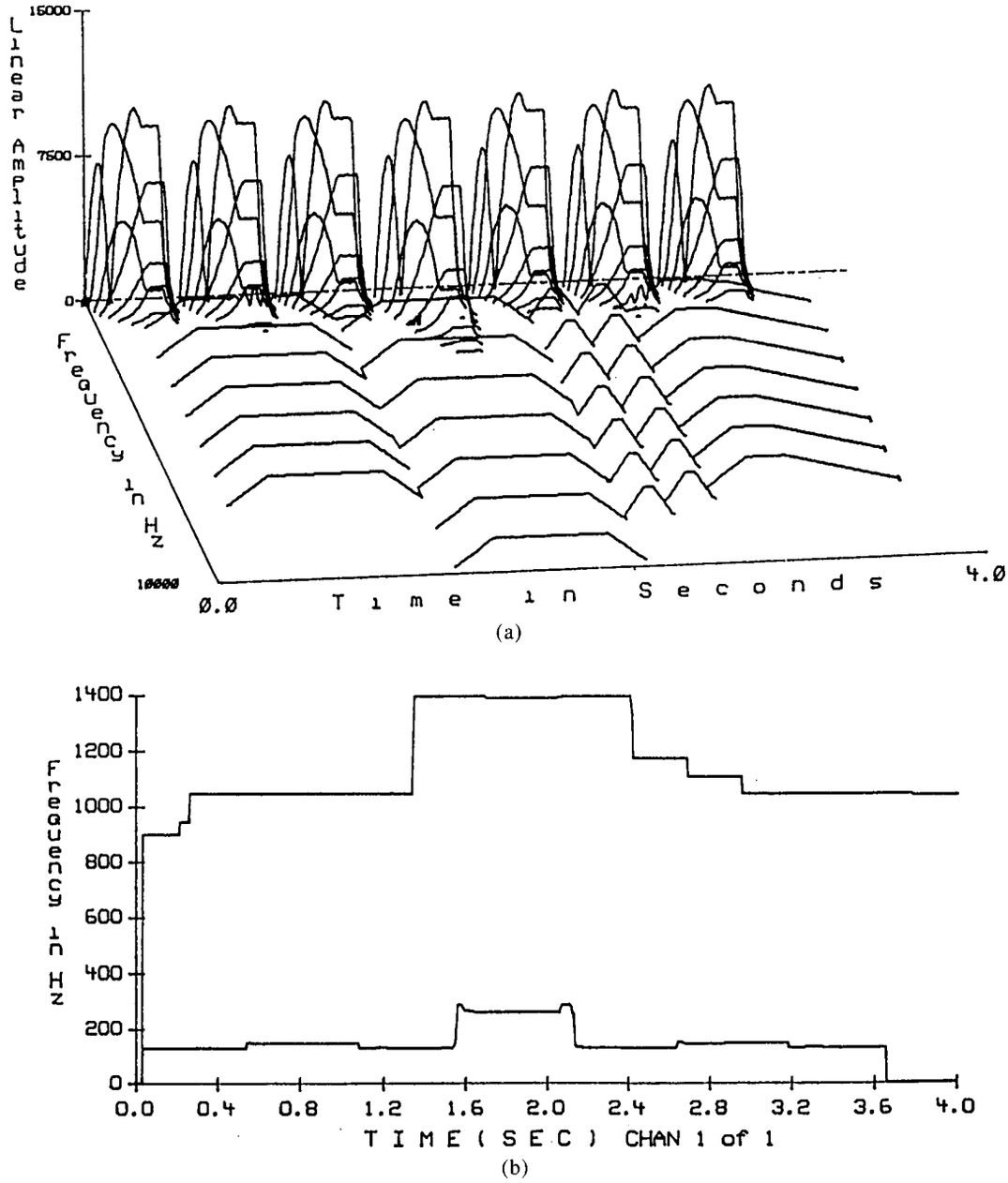


Fig. 14. Artificial duet example 2. (a) Raw MQ analysis. (b) TWM duet frequency-tracking results.

exactly correct only if the sinusoidal components comprising the signal *do not change their frequency or amplitude during the interval covered by the analysis window*. The fundamental frequency of the glissando voice in this example changes from 1200 to 880 Hz ($D = 320$ Hz) in 1 s, or 0.32 Hz/ms. The MQ analysis window used in this example is 25.6 ms in duration, yielding a frequency change during the window interval of 8.192 Hz for the fundamental. Also note that because the overtones are multiples of the fundamental frequency, the frequency of the second partial changes 16.384 Hz during the window interval, the third partial changes 24.576 Hz, and so on. Thus the assumption of constant partial frequencies—and therefore constant spectral window shape—is clearly violated in this case. Unless an explicit prediction of the short-time spectrum is made for every swept-frequency component identified in every input frame, the linear equation solution strategy is not a truly valid approach for this example.

The main effect of rapidly changing frequencies observed in the short-time spectrum is convolutional broadening of the peak corresponding to the changing frequency component, or even FM sidebands in extreme cases. The increased frequency extent of the spectral peak (particularly for upper partials) increases the likelihood of a collision between the broadened peak and any adjacent components, requiring extra care in the separation process. The underlying issues here is the fundamental time–bandwidth limitation of short-time analysis techniques.

The second artificial duet contains several notes where the voices are in octave alignment but with little spectral overlap. These results are shown in Fig. 16.

Interpretation of the separation results for these and

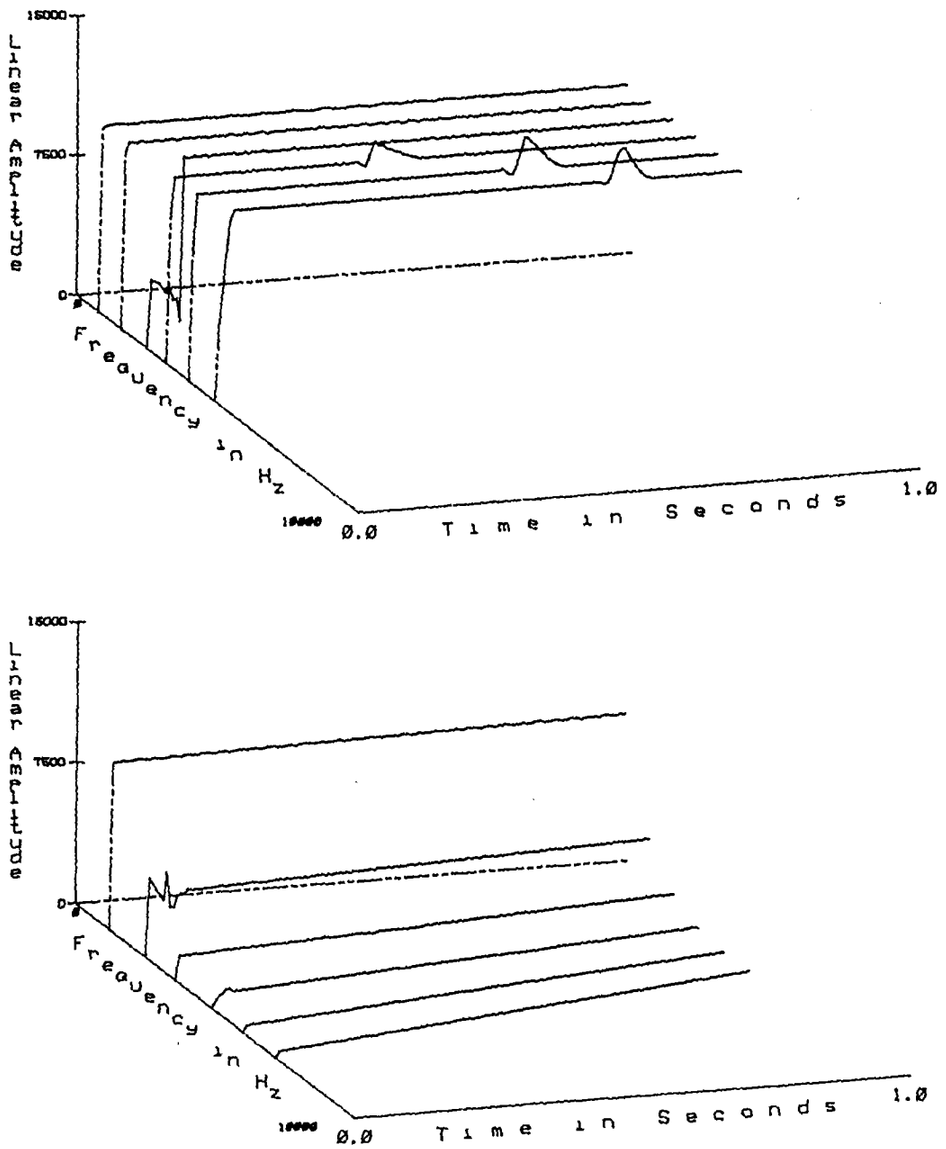


Fig. 15. Separation results for artificial duet example 1 using a priori frequency information.

other examples indicates the following conclusions.

1) The voice separation process is most effective for time intervals in which the fundamental frequencies of both voices remain constant.

2) The separation results may not be perfect—even in the best-case situation of a priori frequency knowledge. The discrepancies are primarily attributable to the time–bandwidth limitations of the short-time Fourier transform used in the analysis.

3) Reliable separation requires frequency estimates of each partial to be within a few hertz of the true value in order for accurate collision prediction and repair. If the estimate of the fundamental frequency of a voice is in error by a small amount, say f_e Hz, the frequency error for the Jth partial will be Jf_e Hz. Thus the best-case situation of known frequencies may not always be attained in practice.

3.4 Evaluation of Voice Separation with TWM Frequency Tracking

The complete automatic duet separation procedure was applied to many example duet recordings. The results of the artificial duet examples were found to be very similar to the results with a priori frequency knowledge, indicating that the TWM tracker performed well using the simple artificial inputs.

The real duet test examples obtained from analog record albums contain appropriate levels of reverberation. This reverberation is a source of trouble because the duet separation procedure assumes that no more than two sets of harmonic partials are present in a single analysis frame. At the transition from one note to the next, the frequency tracker may 1) begin to track the new note, 2) continue to follow the reverberation "tail"

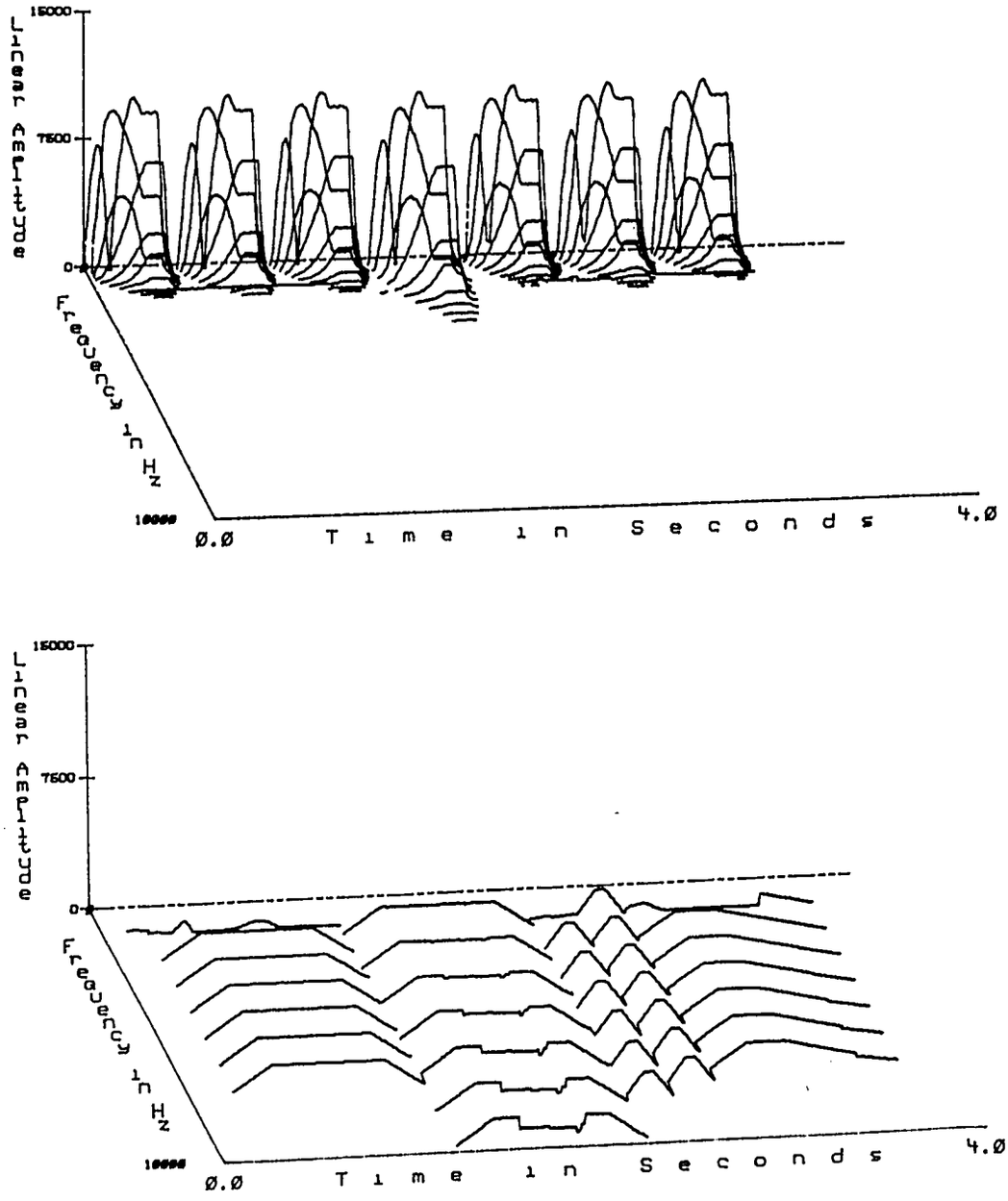


Fig. 16. Separation results for artificial duet example 2 using a priori frequency information.

of the previous note, or 3) hop back and forth between the two choices in some unpredictable way. In the first case the reverberation tails of the released note may collide with the partials of new (or sustained) notes, causing interference that is not included in the two-voice separation process and, therefore, not identified and corrected. The second case misses the attack portion of the new note, particularly if the new note starts at a lower amplitude level than the reverberated note. When the third case occurs, the separation results are generally poor.

The separation results for the first acoustic duet example (clarinet and bassoon, Fig. 11) are shown in Fig. 17. The surface noise from the analog record album source did not impair the performances of the TWM and separation procedures. However, the sound quality is degraded by the presence of audible reverberation tails of each note trailing over into the next note due to partial collisions. The reverberation tails may often be a desirable part of the separation output if they blend with the new notes in a natural manner. A less benign effect is the presence of one or more reverberated partials from one voice in the "separated" output of the other voice.

The separation results for the second acoustic duet (tuba and trumpet, Fig. 12) are depicted in Fig. 18. The sharp attacks on each note of the trumpet voice enabled the frequency tracker to follow the musical line with high accuracy and crisp transitions. The tuba line was more difficult to track. This was primarily due to corruption of the closely spaced low-amplitude tuba partials by stronger colliding partials of the trumpet. The separation results could easily by improved by manual editing of the TWM output, but that would

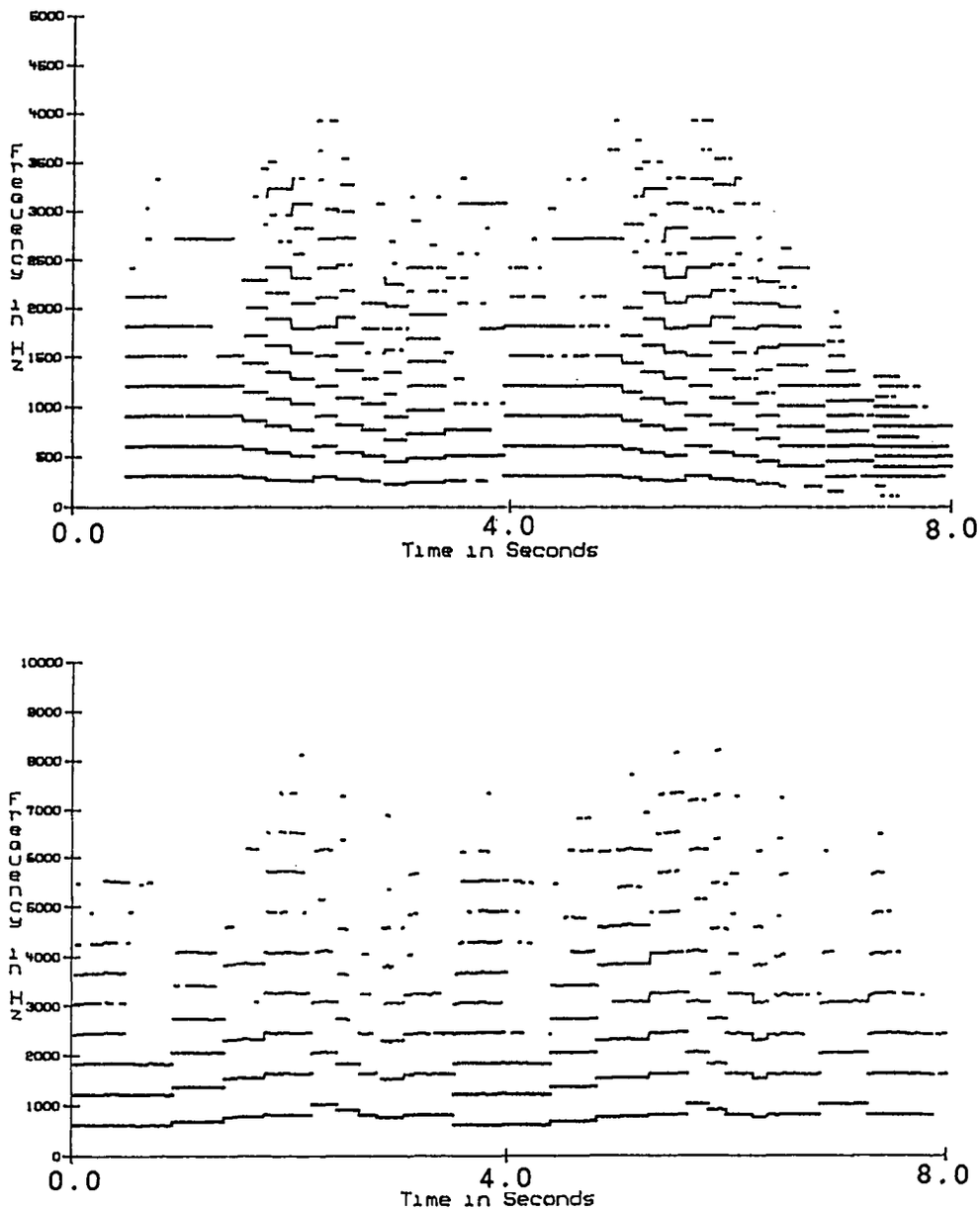


Fig. 17. Separation results for real duet example 1 using TWM frequency tracking and full separation procedure.

violate the definition of an "automatic" separation system. The separated signal quality was reasonably good for the trumpet voice, but the separated tuba voice contained occasional whistles and bleeps as low partials from the trumpet collided with the upper partials of the tuba and were misinterpreted as tuba components.

4 CONCLUSION

This research project developed novel approaches to two basic problems in duet signal separation: estimation of the two fundamental frequencies of a duet from the composite monaural signal, and separation of two voices given the pair of fundamental frequencies.

4.1 Summary of Findings

The separation procedure may be described as follows.

1) The spectral partials of one voice in a typical duet collide with the partials of the other voice. The separation procedure determines the contribution of each colliding partial to the resulting amplitude and frequency interaction observed in a short-time spectral analysis.

2) Level imbalances between the two voices of a duet may make the parameters of the weaker voice difficult to ascertain. The frequency tracking and voice separation procedures supply estimates of any missing information using knowledge of previous and subsequent analysis frames.

3) The voices of a duet may occur simultaneously, one at a time during solo passages, or not at all during shared rests. The voice separation process determines the current voicing paradigm and applies the appropriate separation method.

4) For recordings containing reverberation or other characteristics not strictly within the guidelines set forth

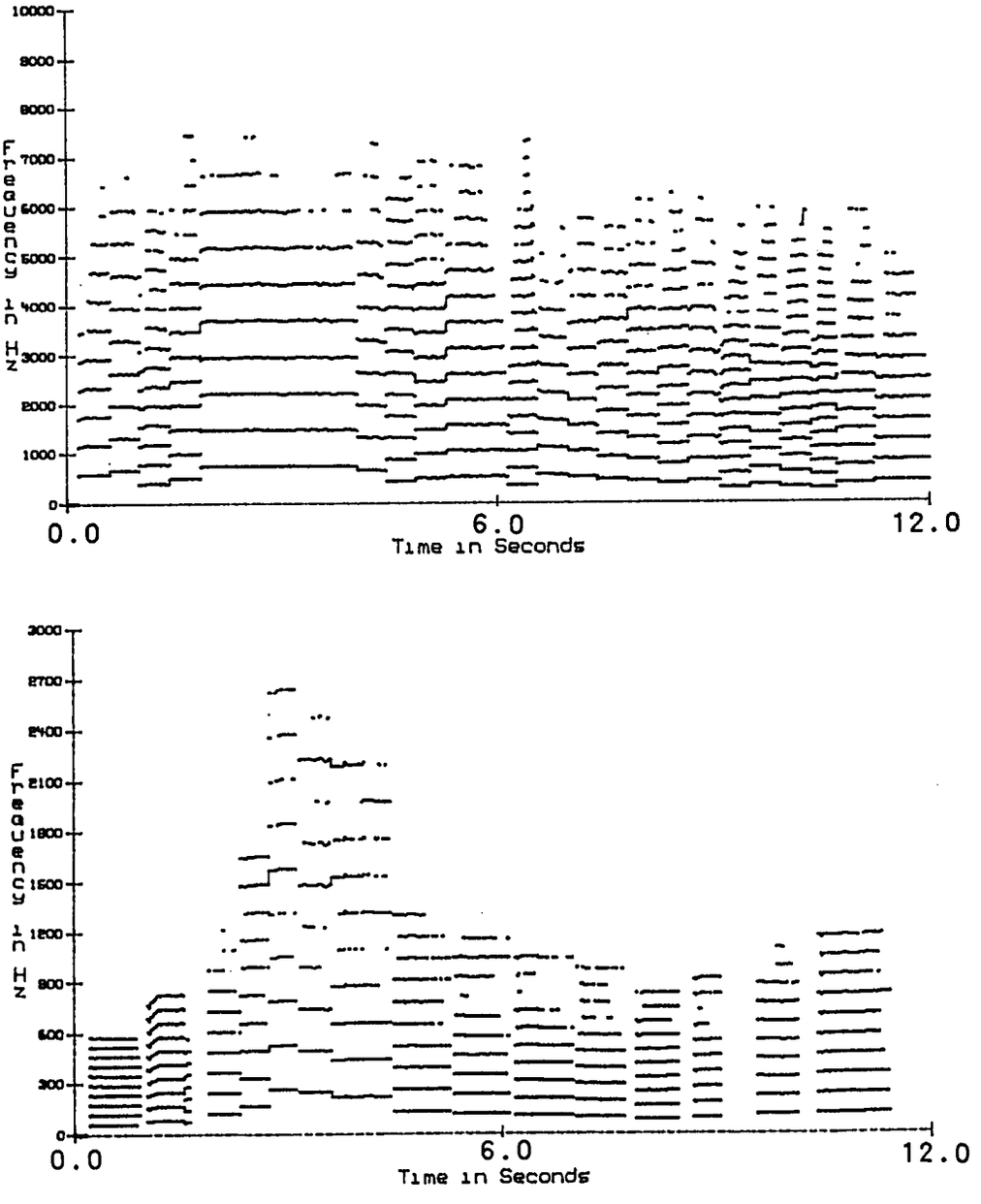


Fig. 18. Separation results for real duet example 2 using TWM frequency tracking and full separation procedure.

in the Introduction, the performance of the frequency tracking and voice separation procedures is degraded. The acceptability of the degraded separation depends on the particular combination of voices and the intended application for the results.

The results were excellent for combinations of voices and frequencies in which the number of collisions between partials of the two voices was small. The performance of the system was less satisfactory for frequency combinations in which one voice had most of its partials coincident with partials of the other voice, such as for combinations of fundamental frequencies in which the upper voice fundamental was an integer multiple of the lower voice. For typical duets the frequency relationship will change from note to note, causing the quality of the separation to vary from note to note as well.

4.2 Future Directions

For a truly practical musical signal separation system, the output signal should never be perceptually "worse" than the input signal. The system must also be robust, with reasonable behavior for a wide range of input signals and minimal operator intervention. These goals imply the need for many levels of knowledge—from short-time spectra and pitch tracking to note segmentation and even analysis of musical form. Moreover, such a system should be capable of adaptive behavior in response to the changing characteristics of its input signal. The system described in this paper is another step in this direction. Finally, extension of this separation technique to ensemble recordings with more than two voices (or to polyphonic instruments) would be necessary for a truly useful system.

The problems associated with acoustic signals need further study. In particular, the detrimental effect of reverberation encountered in this investigation needs to be resolved, since most music is recorded with natural or artificial reverberation. Further understanding and incorporation of human perception and perceptual strategies might provide the necessary breakthroughs in this area.

A different approach to the separation problem could be developed in which the time-variant analysis would be used for *note segmentation* only, with signals themselves generated according to a prespecified artificial synthesis method. This approach would be an extension of the modeling concepts considered in Sec. 2.2.3.

5 ACKNOWLEDGMENT

This work was supported, in part, by an Audio Engineering Society educational grant, a National Science Foundation graduate fellowship, and a grant from the University of Illinois Research Board. The learned and lucid comments of two reviewers are truly appreciated. The author is also pleased to acknowledge the guidance and assistance of Dr. James Beauchamp, director of the University of Illinois Computer Music Project, and the technical assistance of George Gaspar and Dr. Kurt Hebel.

6 REFERENCES

[1] J. Everton, "The Separation of the Voice Signals of Simultaneous Speakers," Ph.D. dissertation, University of Utah, Salt Lake City (1975).

[2] T. W. Parsons, "Separation of Speech from Interfering Speech by Means of Harmonic Selection," *J. Acoust. Soc. Am.*, vol. 60, pp. 911—918 (1976).

[3] B. A. Hanson and D. Y. Wong, "The Harmonic Magnitude Suppression (HMS) Technique for Intelligibility Enhancement in the Presence of Interfering Speech," *Proc. IEEE ICASSP*, vol. 2, pp. 18A.5.1–4 (1984).

[4] J. A. Naylor and S. F. Boll, "Techniques for Suppression of an Interfering Talker in Co-channel Speech," *Proc. IEEE ICASSP*, vol. 1, pp. 205–208 (1987).

[5] C. K. Lee and D. G. Childers, "Cochannel Speech Separation," *J. Acoust. Soc. Am.*, vol. 83, pp. 274–280 (1988).

[6] 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 Processing*, vol. ASSP-38, pp. 56–69 (1990).

[7] J. A. Moorer, "On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer," Ph.D. dissertation, Rep. STAN-M-3, Dept. of Music, Stanford University, Stanford, CA (1975).

[8] M. Piszczalski and B. A. Galler, "Automatic Music Transcription," *Comput. Music J.*, vol. 1, no. 4, pp. 24–31 (1977).

[9] W. A. Schloss, "On the Automatic Transcription of Percussive Music—From Acoustic Signal to High-Level Analysis," Ph.D. dissertation, Rep. STAN-M-27, Dept. of Music, Stanford University, Stanford, CA (1985).

[10] J. Chowning and B. Mont-Reynaud, "Intelligent Analysis of Composite Acoustic Signals," Rep. STAN-M-36, Dept of Music, Stanford University, Stanford, CA (1986).

[11] A. M. Noll, "Cepstrum Pitch Determination," *J. Acoust. Soc. Am.*, vol. 41, pp. 293–309 (1966).

[12] M. M. Sondhi, "New Methods of Pitch Extraction," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 262–266 (1968).

[13] M. R. Schroeder, "Period Histogram and Product Spectrum: New Methods for Fundamental Frequency Measurement," *J. Acoust. Soc. Am.*, vol. 43, pp. 829–834 (1968).

[14] M. Piszczalski and B. A. Galler, "Predicting Musical Pitch from Component Frequency Ratios," *J. Acoust. Soc. Am.*, vol. 66, pp. 710–720 (1979).

[15] J. A. Moorer, "The Optimum Comb Method of Pitch Period Analysis of Continuous Digitized Speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 330–338 (1974).

[16] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average Magnitude Difference Function Pitch Extractor," *IEEE Trans.*

Acoust., Speech, Signal Processing, vol. ASSP-22, pp. 353–362 (1974).

[17] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals* (Prentice-Hall, Englewood Cliffs, NJ, 1978).

[18] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. (Springer, New York, 1976).

[19] J. W. Beauchamp, "Transient Analysis of Harmonic Muscial Tones by Digital Computer," presented at the 33rd Convention of the Audio Engineering Society, New York, 1966 October 10–14, preprint 479.

[20] J. A. Moorer, "The Use of the Phase Vocoder in Computer Music Applications," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 26, pp. 42–45 (1978 Jan./Feb.).

[21] M. R. Portnoff, "Time–Frequency Representation of Signals and Systems Based on Short-Time Fourier Analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 55–69 (1980).

[22] M. B. Dolson, "A Tracking Phase Vocoder and Its Use in the Analysis of Ensemble Sounds," Ph.D. dissertation, California Institute of Technology, Pasadena (1983).

[23] J. Strawn, "Analysis and Synthesis of Musical Transitions Using the Discrete Short-Time Fourier Transform," *J. Audio Eng. Soc.*, vol. 35, pp. 3–14 (1987 Jan./Feb.).

[24] R. C. Maher, "An Approach for the Separation of Voices in Composite Musical Signals," Ph.D. dissertation, University of Illinois, Urbana (1989).

[25] J. B. Allen and L. R. Rabiner, "A Unified Approach to Short-Time Fourier Analysis and Synthesis," *Proc. IEEE*, vol. 65, pp. 1558–1564 (1977).

[26] F. Harris, "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform," *Proc. IEEE*, vol. 66, pp. 51–83 (1978).

[27] J. B. Allen, "Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 235–238 (1972).

[28] R. J. McAulay and T. F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 744–754 (1986).

[29] J. O. Smith and X. Serra, "PARSHL: An Analysis/Synthesis Program for Non-harmonic Sounds Based on a Sinusoidal Representation," *Proc. Int. Computer Music Conf* (Computer Music Assn., San Francisco, CA, 1987), pp. 290–297.

[30] R. C. Maher and J. W. Beauchamp, "An Investigation of the Vocal Vibrato for Synthesis," *Appl. Acoust.*, vol. 30, no. 2–3, pp. 219–245 (1990).

[31] X. Serra, "A Computer Model for Bar Percussion Instruments," *Proc. Int. Computer Music Conf.* (Computer Music Assn., San Francisco, CA, 1986), pp. 257–262.

[32] R. C. Maher, "A Meditation on a Theme by Dr. Martin Luther King, Jr." (LP recording), part of *New Electronic Works from the University of Illinois* (University of Illinois, Urbana, 1989).

[33] X. Serra, "A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic Plus Stochastic Decomposition," Ph.D. dissertation, Rep. STAN-M-58, Dept. of Music, Stanford University, Stanford, CA (1989).

[34] R. G. Danisewicz and T. F. Quatieri, "An Approach to Co-channel Talker Interference Suppression Using a Sinusoidal Model for Speech," Techn. Rep. 794, MIT Lincoln Laboratory, Lexington, MA (1988).

[35] E. H. Wold and A. M. Despain, "Parameter Estimation of Acoustic Models: Audio Signal Separation," *Proc. 1986 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics* (IEEE, New York, 1986).

[36] K. Hebel, "A Machine-Independent Sound Conversion/Sound Storage System," *Proc. Int. Computer Music Conf.* (Computer Music Assn., San Francisco, CA, 1985).

APPENDIX
IMPLEMENTATION NOTES

This research was implemented using the facilities of the University of Illinois Computer Music Project. The examples used in this investigation were obtained using an analog-to-digital converter running at a 20-kHz sample rate (monaural) and 16-bit linear quantization [36]. All calculations during processing were performed out of real time using 32-bit floating-point arithmetic on an IBM RT-PC model 125 workstation running the AIX operating system. Typical processing-to-real time ratios exceeded 200.

The digitized input signals were converted to floating point and preemphasized using a simple first-order fixed filter of the form

$$H(z) = 1 - Ez^{-1} \tag{17}$$

with $E = 0.95$. This high-pass preemphasis was included to help counteract the typical spectral rolloff of musical sounds with increasing frequency. Without preemphasis the low-amplitude high-frequency partials can be obscured by the analysis sidebands of stronger partials. The preemphasis also helps to equalize the partial amplitudes, compressing the internal dynamic range and noise requirements of the analyzer.

The short-time Fourier transform (STFT) was implemented using a fixed-length Kaiser window either 511 or 1023 points in duration, corresponding to 25.55 or 51.15 ms, respectively. The choice of window length was made according to the lowest note expected in the signals to be analyzed. The longer window was used to resolve fundamental frequencies below approximately 100 Hz, with some corresponding loss of time resolution.

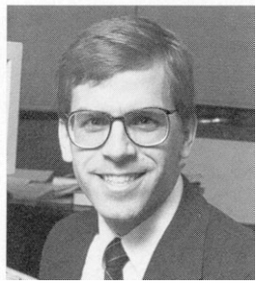
The preemphasized windowed input data were then zero padded by a factor of 2 (to length 1024 or 2048), and a standard fast Fourier transform (FFT) algorithm was used to obtain the discrete Fourier transform (DFT)

for each data frame. The frame hop was set to a fixed increment of 128 samples (6.4 ms), corresponding to one-fourth or one-eighth of the window length. Note that the preemphasis of the input signal could be removed at this point by frequency-domain deemphasis, if necessary.

The MQ analysis procedure (see Sec. 1.5.3) was applied to every frame of the STFT. The spectral peak selection process was limited in two ways: 1) a user-specified minimum peak amplitude was used as a global noise floor, and 2) a floating threshold level 50 dB below the maximum spectral peak in a given frame was used to prevent misinterpretation of window sidebands as signal components. Each peak value from the short-time spectrum was identified and stored for use in the frequency tracking and voice separation procedures.

After separation, a time-domain signal was synthesized directly from the linked-list data structure using an additive procedure. Finally, the synthesized signal was deemphasized using the inverse of the preemphasis filter, converted from 32-bit floating point to a 16-bit integer, and passed through a digital-to-analog converter for evaluation.

THE AUTHOR



Robert C. (Rob) Maher was born in 1962 in Cambridge, England, of American parents. He holds a B.S. degree from Washington University in St. Louis (1984), an M.S. degree from the University of Wisconsin-Madison (1985), and a Ph.D. from the University of Illinois-Urbana (1989), all in electrical engineering. He has received several prestigious academic awards, including a four-year, full-tuition Langsdorf fellowship, a National Science Foundation Graduate Fellowship, a University of Illinois Graduate Fellowship, and an Audio Engineering Society Educational Grant. While a student, he also worked as a graduate research assistant for the University of Illinois Computer Music Project.

Dr. Maher is a member of the Tau Beta Pi, Eta Kappa Nu, and Phi Kappa Phi honor societies, and several profesional organizations including the Audio Engineering Society, IEEE, ASA, ASEE, CMA, and IMA. He is currently an assistant professor of electrical engineering at the University of Nebraska-Lincoln, with teaching and research interests in the application of advanced digital signal processing methods in audio engineering, electroacoustics, and computer music.