Media Lab Perceptual Computing Technical Report #252
presented at the 126th meeting of the Acoustical Society of America, Denver, October 1993

A simulation of vowel segregation based on across-channel glottal-pulse synchrony

Daniel PW Ellis
MIT Media Lab Perceptual Computing, Cambridge, MA 02139
dpwe@media.mit.edu

As part of the broader question of how it is that human listeners can be so successful at extracting a single voice of interest in the most adverse noise conditions (the 'cocktail-party effect'), a great deal of attention has been focused on the problem of separating simultaneously presented vowels, primarily by exploiting assumed differences in fundamental frequency (f_0) (see (de Cheveigné, 1993) for a review).

While acknowledging the very good agreement with experimental data achieved by some of this models (e.g. Meddis & Hewitt, 1992), we propose a different mechanism that does not rely on the different *period* of the two voices, but rather on the assumption that, in the majority of cases, their glottal pitch pulses will occur at *distinct instants*.

A modification of the Meddis & Hewitt model is proposed that segregates the regions of spectral dominance of the different vowels by detecting their synchronization to a common underlying glottal pulse train, as will be the case for each distinct human voice. Although phase dispersion from numerous sources complicates this approach, our results show that with suitable integration across time, it is possible to separate vowels on this basis alone.

The possible advantages of such a mechanism include its ability to exploit the period fluctuations due to frequency modulation and jitter in order to separate voices whose f_0s may otherwise be close and difficult to distinguish. Since small amounts of modulation do indeed improve the prominence of voices (McAdams, 1989), we suggest that human listeners may be employing something akin to this strategy when pitch-based cues are absent or ambiguous.

1. INTRODUCTION

Although we are exposed to many sounds in everyday life, arguably the most important input to our sense of hearing is the speech of others. As with all hearing tasks, a major obstacle to the recognition and interpretation of such sounds is the isolation of the signal of interest (e.g. a particular spoken phrase) from any simultaneous interfering sound. These may be extraneous noise such as traffic or ringing telephones, or they may be other voices which are of less interest than the particular object of our attention. This latter case is particularly common, and in signal detection terms it is the most difficult : isolating a target from noise with the same average characteristics. This makes it difficult to design a mechanism to remove the interference from the target.

Yet this is something we do with an almost incredible effectiveness. The loosely-defined set of processes referred to as the 'Cocktail-Party Effect' enable us to engage in spoken communication in the most adverse circumstances. Although this process has many parts, none of them perfectly understood, we can list a few of the major techniques presumably employed in such circumstances:

(a) Visual verbal cues such as lip-reading;
(b) Other visual cues such as facial expression;
(c) Context, limiting the range of possible interpretations that can possibly be placed upon other evidence;
(d) Acoustic information isolated by spatial cues (interaural time differences and interaural spectral intensity differences). This is most often what is suggested by the 'Cocktail Party Effect';
(e) Acoustic information isolated from the total sound mixture on the basis of cues other than those related to spatial location.

This last category, though defined in a rather indirect manner, is of considerable interest since it is the only method available (apart from semantic context) when visual and spatial information is removed, for instance

over a telephone link or listening to a radio broadcast. Experience in such situations reveals that we are able to do a reasonable job of 'hearing out' individual voices without binaural or other cues.

There has been considerable interest in and research into this problem in the past few decades, motivated by the potential usefulness of devices that could enhance 'desired' but corrupted speech, as well as perhaps by curiosity about the strategy employed by our own perceptual mechanisms. One popular line of inquiry starts with the observation that a major attribute of much speech is its pseudoperiodic nature : 'voiced' speech is perceived as having a pitch, which corresponds to the approximate periodicity of the sound pressure waveform. When we are presented with a mixture of voiced speech from two or more independent sources, we can perhaps exploit the fact that the different voices are likely to have distinct pitches as a basis for attending to one and discarding the rest. Informal introspection reinforces this idea of pitch as a primary organizing attribute in speech.

While a number of different approaches have been taken to the problem of automatically separating voices based on the differences in their periodicity, this paper describes what we believe to be a new, additional method to help in separation. We have been particularly struck by the enhanced prominence of sounds with slight frequency modulation compared to completely unmodulated tones, as very plainly demonstrated by McAdams (1984, 1989). Considering the frequency modulation characteristics of natural speech sounds, we speculated that there may be mechanisms in the auditory system that are able to detect the short-term cycle-to-cycle fluctuations in the fundamental period of real speech, and use these to help separate distinct, simultaneous voices. We will describe the algorithm that we have developed to exhibit these qualities, and its effectiveness at this task, which exceeded our preliminary expectations. Due to its dependence on variability over a very short time scale, this approach may be considered a time-domain algorithm in contrast to the harmonic-tracking frequency-domain approaches which have been popular.

In section 2, we discuss the nature of pitch-period variation in real speech as motivation for the new approach. In section 3, we briefly review some previous work in vowel separation, then explain the basis of our new technique. Section 4 gives some examples of the preliminary results we have obtained. The issues raised by the model are discussed in section 5. We conclude in section 6 with suggestions concerning how this work might be further validated and developed.

2. A MOTIVATION — PITCH-PULSE VARIATION IN REAL SPEECH

McAdams made a very compelling demonstration of the capacity of the auditory system to segregate vowels based on differences in fundamental frequency (McAdams 1984, 1989). Three different synthetic vowels with different fundamental frequencies are mixed together. The resulting sound is a dense chord with no obvious vowel identity. However, by adding frequency modulation at a rate of a few Hertz and a depth of a few percent to any one of the vowels, it can be made to 'jump out' of the mixture, gaining a very clearly discernible pitch and phonemic identity. This demonstrates that the auditory system is able to hear out different vowels without any visual, spatial or contextual cues, but that some other cue, such as frequency modulation, is needed to identify the spectral energy belonging to a particular voice.

This result has generated a great deal of interest. In the context of 'Auditory Scene Analysis', it is interpreted as the common frequency modulation attribute of the distinct spectral regions causing the vowel to be grouped together. This suggests a mechanism where modulation rate is calculated for each frequency channel in the auditory system and compared across channels for matches. However, experiments by Carlyon (1991) showed the situation to be more complex, since prominence due to modulation can only be obtained for harmonically-related partials, not for inharmonic complexes. Still other results by Summerfield & Culling (1992) found that, while modulation of one vowel against a static background aided identification, if both target and background are modulated, even at different rates, the segregation falls back to that achievable with static target and masker. Thus the rules governing the benefit of frequency modulation must be quite complex to account for these results.

Nevertheless, the original McAdams demonstration remains a fascinating piece of evidence. While there are doubtless several different reasons why a modulated tone might be more 'prominent' than a static one, we were interested in the possibility that there might be grouping or fusion mechanisms in the auditory system that actually worked better in the presence of modulation than on purely static tones. This speculation is based in part on the informal observation that even for nominally unmodulated tones (such as a slowly-pronounced word, or a sung musical note), there is a significant qualitative difference between the sound produced by a human speaker, and the sound of an absolutely static machine-generated tone although their spectral envelopes may be closely matched. The major factor in this distinction is that even when producing a steady pitch, there is a certain amount of cycle-to-cycle variation in human vocalization which is absent in the machine-generated version. Thus we were curious to investigate the possibility of an auditory grouping scheme that might be particularly well suited to handling this kind of natural sound, where every cycle has some variation. We sought to develop a model of a mechanism of this kind.

If we are interested in a process that is sensitive to the short-term fluctuations found in natural speech, it is useful to have some quantitative results describing the nature of

this variation, which McAdams called 'jitter'. We recorded some examples of human-produced steady vowels (where the speaker was attempting to produce an unmodulated sound). We then extracted the cycle lengths for these sounds using a matched filter based on an 'average' pitch cycle waveform to locate a fixed reference point in each cycle. Since we used the vowel /ah/ with a relatively flat spectrum, it was possible to extract cycle lengths with good accuracy.

The result was a series of cycle times with peak variation of about 2% of the average length. Treating these as samples of a continuous modulation function (a slight approximation since the sampling was not exactly uniform), we were able to calculate the spectrum of the modulation. A typical modulation contour and its spectrum are shown in figure 1.

We see that the modulation is broad-band. While there is a strong low-frequency component to the pitch variation peaking at around 3 Hz (which we might call vibrato or pitch modulation, since the tendencies extend across enough cycles to be perceived as a change in pitch), there is almost as much variation on a cycle-to-cycle basis, with a largely white (flat) spectrum. This is the particular component to which we were referring by the term 'jitter' above.

Smoothing (i.e. low-pass filtering) the modulation contour produces a pitch modulation curve, and the difference between this and the actual modulation contour can be thought of as the high-frequency (6-30 Hz) jitter residual. We experimented with constructing new vowels based on combinations of different amounts of both these components, resynthesizing vowels that varied from having no modulation at all, to vowels where either or both of the modulation components were exaggerated by a factor of up to eight. Informal listening suggested that while either type of modulation increased the naturalness of the synthetic vowels, it was only the combination of them both — i.e. a broadband cycle length variation — that sounded truly natural and non-mechanical. (Other research describing the nature of modulation in natural voicings includes the work by Cook (1991) on singing voices.)

This result is significant because the experiments of Carlyon and Summerfield that measured the impact of 'frequency modulation' on various perceptual abilities used simple sinusoidal modulation, in order to have stimuli that were easy to describe, generate and control. However, it might be that the auditory system has evolved to make use of some of the properties of natural sounds that were not well duplicated by such narrowband modulation, such as the short term variation we have termed jitter. If this is the case (and we will argue that the main model of this paper has this property) then it may be that these experiments have involved stimuli that prevented the full segregation benefits of modulation from being obtained.

A possible distinction between slow frequency modulation and jitter stimuli

By way of supporting our suggestion that complex tones with low frequency (≤ 8 Hz) narrowband frequency modulation might not allow the optimum performance of modulation-detecting mechanisms in the auditory system, we will sketch an idea of the kind of processing that might find broadband jitter-like modulation easier to segregate than smooth fm. Consider a system that can detect period changes between adjacent cycles that exceed some fixed percentage. With sinusoidal modulation, the peak adjacent cycle-length variation occurs at the maximum slope i.e. at the zero crossing of the modulation function. Broadband 'white' modulation - for instance, samples from a Gaussian distribution - will achieve the same level of cycle-to-cycle modulation with a far smaller 'mean' variation; however, if the threshold for detecting adjacent-cycle variation is bigger than the mean variation, steps of this size will occur relatively infrequently, with the expected time between such extreme modulation events depending on how far up the 'tail' of the distribution the threshold lies.

Now consider an auditory grouping unit that is observing various partials and trying to organize them into separate sounds. The aspect of this unit crucial to the current argument is that it accumulates evidence across time. In particular, if it has evidence that certain partials should be grouped together in time slice tn, it can remember this evidence and continue to group the partials involved for

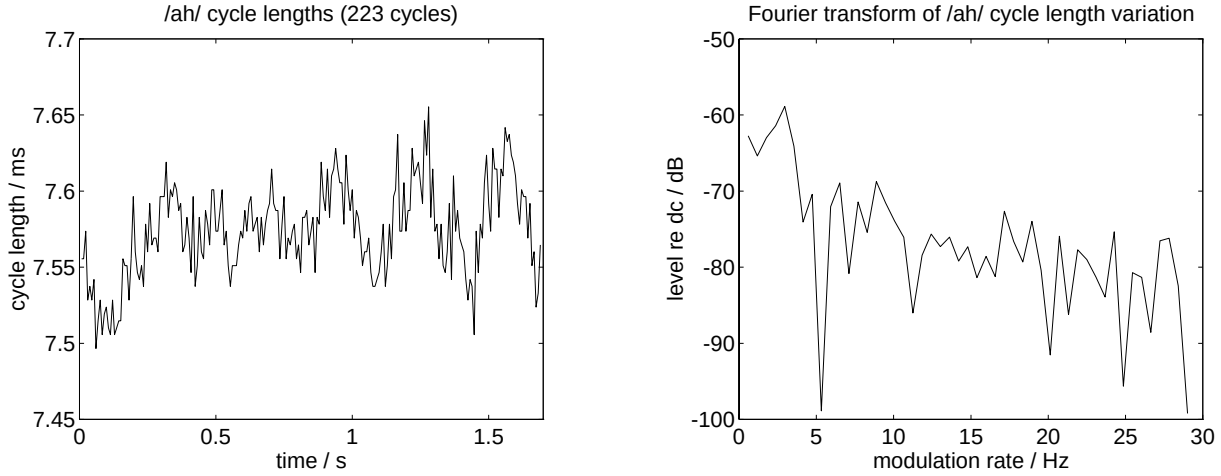

figure 1: *Cycle length modulation function extracted from real speech (a sustained /ah/), and its Fourier transform.*

many subsequent time steps, even though no further evidence may be observed. (In the interests of robustness, this association should 'wear off' eventually, in some unspecified fashion).

We present this unit with a combination of two modulated vowels. If the vowels are sinusoidally modulated to a sufficient depth that the adjacent-cycle-variation detector is triggered, or if the sensitivity of this detector is adjusted to trigger at some minimum rate, this triggering will occur for a significant chunk of every half cycle (since the second derivative of this modulation, the rate at which the cycle-to-cycle variation changes, is smallest when the variation itself is at a maximum). We might imagine that the modulation detector is overloaded, and since both vowels are highly modulated for quite a large proportion of each cycle, it may be harder to find disambiguating time-slices where the harmonics of only one of the vowels are highly modulated thereby establishing the independent identity of that vowel. The size of the hypothetical time slice used for evidence collection relative to the modulation period will be a factor in this situation.

By contrast, if the modulation of the two vowels is according to a white Gaussian jitter function, then the threshold of the adjacent-cycle-variation detector can be adjusted to provide events that are sufficiently infrequent to allow the evidence observer a clear look at each event, yet sufficiently regular that enough evidence is collected to maintain a confident arrangement of the stimuli across time. This arises because of the different probability distribution of instantaneous cycle-to-cycle variation under the two schemes, as illustrated in figure 2 below: The sinusoidal modulation is sharply bimodal, making it very difficult to adjust a threshold to achieve a particular density of cycle-variation events. By contrast, the Gaussian modulation has long, shallow tails making such a task far easier.

This example is intended merely to illustrate a simple scenario that highlights an important difference between narrowband and broadband cycle length modulation. This example is not a particularly good description of the model that is the main subject of this paper, since the tracking of individual components is a very inappropriate way to consider the perception of formant-structured sounds like vowels. However, the general idea of irregularly spaced disambiguation events is central to our model, and will be re-introduced later.

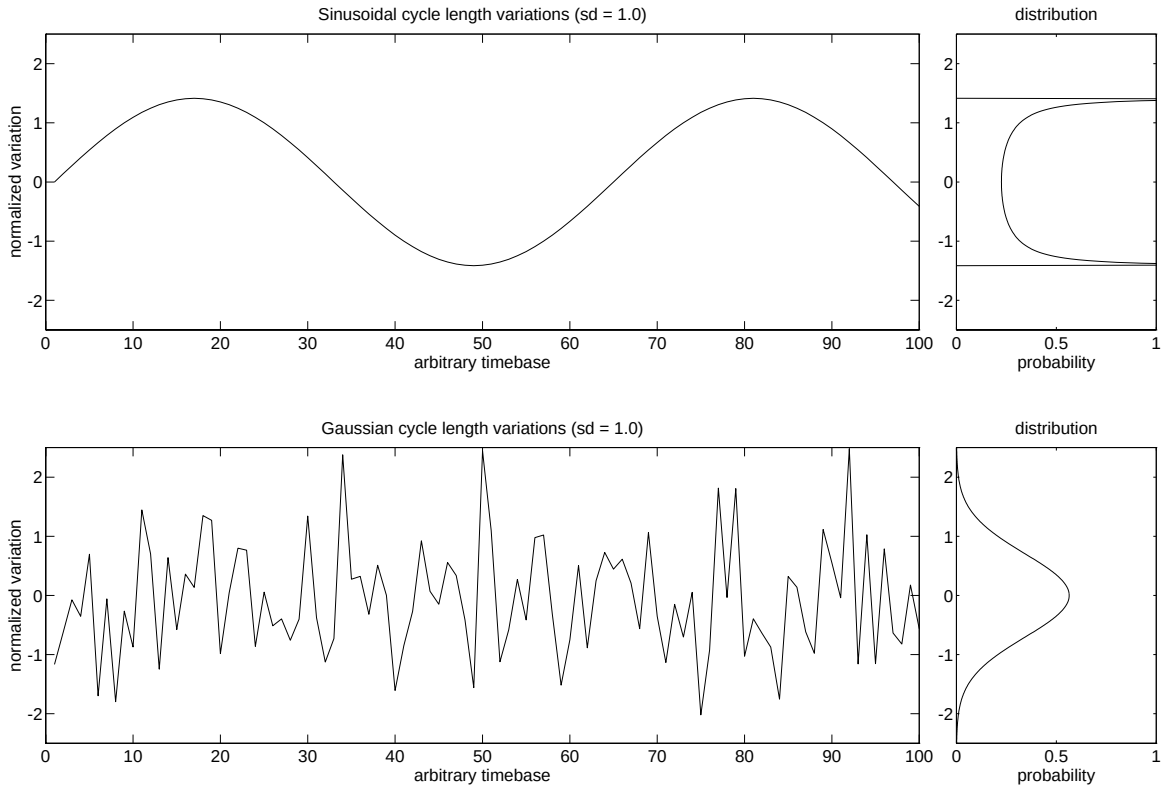

figure 2: *Examples of modulation patterns according to sinusoidal and Gaussian functions, along with their theoretical distribution functions.*

3. THE GLOTTAL-PULSE SYNCHRONY MODEL AS AN ALTERNATIVE TO PERIODICITY METHODS

A brief review of previous work in double vowel separation

Vowel separation seems a tantalizingly tractable problem, since the presumed different fundamental frequencies should form a powerful basis for segregation. Such a technique was described by Parsons (1976), which sought to separate two voices by identifying both pitches and using this to segregate the harmonics belonging to each talker in a narrow-band short-time Fourier transform.

de Cheveigné (1993) describes this and many subsequent models based both on that kind of frequency domain representation, and on equivalent time-domain processing such as comb filters, which he prefers largely for reasons of physiologic plausibility. He classifies the processing performed by these methods as either *enhancement* (where the target is identified and boosted relative to the interference) or *cancellation* (where the interfering voice is identified and subtracted). He proposes a model of the latter, including hypothetical neural circuitry, on the basis that signal separation is most useful when the interference, but not the target, is readily identified.

Some recent research has focused on modeling the separation strategies employed by the auditory system, which is also our objective in the current paper. Some approaches seek to duplicate the behavior of human subjects in given psychoacoustic tests. The models of Assmann and Summerfield (1989, 1990) match human identification rates for static vowel pairs by finding vowel pitches and extracting the harmonic spectra, but work within the constraints of an 'auditory filterbank' (Moore & Glasberg, 1983). Since this filterbank is unable to resolve the individual high-frequency harmonics, this information is extracted from a subsequent autocorrelation stage at the lag corresponding to the appropriate pitch.

The Meddis & Hewitt (1992) model similarly consists of an auditory filterbank followed by a nonlinear 'inner hair cell' model providing rectification and dynamic range compression and generating auditory nerve 'firing probabilities' which are autocorrelated in each channel. Their vowel separation strategy sorted channels of the filterbank according to the dominant periodicity observed in the autocorrelation - i.e. the 'pitch' in that particular channel. Autocorrelation provides a particularly effective mechanism for projecting the shared fundamental periodicity in different frequency channels onto a common feature, since every formant of a given vowel will have a peak in its autocorrelation centered at the lag corresponding to the period - reflecting the fact that the vowel sound, and hence any subrange in its spectrum - is periodic in that cycle time. In the Meddis & Hewitt model, the dominant periods from each channel are pooled to provide the pitch estimates, and then individual channels are attached to one or other of the detected pitches accordingly. This model was capable of very close agreement with the performance of human listeners for identification rate as a function of fundamental frequency separation of static vowels.

Summerfield and Assmann (1991) attempted to isolate which particular consequence of pitch difference aided the separation of two voices. By careful control of each harmonic, they constructed artificial stimuli which exhibited either harmonic frequency separation or formant burst temporal separation without most other attributes of double vowels. They found that neither of these qualities was sufficient to obtain the gain in recognition accuracy afforded by conventional synthesis of vowels with 6% frequency difference. However, they did not experiment with modulation of the artificial vowel spectra, which would have been particularly relevant to this paper.

All the methods make a pseudoperiodic assumption for the input signal i.e. they treat the signal as perfectly periodic, at least over some short analysis window. (However, the time-varying 'corellograms' of Duda et al (1990) are intended to exploit the dynamic nature of the sound as a basis for separation). All the separation procedures involve the explicit extraction of the pitch of one or both vowels. These aspects are very reasonable, since aperiodic 'vowels' (a formant synthesizer excited by a free-running Poisson process perhaps) would not be perceived as voice-like. However, we were curious about the possibility of techniques for detecting and extracting sounds that did not assume perfect periodicity over some analysis window, with the attendant loss of efficiency when handling real, jittered, sounds — particularly as such real sounds are sometimes easier to segregate than their truly periodic imitations.

The new method - detecting formants by glottal-pulse synchrony (GPS)

The central idea behind the algorithm we are going to describe is that the different regions of spectral energy that originate in a particular voice may be grouped together by observing the time synchronization of different formant regions excited by the same glottal pulse. Whereas methods such as those mentioned above have exploited features in the Fourier domain or the autocorrelation arising from the periodic nature of glottal excitation, we are interested in grouping the energy associated with each individual pulse, not the combined effect of successive pulses. This follows naturally from our interest in variations between individual cycles, but it seems to preclude us from using the strong cues of periodicity normally employed. We will explain how we can manage without them below, but first we review some important aspects of voiced sounds upon which the new method relies.

The method of production of voiced sounds in people is approximately as follows: air pressure from the lungs is obstructed by the vocal cords in the glottis, which periodically open and flap shut in a form of relaxation oscillator. The resulting pulses of air flow form an excitation to the upper throat, mouth and nose cavities which act as resonators, so that the resulting pressure radiation has the same periodicity as the glottal pulses, but with a dynamically modified spectrum. By moving the tongue and other articulators we change the nature of this spectrum and thereby produce the different vowels of speech. Mathematically, we may describe this production in the well-known source-filter formulation:

$$s(t) = e(t) * v(t) \quad (1)$$

where $s(t)$ is the radiated voice signal, $e(t)$ is the glottal pulse excitation. $v(t)$ is the slowly-varying vocal tract transfer function, here approximated as a time-invariant impulse response. $e(t)$ is well approximated as an impulse train at least over a bandwidth of a few kilohertz, yielding:-

$$e(t) = \sum_i A_i \delta(t - t_i) \quad (2)$$

where $\{t_i\}$ are the near-periodic instants of glottal pulsation, each of amplitude $\{A_i\}$, and thus

$$s(t) = \sum_i A_i v(t - t_i) \quad (3)$$

i.e. the output waveform is the superposition of individual vocal tract impulse responses shifted in time by the glottal pulses. Ignoring the nasal branch, the vocal tract can be modeled as a single concatenation of tubes of varying cross-section, which results in an all-pole structure for the filter's Laplace transform,

$$V(s) = \frac{K}{\prod_j (s - s_j)} \quad (4)$$

or, equivalently, an impulse response consisting solely of damped resonances:

$$v(t) = \sum_j a_j \cdot \exp(-t/\tau_j) \cdot \cos(\omega_j t + \phi_j) \quad (5)$$

where the resonant frequencies $\{\omega_j\}$, amplitudes $\{a_j\}$ and decay times $\{\tau_j\}$ are determined by the geometry and losses associated with the resonant cavities in the vocal tract.

If we now look at the time waveform and time-frequency analysis (spectrogram) of some example speech, we can see the kinds of features we would expect according to this analysis. Figure 3 shows 100 milliseconds (about 10 cycles) of a male voice pronouncing the vowel /ah/. The top panel is the pressure waveform as a function of time, showing the large amplitude burst at the start of each cycle, as the vocal tract is initially excited by a release of pressure through the glottis. This then decays away approximately exponentially in a complex mixture of resonances until the start of the next cycle. The lower panel shows a wideband short-time Fourier analysis on the same time scale, with a vertical (linear) frequency scale, and gray density showing the energy present at each time-frequency co-ordinate. Since the analysis window was much shorter than the cycle length, we do not see any of the harmonics of the psuedoperiodic voice, but rather we see a broadband stripe of energy at each glottal pulse. This stripe varies in intensity with frequency, showing peaks at (in this case) around 1000, 4000 and 7000 Hz; these are the center frequencies of the vocal tract resonances we have described above, normally known as the formants of the voiced speech.

The decomposition of the voice into a succession of time-shifted impulse responses implied in (3) above leads us to consider this same display applied to just one glottal pulse. The signal in figure 4 was generated by exciting a 10-pole LPC model of the voice above with a single impulse. What we see is an approximation to an isolated stripe out of the pattern on a somewhat stretched time-base. It is not hard to imagine reconstructing the entire pattern by superposing time-shifted copies of this lone example, as suggested by equation 3.

In the context of this presentation of speech sound, the idea behind the glottal-pulse synchrony (GPS) technique is simple to explain: The problem of extracting a voice from interfering sound is essentially the problem of identifying the formant peaks of that voice, since the particular phoneme being conveyed is encoded in the frequencies of those peaks. If there are numerous energy peaks in the received spectrum, how can segregate target and interference? In light of the images and analysis above, perhaps we can group together all the formants that belong to a particular voice by detecting the way that they are all time-aligned in vertical stripe structures, which is to say that they are all synchronized to the same pulse of glottal energy exciting the speaker's vocal tract. If we can at least organize the composite sound into disjoint sets of formants arising from different speakers, our segregation problem is simplified into choosing which of these is the target voice, and which are interference. While this process is clearly easier to postulate than to accomplish, it is the basis of the technique we have developed which we will describe in more practical terms presently. Specifically, our algorithm detects sets of spectral regions that appear to exhibit synchronized peaks in energy, as if in response to a common excitation. Although these associations exist only over a very brief time, it is possible to use them to separate out voices across time by imposing weak continuity constraints on the formants in a voice to extend these near-instantaneous spectral associations across time into complete phonemes or words.

It is interesting to contrast this approach with autocorrelation methods, in particular the vowel segregation model of Meddis and Hewitt (1992) (MH), since their model was the inspiration for this work and shares several components (thanks to their generous policy of making source code available). Both models start with an approximation to the cochlea filterbank followed by an element that calculates the probability of firing of a given inner hair cell on the Basilar Membrane. The MH model then calculates the autocorrelation of this probability function in each channel to find the dominant periodicity within each channel. Thus each channel is first processed along the *time* dimension to measure the period between moments of high firing probability, and only then are the channels that share a given periodicity grouped across *frequency* to form the complete vowel. (In the two vowel case, this is done for the 'dominant' periodicity or autocorrelation peak of the entire population; the remaining channels are presumed to belong to the other vowel). However in the GPS method

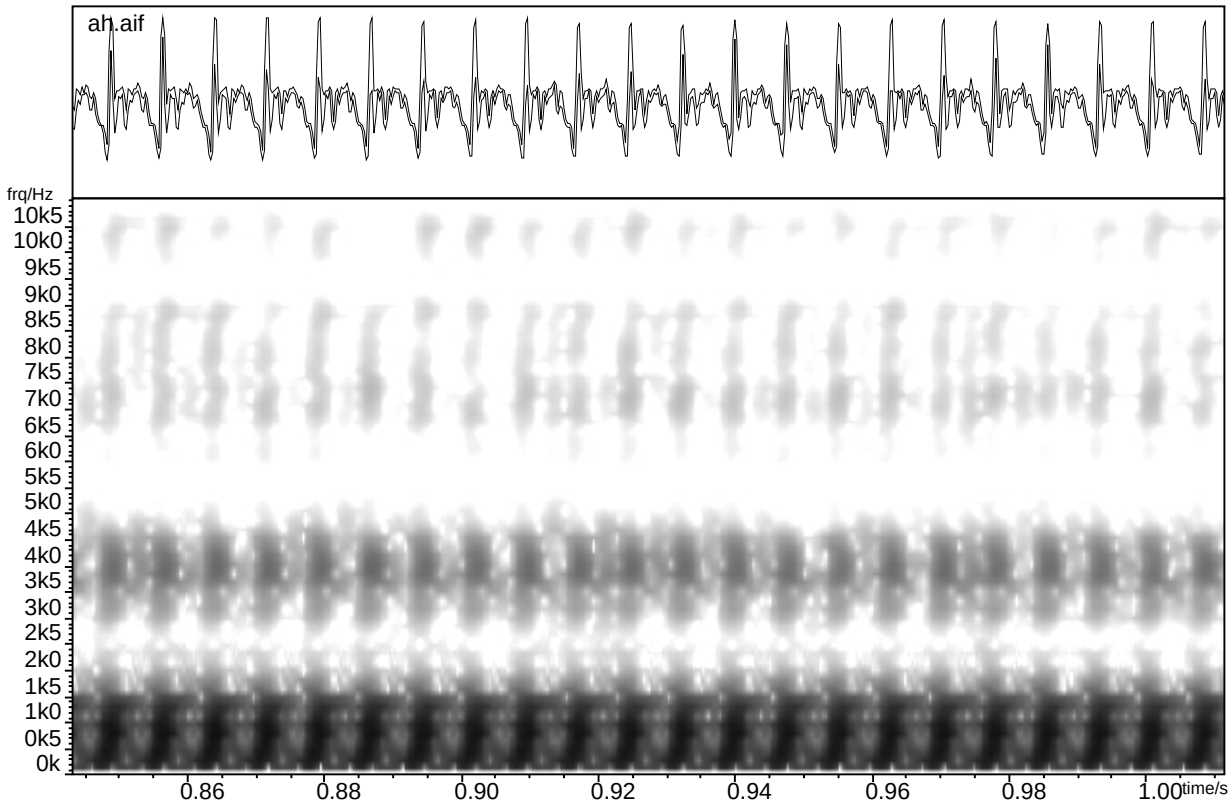

figure 3: *Wideband spectrogram of a section of the long vowel /ah/. Time goes left to right; the top panel shows the actual waveform; the bottom panel has (linear) frequency as the vertical axis, and the gray density shows the amount*

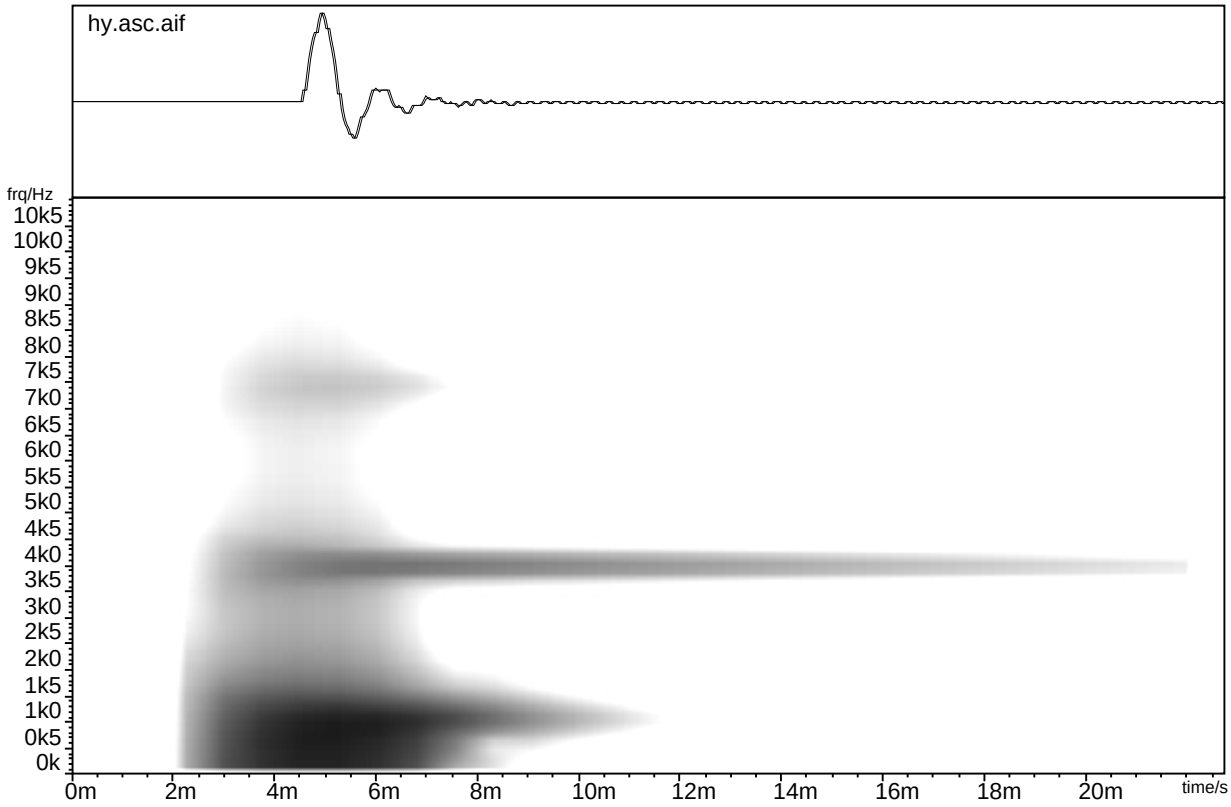

figure 4: *As figure 3, but showing the impulse response of an all-pole model based on the /ah/ vowel. Note the expanded timescale (22 milliseconds across the figure compared with 170 ms in figure 3).*

the first dimension of processing for each burst is across *frequency* to find other, co-synchronous events. Only after grouping on this basis are the spectra traced through time to form more complete sounds. Thus there is a contrast between the order of across-frequency and along-time processing between the two models.

The importance of this difference becomes apparent when one considers the processing of signals with significant cycle-to-cycle variation (such as the jitter discussed in section 2). In order to obtain a reliable and consistent estimate of the periodicity in a given channel, the MH autocorrelation function must have a time window extending over several cycles. But variation within those cycles will compromise the calculation of a single peak periodicity; since the MH method is rather sensitive to the form of the peaks in the autocorrelation, any jitter in the input signal can only serve to hinder their grouping scheme by blurring the period estimates for each channel. This appears to be at odds with the enhanced perceptual prominence and naturalness conferred by moderate jitter in signals.

On the other hand, the GPS method is quite impervious to the spacing between successive pulses, and any variation that might occur. All it relies upon is the consistent alignment of the formant bursts on the individual glottal pulses, and sufficiently frequent bursts to allow tracking through time. In this way, cycle length variation will not compromise the detection of the voice. Indeed, as we will argue below, it can be a strong cue to the segregation between two close voice sounds.

Synchrony skew

We have suggested an algorithm which distinguishes a voice burst from background noise by the consistent time alignment of each of its spectral formant peaks. Implicit in this is the assumption that 'time alignment' denotes exact synchronization, and if this was in fact the nature of the signal, the grouping of formants might be a simple matter: whenever a formant burst is detected, simply look across the spectrum for other simultaneous bursts. However, the real situation is not so simple, as formant bursts at different frequencies, despite originating in a single compact glottal event, may be quite widely dispersed in time before they reach the auditory processing centers. Some of the factors contributing to this dispersion are:-

(a) The phase spectrum of the vocal tract. Even the simple all-pole cascade model can introduce significant dispersion across frequency;

(b) The dispersive effects of the transmission path from speaker to ear, including effects of reflections and diffractions;

(c) Frequency or place-dependent delays in the function of the ear, most notably the propagation delay down the Basilar membrane.

While (c) should be constant and is doubtless subject to static compensation in the auditory system, is does strike a final blow against the superficially attractive idea that we might be able to implement the GPS algorithm without resorting to delay lines. But in any case the unpredictable and very significant effects of (a) and (b) mean that no simple solution will suffice; rather than simply causing simultaneous formant bursts in different channels, the cue of common glottal excitation will appear as a constant, often small, but initially unknown *time skew* between the channels. Thus the problem is complicated into finding both the channels that carry bursts in a synchronized pattern, and the interchannel timing relations that constitute that pattern.

Evidently, we will not be able to deduce these timings by observing a single vowel burst, even if we make assumptions about the maximum possible interchannel timing difference; we cannot distinguish between formants of the target voice and energy peaks from the interfering noise which happen to fall inside our window, suggesting a completely spurious formant. (In the two-vowel situation, such an occurrence is very likely.) The solution is to store all the timings implied by the energy bursts in the window as *potential* components of a formant spectrum, but to refrain from actually using them as the basis for a separately grouped voice until the timings have been confirmed by repeated observations.

This process may be described more formally as seeking to estimate the conditional probability of a firing in one channel at a given timing relative to a firing in another channel. If peripheral channel n exhibits firing events at times $\{t_{n,i}\}$, we can define a 'firing function' for that channel:

$$F_n(t) = \begin{cases} 1 & t_{n,i} \le t < t_{n,i} + \Delta t \quad \text{for some i} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

(where Δt accounts for the temporal resolution of the detection system). We would like to estimate the conditional probability of a firing occurring in channel m at a specific time-skew τ relative to a firing in channel n i.e.

$$PF(n,m,\tau) = \Pr\left\{ F_n(t+\tau) = 1 \,\middle|\, F_m(t) = 1 \right\} \tag{7}$$

If we have a set of formants at different frequencies being excited by the same glottal pulses, we would expect this conditional probability to be very high for the pairs of channels involved at the appropriate relative timings. Thus the approach undertaken by the GPS method is to continually estimate this conditional probability for many channel combinations and time skews to find the pairs of channels that do exhibit a high correlation by this measure.

Since the vocal tract and hence the formant spectrum of a given voice are dynamic, the function $PF(n,m,\tau)$ is time varying and can be estimated at a given instant based on nearby events. We can calculate the estimate at a particular time T as the ratio of coincidences to the total time the condition was true:

$$\hat{PF}_T(n,m,\tau) = \frac{\int e^{-|T-t|/\alpha} F_n(t+\tau) \cdot F_m(t)\, dt}{\int e^{-|T-t|/\alpha} F_m(t)\, dt} \tag{8}$$

Here each event has been weighted by $e^{-|T-t|/\alpha}$ which increases the influence of events close to the time for which we are evaluating the estimated probability. The time constant for this simple weighting function is α, which must be large enough to span several pitch cycles if we are to include a sufficient number of events in the basis for our estimate to make it reasonably stable, while still remaining short enough to allow the formant spectrum to be considered invariant for those events. (Note however that the stability of the pitch *period* over this window has practically no influence on the result, as we desire). Thus α is around 10-20 milliseconds, larger than any τ we will use, and far larger than Δt. We may simplify this calculation as the sampling of one firing function F_m at the firing times of the other channel, $\{t_{n,i}\}$, normalized to the total number of firings:

$$\hat{PF}_T(n,m,\tau) \approx \frac{\sum_i e^{-|T-t_{n,i}|/\alpha} F_m(t_{n,i})}{\sum_i e^{-|T-t_{n,i}|/\alpha}} \quad (9)$$

If we assume that the estimate at time T can only look backwards in time i.e. depend on events for which $t_{n,i} < T$, and if we only update our estimate each time we get a conditioning event i.e. at each of the $\{t_{n,i}\}$, we can recursively calculate $\hat{PF}_{t_{n,i}}(n,m,\tau)$ from its previous value, $\hat{PF}_{t_{n,i-1}}(n,m,\tau)$, as:

$$\hat{PF}_{t_{n,i}}(n,m,\tau) = F_m(t_{n,i}) + e^{-|t_{n,i}-t_{n,i-1}|/\alpha} \cdot \hat{PF}_{t_{n,i-1}}(n,m,\tau) \quad (10)$$

This is effectively the calculation we perform with the histogram 'plane update' described in the next subsection.

The complete Glottal Pulse Synchrony strategy for grouping vowels

We have now explained the principle of operation of the model, and can present the functional structure in detail. Figure 5 shows the entire model in block-diagram form, from the acoustic signal at the left to a resynthesized isolated vowel on the right. As noted above, the first two stages are taken from the MH model, but the remainder is different. We explain each block in turn.

The first stage simulates the frequency analysis of the cochlea with a linear filterbank based on the 'gammatone' approximation to auditory filter shapes (Patterson & Moore, 1986, as referenced in Meddis & Hewitt, 1991). We used a filter spacing of between two and five per ERB unit, over a range 100 Hz to 4 kHz - on the order of 60 overlapping channels. The filters are real (not complex), and the outputs are undecimated, so the data carried to the next stage is a set of bandpass signals, one per filter.

The next stage is the MH cochlea inner hair cell model, nominally translating the basilar membrane displacement (the output of the linear filter) into a probability of firing for the associated nerve fiber. This stage approximates the rectification and adaptation effects observed in the firings of actual auditory nerves in response to acoustic stimuli.

Following that is our pseudo-'firing' module. The GPS model is described in terms of detecting formant bursts for each pitch pulse at different frequencies, therefore we desire the input spectrum transformed into a representation that has discrete events for each such burst (at least within each channel). We experimented with a variety of models of hair-cell firing, ranging from a simple Poisson event generator driven by the firing probability, to models that included refractory effects such as the Ross (1982) 3rd-order Poisson model. However, the noise added by such stochastic elements was quite overwhelming, presumably because the effective number of nerve fibers was very small (one per peripheral channel). Eventually we decided to sacrifice physiological resemblance at this stage and devised a deterministic formant-burst event generator that produces one firing for each local maximum in the firing probability function over a sliding time window of typically 5 ms. This does a very satisfactory job of producing one pulse every pitch cycle in channels where given formants are dominant. It is not clear how this function could be generated in an actual physiological system; nerve fibers do of course generate firings synchronized to their inputs

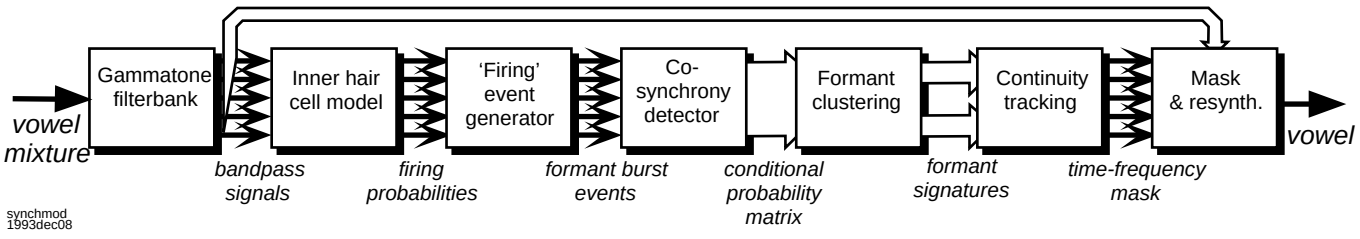

figure 5: *Block diagram of the complete Glottal Pulse Synchrony (GPS) vowel separation scheme.*

at this level of resolution, though not with such consistency.

It is thus a very reduced spectral representation, consisting of a set of frequency channels containing binary events once every few milliseconds, that is passed to the next stage. This is the co-synchrony detector that is the heart of the method. Essentially, it is a three-dimensional histogram, as illustrated in figure 6. The two major dimensions are both indexed by peripheral channel (i.e. frequency), and the third axis is time skew or lag. Thus each bin is calculating a value of the estimated conditional probability function $\hat{PF}_T(n,m,\tau)$ described above i.e. the likelihood that a firing in peripheral channel n (indexed along the vertical axis of the histogram) will follow a firing in channel m (the horizontal axis) at a relative time τ (the short axis going into the page). The actual method of updating this matrix is as follows: when an event is detected in channel n, an entire plane of the histogram is updated. Every other channel, indexed across m, corresponds to a row in this plane with a bin for each possible time skew τ. If there is an event in the other channel within the range of time skews considered (i.e. within a couple of milliseconds of the burst causing the update), the bin appropriate to that time skew is incremented up to some saturation limit. All the bins that do not correspond to events are decremented unless they are already empty. Thus every bin in the plane is changed, unless it is already at a limit of its excursion. (In practice, most bins will spend most of the time at zero). This saturating count is a crude approximation to the exponentially-weighted estimate of equation (10), with the advantage that it permits the use of small integers for each element in the array.

We note in passing the similarity between this matrix detecting co-occurrences of firings in different frequency channels to the binaural co-incidence detector of Colburn (1977), which detects consistent timing skew between channels of the same center frequency, but from opposing ears, as a factor contributing to perceived spatial origin.

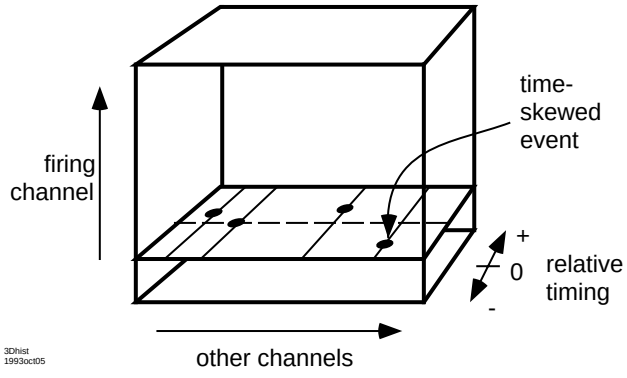

figure 6: *The three-dimensional histogram for estimating the conditional probabilities of firing in pairs of channels at particular relative timings.*

This procedure behaves as follows: Presented with a succession of pitch bursts with a consistent formant pattern, the bins corresponding to the relative timings between the formants for each pair will be incremented on every pitch cycle, and the other bins will remain at zero. Thus by looking for the maximum bin contents along the time skew axis for each channel m compared to a single channel n (i.e. one 'plane'), we will see which channels have occurred with a consistent relative timing. If channel n contains a formant of a particular voice, the ms that have large maximum bin contents (across τ) will be the other formants of the same voice, and for each channel pair the τ of the bin that contains the maximum will give the relative timing (phase difference) between those two formants (of no particular interest to the grouping mechanism, which is mainly interested in the existence of some consistent relative timing).

An example of the values in this matrix during a typical mixed voice sound is shown in figure 7. Here we see the histogram collapsed along the time skew dimension. The vertical and horizontal co-ordinates are the two frequency channels n and m respectively, and the level of gray shows the maximum histogram score across τ for that pair of channels. Because the histogram is updated by planes (horizontal rows in the figure), the image is not quite

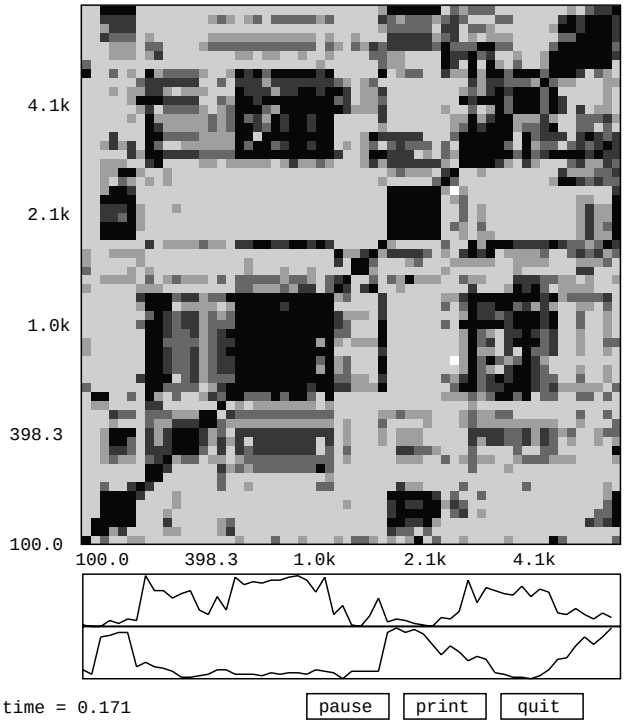

figure 7: *A 'snapshot' of the histogram collapsed along the time skew axis 0.171 seconds into a mixed vowel. Each axis is labeled by the center frequencies of the channels. Gray density shows the estimated conditional firing probability.*

symmetric about the $n=m$ axis; if there are more firings in channel A than channel B, then there will be more updates to the bins in channel A's row that correspond to channel B than there are of the bin for channel A in channel B's row - although for positive counts, we expect symmetry i.e. an event in channel A at a time t relative to an event in channel B will be recorded as counts in the two histogram bins (A, B, t) and (B, A, -t).

In figure 7, we can see the presence of two independent sets of synchronized formants, appearing as grid patterns of dark spots in the grayscale image. Thus frequencies 120 Hz, 2000 Hz and 6000 Hz (approx.) tend to occur with consistent time skew, as do 400 Hz, 1000 Hz and 4000 Hz, but not the intersection i.e. the bins indexed by 1000 Hz and 2000 Hz are pale indicating no identified regular timing skew, even though there are evidently firings in both channels, synchronized to other frequencies.

The next stage of processing is to convert this large map of conditional probability estimates into a few inferred formant spectra, accomplished by the next block from figure 5, the 'formant clustering' unit. This considers each row in the matrix from the previous stage and attempts to find a small number of formant spectra that, between them, resemble most of the rows in the matrix. The algorithm is a simplified version of k-means clustering: Starting with an empty set of output formant spectra or 'signatures', we go through each row of the matrix: If it has a peak value above a threshold (indicating that it contains some useful co-occurrence information), it is compared using a distance metric to each of the current signatures. If it is close to one, it is added to that cluster, and the signature is modified to reflect the contribution of the new row (a shift of the mean of the cluster). If the it is not sufficiently close to any of the signatures, a new signature is created based on that row. The result of this clustering is a small number (between zero and five) of candidate formant signatures for each time step when the clustering is run (every 10 or 20 milliseconds). Note that since these signatures have a frequency axis, we are referring to as spectra. However, the function of frequency is average conditional firing probability, and does not reflect the relative intensity of the actual formants in the signal (except in so far as more intense formants will be less affected by interference and thus more easily detected by synchrony).

The penultimate stage of the model attempts to form complete voices by tracking formant signatures through time. The procedure is that once a voice has been identified, the formant signature most similar to the last 'sighting' of the voice is chosen as a continuation. Exactly when to start tracking a new voice is still an unresolved issue; currently, we track only one voice at a time, and the system is given a 'seed' (a known formant of the target voice at a particular time) so that it may lock on to the correct set of signatures. The result of this tracking is a two dimensional map, a function of frequency (channel) and time, showing the contribution of each time-frequency tile to the separately-identified voice. This can then be used as a 'mask' for the time-frequency decomposition from the original linear filterbank, and subband signals masked this way can be passed through a reconstruction filter to resynthesize a sound approximating the separated voice. (This approach to resynthesis is borrowed from Brown (1992)). This is represented by the final block of figure 5, 'Mask & resynth'.

4. RESULTS FROM A PRELIMINARY IMPLEMENTATION

In this section we describe our first results with this method. We distinguish this implementation from the description in the previous section owing to some simplifications made in the interests of computational efficiency. The most significant of these is that we did not maintain the full three-dimensional histogram described above; rather, for each pair of frequency channels we tracked the score only for the two most recently observed inter-firing time skews, but not all the other possible skews. If, during a plane update, the calculated time skew to another channel was close or equal to one of the time skews being tracked, the histogram score for that skew was incremented accordingly. If the new relative timing measurement was not one of the two being tracked, the worse-scoring of those two was discarded, and the new time skew tracked instead from then on. While this modification complicates the analysis of the process, it is intended that 'correct' time skew will be detected in situations where there is a strong correlation between channels, even if there is significant interference between two voices in one of the channels.

The other simplification concerns the way that voices are formed through time. Rather than creating all possible component voices automatically, or seeding a voice at a particular time-frequency point and following it forward from there, our tests involved static vowels that could be reliably associated with a particular formant throughout their extent. It was not necessary to track; we simply formed voices by taking the vowel signature that contained the known formant at each time. Thus these results do not test the mechanism of tracking voices through time suggested for the 'Continuity tracking' block in figure 5.

Despite these simplifications, the results are in line with our intentions. Figure 8 illustrates two extended vowels both separately and mixed together. The top image of each pair shows the output of the inner-hair-cell firing probability stage, which is approximately a rectified version of the output of the filterbank. The horizontal dimension is time and the vertical axis shows the center frequency of the corresponding channels of the filterbank. The firing probability is shown as gray density, and the resulting image is similar to a wideband spectrogram, with the addition that individual cycles of the ringing of each filter can sometimes be resolved (as a result of the half-wave rectification). The second image of each pair is the same sound as represented by the firing event generator. The axes are the same as for the upper image. Here we can see how each pitch cycle gives rise to an approximately vertical structure of burst events, with the white space showing the absence of firing events between pitch bursts.

In the bottom pair of images, we see the analysis of a mixture of the vowels at approximately equal perceived loudness. The resulting sound is quite complex, although certain frequency regions are visibly dominated by one or other of the vowels, such as the 700 Hz formant for the

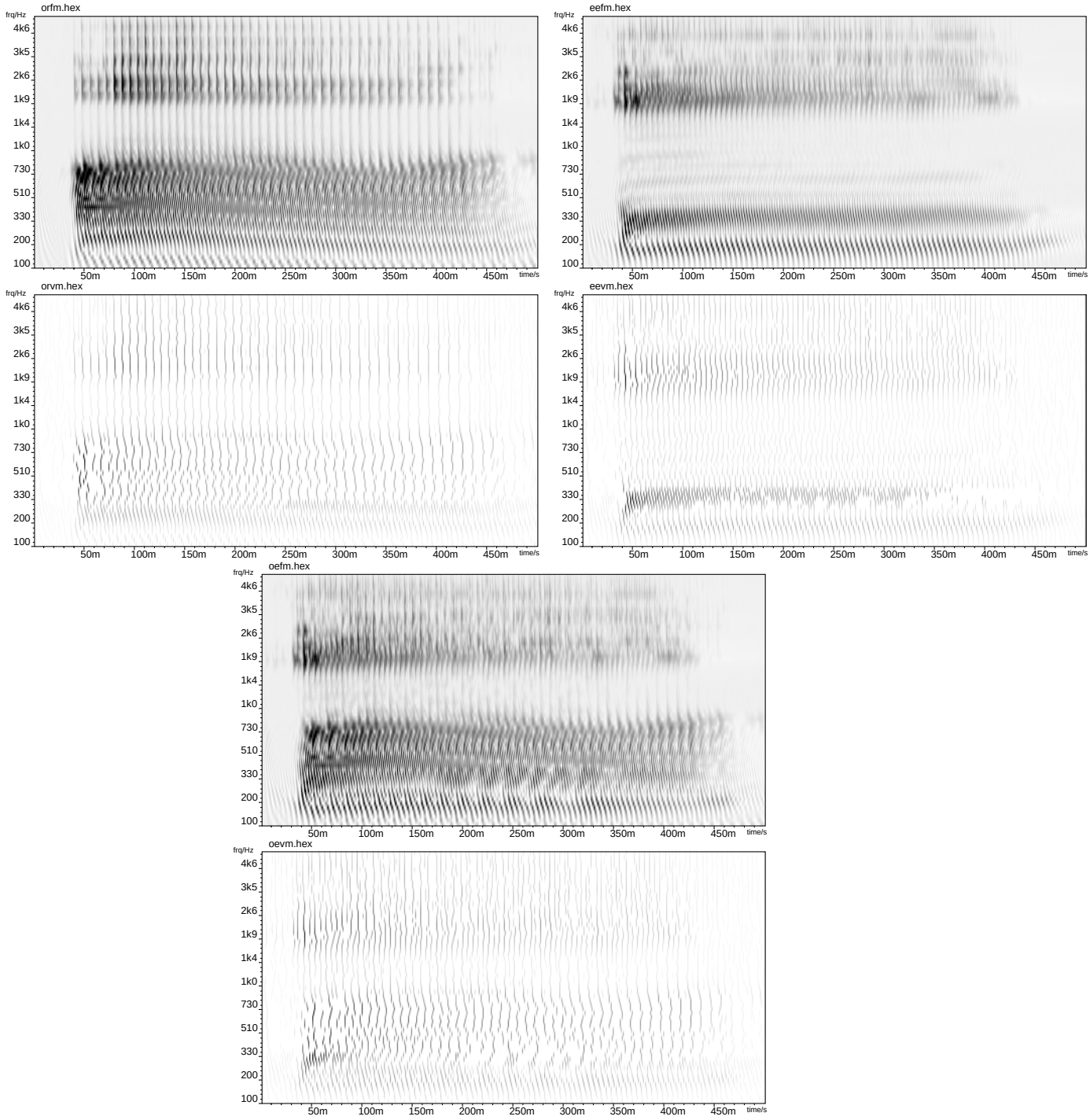

figure 8: *Firing probability (top of each pair) and firing events (bottom) for the vowels /or/ and /ee/, and their mixture.*

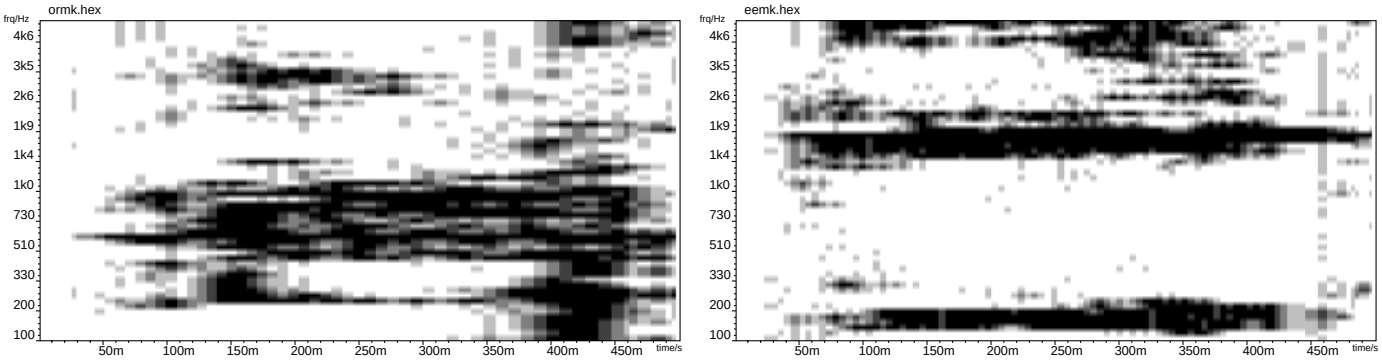

figure 9: *Extracted time-frequency masks for the vowels in figure 8.*

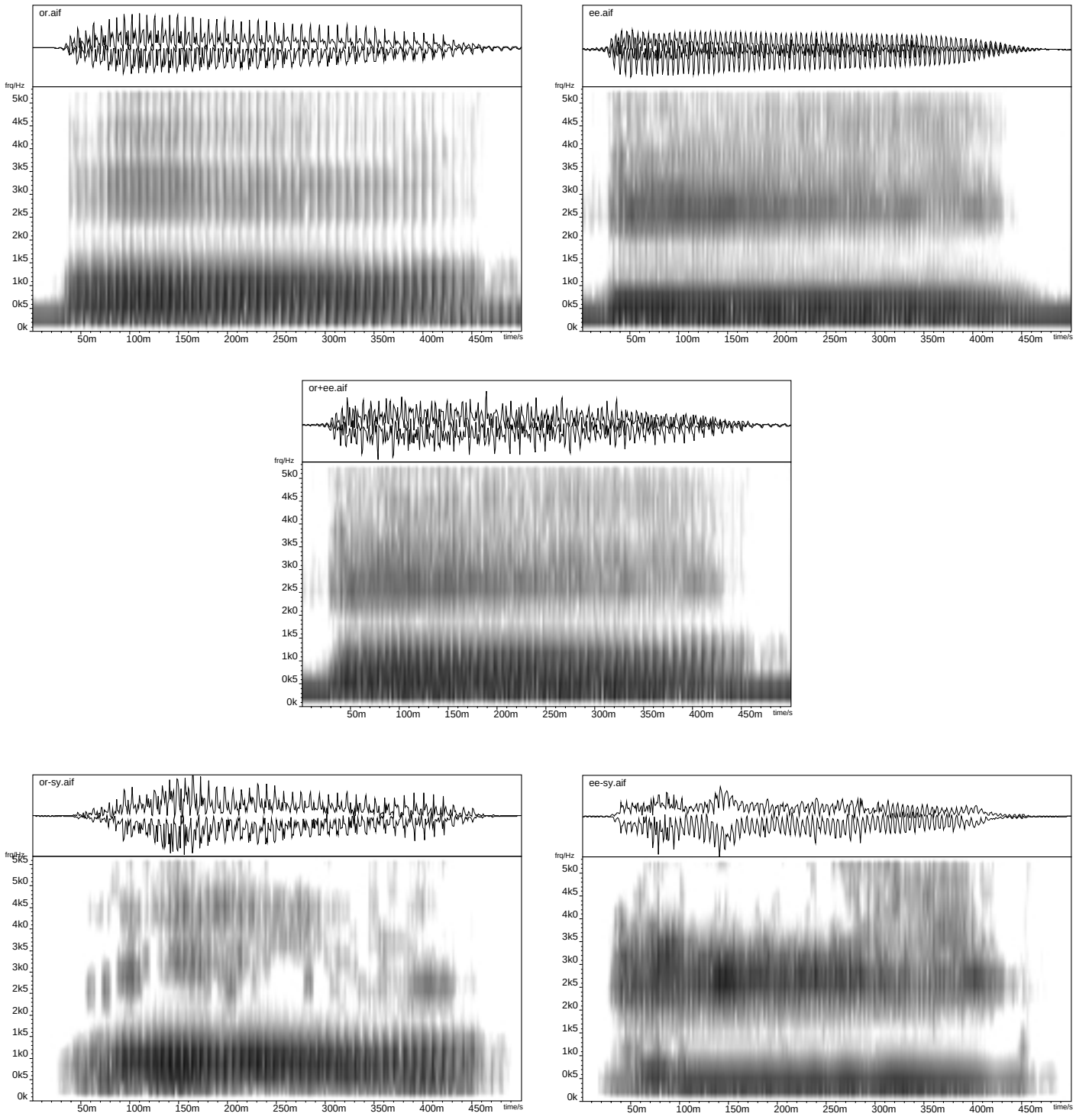

figure 10: *Wideband spectrograms for the two original vowels (top row), their mixture (middle), and the reconstructed vowels from the separation (bottom row).*

vowel /or/ shown at the left of the figure and the 2 kHz formant for /ee/.

The discriminating formants in these two regions were in fact the basis of the simplified continuity matching used in these schemes. Figure 9 shows the time-frequency masks derived for the two vowels, by selecting vowel signatures from the output of the clustering stage that exhibited the relevant formants. These masks are dark for time-frequency regions exhibiting high consistency of timing skew to the 'seed' formants, and light for areas that show negligible correlation. The masks are then multiplied by the corresponding outputs of the original cochlear filterbank, which are then combined into a single output signal by an inverse filter bank. The sounds produced by these two masks, i.e. the separated vowels, are shown in figure 10 as conventional wide-band spectrograms, along with the spectrograms of the original vowels and their mixture. As can be seen, although each separated signal is missing significant regions of energy where it was masked by the other vowel, the net result has captured much of the character of the sound, and successfully rejected the interfering sound. Listening to the outputs confirm that the separation is good, and the vowels are easily recognized, although obviously distorted in comparison to the originals.

5. DISCUSSION

We have introduced a detailed model with complex behavior. It is useful to consider some of the tuning parameters of the model and their influence.

The cochlear filterbank and inner hair cell firing probability models have a profound influence on the model behavior, since they determine what information from the raw stimulus is available to the rest of the processing. The frequency selectivity, level adaptation and signal synchrony of these stages are particularly important. However, we hope that by using the relatively established Meddis and Hewitt modules for these stages we can circumvent extensive arguments and justifications.

The simple operation of the burst-event generation stage, described in section 3, is dependent only on the time window over which it operates. This was set to 5 milliseconds to be approximately the period of the highest pitch for which the system will operate. Since at most one event can be generated within this window, this helps ensure the desired behavior of a single event per channel per pitch cycle.

In the implementation of the timing-skew histogram, there are two important parameters of the relative timing dimension: the maximum skew that is registered, and the timing resolution. In our implementation, we looked for events for 5 milliseconds on either side of the update event: Again, this is intended to capture all events within one pitch cycle, since it is difficult to imagine a reliable mechanism for tracking time skews that exceed the pitch cycle time. The resolution was about one-hundredth of this i.e. 50 microseconds. If this is made too large, then fine jitter effects may not be noticed, and the segregation ability of the process is reduced. If it is too small, it becomes overly sensitive to timing noise in the burst detection process, and also raises questions of plausibility. However, the human ability to discriminate azimuth to a few degrees based on interaural timing differences proves that this level of timing sensitivity can occur at least in some parts of the auditory system. The actual implementation of resolution in the histogram bins was side-stepped by the parameterized way in which timing was represented in our implementation. For a real histogram with actual timing bins, it would be necessary to use some kind of point-spread function when incrementing bins to avoid artifactual behavior on the boundaries between bins.

Perhaps the most interesting parameter is the sluggishness of the probability estimation process, equivalent to the time window upon which estimate is based. In the integer-only histogram implementation, the exponential behavior of the recursive estimator in equation (10) has been approximated with a straight line segment that causes the time support of the estimate to scale with the pitch cycle, such that a sudden change in the vocal tract parameters will be fully reflected in the histogram after five complete cycles. While this only approximates the form of adaptation to new evidence, and makes an implicit assumption that the spacing between events is some constant period, it possibly provides adaptive behavior more suitable to our ultimate task of separating the vowels. A more careful analysis of the nature of this approximation is required.

In the original formulation of equations (8) through (10), the sluggishness was represented by the time window of the conditional firing probability estimates. A larger α (which would be approximated by a more shallow line segment, i.e. saturation at a larger level) would give a more conservative result, requiring more instances to fully accept a particular timing. While this would be less often wrong, it would also be in danger of losing track of rapidly modulated voice. In fact, we consider the most interesting aspect of the present model to be the adaptation to evidence across time, in contrast to the pseudo-static nature of previous models. We can be quite confident that the actual perceptual system has far more sophisticated strategies for temporal evidence accumulation than simple one-pole integrators. It is interesting none the less to begin to model these effects.

Given that the GPS model relies exclusively on the temporal repetitions of between-channel patterns to make its groupings, it is worth asking whether real speech is sufficiently static to make such a principle feasible. Our current model requires some four pitch cycles with essentially constant vocal tract response to lock-on to the pattern. This is only 50 milliseconds even at low voice pitches, a reasonable minimum syllabic duration. However, transitions can be shorter than this, so it will be interesting to see how a complete system, including voice tracking through time, deals with speech that modulates at realistic rates. With very fast transitions, the speech may be chopped into sections whose internal connection is reasonably evident, but with no clear continuity across the transitions. However, other mechanisms may be postulated to mend these discontinuities : two sections of speech with abutting end and start times are showing evidence of common origin.

In section two we explored the usefulness of jitter, arguing that a modulation detector with a variable threshold but limited time resolution would be able to adjust itself to an optimal rate of detected events if the modulation was governed by a Gaussian jitter function (with gradual roll-off at the extremes of the modulation probability density function), but such adjustment would be difficult or impossible for sinusoidal modulation with its sharp, bimodal distribution. There is another qualitative argument that may be made in favor of the utility of jitter, in the context of a sensitive relative-timing detection scheme such as the current one. If a periodic signal is composed energy in different frequency bands which show a consistent relative timing, then a system to detect relative timing will correctly observe the signal, as will an alternative system that first calculates periodicities within each channel (such as the Meddis & Hewitt model). However, if the signal departs from periodicity by some jitter mechanism that, say, elongates a particular cycle, the relative-timing-based scheme will not be in the least perturbed, since the delayed glottal pulse will most likely still excite the same relative timing amongst the formants — each channel will have its burst delayed by the same amount, and the pattern is preserved. If the system is attempting to segregate two voices that are close in pitch, this consistency of jitter-derived displacement of one voice's formants can effectively disambiguate the voices where period measurements fail. Thus the presence of jitter, and a detection scheme that is impervious to its effects, can be particularly crucial for the perceptual isolation of voices with poor pitch separation.

It is important to recognize that the current model is completely insensitive to pitch cycle length and its stability. This is quite unrealistic, since random jitter at high levels will rapidly disrupt the stability and clarity of a sound; something approximating smooth pitch variation is important for the perception of voice. Since the GPS model cannot provide this behavior, it cannot be considered a sufficient model of voice separation. Rather, it may be one of a variety of techniques available to the auditory system, each used when appropriate to the particular problem at hand. Periodicity-based methods may comprise other alternatives, and there may be still other principles. This is entirely consistent with a theory of perception and cognition based on corroboration between multiple, disparate mechanisms of indifferent individual reliability, but capable of excellent performance when successfully combined (a 'society' in the sense of Minsky (1986)).

A final comment regards the physiological plausibility of the method. One inspiration for the design was the known cross-correlation mechanisms identified in bat echolocation systems by Suga (1990). There, correlations between certain features of a reflected sonar pulse were mapped across two dimensions of individual neurons. The three-dimensional histogram of the GPS model could perhaps be implemented with the same building blocks, but the extra dimension certainly reduces its likelihood of it being a direct description of a neurological system since the number of elements involved would be considerable. It might be that a low-resolution time-skew axis would make the third dimension relatively shallow, yet still retain useful behavior. The 'two-and-a-half' dimensional modification we actually used in our implementation (where the timing skew is represented explicitly rather than implicitly in an array of bins) requires less computational complexity, but a different kind of representation that compromises its plausibility also. It may also be unnecessary to compare *every* frequency channel against every other, but instead reduce computation by using a local window to calculate a diagonal 'band' from the middle of our histogram. This would still find adjacent pairs of formants (provided both lay within the window). These could then be assembled into complete vowels by transitivity.

6. CONCLUSIONS AND FUTURE WORK

Future tests of the model

This report describes a very preliminary stage of development of this model. There are a number of tests of immediate interest whose results would guide further development and refinement. There are also some interesting psychoacoustical experiments suggested by the presentation of this algorithm as a potential strategy employed by the brain. We will describe some of this planned future work.

The one example we tried was not terribly difficult: the two vowels were very distinct in formant structure, and they had wide pitch separation. However, the prediction made in section 5, that exploiting jitter should allow segregation even with little difference in fundamental frequency, is relatively easy to test and could reveal a particularly advantageous application of this technique.

While Meddis & Hewitt's model achieved an extremely good match to experimental results on the confusion of vowels as a function of pitch separation by actual listeners, their artificial stimuli lacked the kind of jitter that should be very important in such situations, but which is not well detected by their model. Our contention that the currently proposed system gains advantage by exploiting this jitter would be tested by generating sets of artificial vowels with and without simulated jitter of various intensities and types. We would expect the GPS model to perform much better in the presence of wideband jitter, and the MH model to perform perhaps a little worse compared to unmodulated vowels. To be able to conduct the same kinds of tests as described in the Meddis and Hewitt paper, we would need to add some kind of vowel classification stage to the output of the GPS. This could be done in several ways, for instance with a low-order cepstral analysis of the separated, resynthesized signal.

The perceptual assertion upon which this entire paper is founded is that the human auditory system is indeed sensitive to correlated modulation as caused by glottal pulse train jitter. We have argued circumstantially for this, but it would be more satisfying to have direct psychophysical evidence. One test would be to use the same collection of synthetic vowel combinations with various degrees of modulation of different kinds, and to test human ability to segregate as a function of these parameters. This

evidence could provide very strong validation of the physiological relevance of our model, or alternatively demonstrate its irrelevance!

Another prediction of this kind of mechanism is that excessive phase dispersion should degrade the coherence of a signal; if there is more than one pitch-cycle of additional timing skew between different formants, then it becomes much harder to correlate between appropriate pairs of events (originating from the same glottal pulse). Although this could still be achieved with a sufficiently large time dimension for the three-dimensional histogram, short cuts of the kind we actually used in section 4 would not be applicable. We know that moderate phase dispersion has little impact on the perceptual quality of a vowel, but it would be interesting to investigate the point at which dispersion becomes deleterious, and the perceptual characterization of the impact.

On a similar theme, and in the spirit of Summerfield & Assmann (1991), it would be possible to generate synthetic vowels where each formant was excited by a separate pulse train. If these pulse trains had the same underlying period but different superimposed jitter functions, the resulting synthetic voice would have average jitter characteristics similar to those of real speech, but without any of the coherent timing skew detected by the GPS model. Such sounds would be poorly fused by the model; it would be interesting to test if they were similarly 'fragile' for human listeners.

Conclusions

(1) An argument was made that for certain possible perceptual mechanisms the random nature of cycle-to-cycle variation in natural speech could be materially different from the sinusoidal modulation typically employed in experiments.

(2) We described in some detail an algorithm by which the human auditory system might be able to segregate the contributions of each of several voices in an acoustic mixture based upon the coherent timing skew between different regions of spectral dominance of a single voice. A particular aspect of this strategy is that it must combine evidence over time (i.e. several pitch cycles) in a nonlinear fashion to form a confident output.

(3) Results from a preliminary implementation of this model tested with real voice samples show a surprising ability to segregate voices on this basis alone.

There can not, however, be any doubt that other cues such as voice pitch are extremely important to the perception and segregation of voices. Therefore we propose that if the current model does in fact reflect some genuine aspect of human auditory processing (as our future tests may reveal), it is only one of a collection of voice-segregation methods available to human listeners. Future research should concentrate not only on the discovery and refinement of individual techniques, but also on the interesting question of how the auditory system might combine the results of different processes.

ACKNOWLEDGMENT

This work was generously supported by the MIT Media Laboratory, in particular the Television of Tomorrow consortium. The author is in the United States as a Harkness Fellow of the Commonwealth Fund of New York, whose support is gratefully acknowledged.

We would like to thank Dr. Lowel P. O'Mard and the Speech and Hearing Laboratory at the Loughborough University of Technology for making their 'LUTEar' auditory model software available, which constituted a significant portion of this project.

REFERENCES

Assmann, P. F. and Summerfield, Q. (1989). "Modeling the perception of concurrent vowels: Vowels with the same fundamental frequency," JASA 85(1), 327-338.

Assmann, P. F. and Summerfield, Q. (1989). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," JASA 88(2), 680-697.

Brown, G. J. (1992). "Computational auditory scene analysis: A representational approach," Ph.D. thesis CS-92-22, CS dept., Univ. of Sheffield.

Carlyon, R. P. (1991). "Discriminating between coherent and incoherent frequency modulation of complex tones," JASA 89(1), 329-340.

de Cheveigné, A. (1993). "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," JASA 93(6), 3271-3290.

Colburn, H. S. (1977). "Theory of binaural interaction based on auditory-nerve data. II. Detection of tones in noise," JASA 61(2), 525-533.

Cook, P. R. (1991). "Identification of control parameters in an articulatory vocal tract model with applications to the synthesis of singing," Ph.D. thesis, CCRMA, Stanford Univ.

Duda, R. O., Lyon, R. F., Slaney, M. (1990). "Correlograms and the separation of sounds," Proc. IEEE Asilomar conf. on sigs., sys. & computers.

McAdams, S. (1984). "Spectral fusion, spectral parsing and the formation of auditory images," Ph.D. thesis, CCRMA, Stanford Univ.

McAdams, S. (1989). "Segregation of concurrent sounds. I: Effects of frequency modulation coherence," JASA 86(6), 2148-2159.

Meddis, R. and Hewitt, M. J. (1991). "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," JASA 89(6), 2866-2882.

Meddis, R. and Hewitt, M. J. (1992). "Modeling the identification of concurrent vowels with different fundamental frequencies," JASA 91(1), 233-245.

Minsky, M. (1986). *The Society of Mind*, (Simon and Schuster, New York).

Moore, B. C. J. and Glasberg, B. R. (1983). “Suggested formulae for calculating auditory-filter bandwidths and excitation patterns,” JASA 74(3), 750-753.

Parsons, T. W. (1976). “Separation of speech from interfering speech by means of harmonic selection,” JASA 60(4), 911-918.

Patterson, R. D. and Moore, B. C. J. (1986). “Auditory filters and excitation patterns as representations of frequency resolution,” in *Frequency Selectivity in Hearing*, edited by B. C. J. Moore (Academic, London).

Ross, S. (1982). “A model of the hair cell-primary fiber complex,” JASA 71(4), 926-941.

Suga, N. (1990). “Cortical computational maps for auditory imaging,” Neural Networks 3, 3-21.

Summerfield, Q., and Assmann, P. F. (1991). “Perception of concurrent vowels: Effects of harmonic misalignment and pitch-period asynchrony,” JASA 89(3), 1364-1377.

Summerfield, Q. and Culling, J. F. (1992). “Auditory segregation of competing voices: absence of effects of FM or AM coherence,” Phil. Trans. R. Soc. Lond. B 336, 357-366.