MIT-CSAIL-TR-2005-005
AITR-2005-001

January 28, 2005

Determining Articulator Configuration in Voiced Stop Consonants by Matching Time-domain Patterns in Pitch Periods

Attila Kondacs

Determining articulator configuration in voiced stop consonants by matching time-domain patterns in pitch periods

by

Attila Kondacs
MSc in Mathematics, Eötvös University, Hungary
Diploma in Economics, University of Cambridge, UK

Submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2005

Certified by: Gerald J. Sussman
Matsushita Professor of Electrical Engineering
Thesis Supervisor

Accepted by: Arthur C. Smith
Chairman, Department Committee on Graduate Students

Determining articulator configuration in voiced stop consonants by matching time-domain patterns in pitch periods

by

Attila Kondacs
MSc in Mathematics, Eötvös University, Hungary
Diploma in Economics, University of Cambridge, UK

Submitted to the Department of Electrical Engineering and Computer Science on January 7, 2005, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science and Engineering

Abstract

In this thesis I will be concerned with linking the observed speech signal to the configuration of articulators.

Due to the potentially rapid motion of the articulators, the speech signal can be highly non-stationary. The typical linear analysis techniques that assume quasi-stationarity may not have sufficient time-frequency resolution to determine the place of articulation.

I argue that the traditional low and high-level primitives of speech processing, frequency and phonemes, are inadequate and should be replaced by a representation with three layers: 1. short pitch period resonances and other spatio-temporal patterns 2. articulator configuration trajectories 3. syllables. The patterns indicate articulator configuration trajectories (how the tongue, jaws, etc. are moving), which are interpreted as syllables and words.

My patterns are an alternative to frequency. I use short time-domain features of the sound waveform, which can be extracted from each vowel pitch period pattern, to identify the positions of the articulators with high reliability. These features are important because by capitalizing on detailed measurements within a single pitch period, the rapid articulator movements can be tracked. No linear signal processing approach can achieve the combination of sensitivity to short term changes and measurement accuracy resulting from these nonlinear techniques.

The measurements I use are neurophysiologically plausible: the auditory system could be using similar methods.

I have demonstrated this approach by constructing a robust technique for categorizing the English voiced stops as the consonants B, D, or G based on the vocalic portions of their releases. The classification recognizes 93.5%, 81.8% and 86.1% of the *b*, *d* and *g* to *ae* transitions with false positive rates 2.9%, 8.7% and 2.6% respectively.

Thesis Supervisor: Gerald J. Sussman
Title: Matsushita Professor of Electrical Engineering

Acknowledgments

My many thanks go to Gerry Sussman and Tom Knight - these remarkable individuals acted as my supervisors. They insisted that I should do something I am really interested in, and have exhibited infinite patience and faith in that one day I would produce something useful (perhaps hoping that this day would come before we exhaust all our funding options). Without their help I would not have written this thesis. I would also like to thank my other two committee members, Denny Freeman and Jim Glass, whose help was vital at the project's various phases. I am also thankful to the Computer Science and Artificial Intelligence Lab (CSAIL) at MIT, the Department of Mathematics at the Eötvös University in Budapest, and my high school class in the Arpad Secondary School in Budapest for providing vibrant intellectual environment that paved the way to this thesis. The ultimate blame for this thesis, however, lies with my parents.

It has been rather difficult to find funding for this research, as it did not fit into any of the already existing projets at MIT. I am grateful for support coming from the Grass Instrument Company; the Deshpande Center at MIT (Non-linear time-domain front-end for speech understanding, Tom Knight); U.S. Department of Interior - Fort Huachuca (Natural tasking of robots on human interaction cues; Rodney Brooks); the University of Singapore (Faculty Fellowship, Leslie Kaelbling); two MIT EECS teaching assistanships; NSF grant #EIA-0130391 (Robust engineering using biologically inspired cell-differentiation and morphogenesis, Radhika Nagpal); NSF grant #6891318 (CISE Experimental Partnerships: Mit Raw project, Anant Agarwal); and the Nippon Telegraph and Telephone Corporation, and my savings account.

Contents

1 Introduction **7**

2 Problems With the Traditional Way **9**

2.1 Stationarity and Frequency Analysis 9

2.2 Do We Really Hear Phonemes? 11

2.3 Non-Linearity or A Priori Knowledge 13

3 The Auditory System **15**

3.1 The Outer and the Middle Ear 15

3.2 The Cochlea 17

3.3 The Auditory Nerve 21

3.4 Temporal Attributes of Processing in the Auditory Nerve 23

3.5 Time-Domain and Auditory Front Ends in Automatic Speech Recognition 29

4 Patterns **31**

4.1 Grouping Principles 31

4.1.1 The dot abstraction 32

4.2 Energy 36

4.3 Time-Patterns 39

4.4 Patterns 40

5 Observations **42**

5.1 Sub-Pitch-Period Time-Patterns Following Voiced Stops 42

5.1.1 Aeroacoustic flow in the vocal tract 44

6 Experiments **48**

6.1 The algorithm 48

6.1.1 Linear Pre-Processing 48

6.1.2 Time and channel-space segmentation 49

6.1.3 Feature extraction 62

6.1.4 Categorization 67
6.2 Experiments and Statistical Results 68
6.3 Prior work on stop consonants 71

7 Summary 74

A Response time of low-pass filter with sharp cutoff frequency 77

B Sub–pitch-period time-pattern examples 86

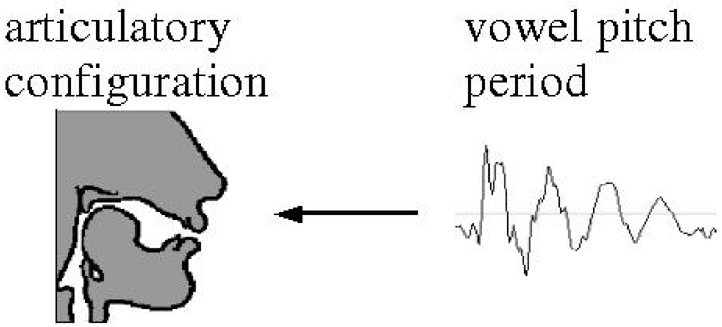
Chapter 1

Introduction

Signal processing front ends currently in use do a poor job of differentiating phones. It is a testament to the quality of the high-level algorithms that current speech recognition systems work at all. Without the high level algorithms, speech recognizers have phone error rates in the 10-20% range under perfect conditions, and 40-60% in noise. Based on the front-end output, every second word of noiseless speech and every word of noisy speech would be misunderstood. The recognizers' performance also breaks down when presented with previously unheard speakers and dialects. The striking difference between machine and human performance [33] suggests that we have not yet fully understood what really differentiates particular sounds, and how our auditory system processes those sounds.

In this thesis we will challenge two core assumptions of the traditional speech processing approach: one is the use of frequency as the primitive of the low level representation for speech. The other is that speech can be segmented into phonemes and thus that phonemes are useful high level representation units. We will argue instead that a representation based on the following three layers of primitives is better: syllables as high level units, articulator configuration as mid-level primitives and pitch-period resonances or spatio-temporal patterns as low-level primitives. In this more robust representation vowel pitch-period resonance patterns are used to determine the current articulator configuration. Trajectories of articulator configurations are then interpreted as syllables and words.

To support some of the ideas above we have built a system that extracts attributes of pitch-period patterns and uses them to robustly differentiate between three articulator configurations: the ones that



immediately follow the burst of English stop consonants *b*, *d* and *g* as they are released into certain vowels. We will demonstrate how the system performs on a subset of the TIMIT database.

Chapter 2

Problems With the Traditional Way

2.1 Stationarity and Frequency Analysis

The current popular model of speech sound production is: glottal pulses $\rightarrow F \rightarrow$ sound, where F is the transfer function of the vocal tract. Frequency-domain analysis is based on the assumption that F is *semi-stationary* – changing so slowly that it can be regarded as constant for short time segments. This is not true. The tongue may move rather quickly and often assumes the position of a particular phone for only one or two glottal pulses, or only "points" or gestures toward that position. There have been reports on vowel recognition with accuracy close to normal even when there is only one glottal pulse or a section of a glottal pulse present [17, 37]. Frequency-domain analysis of such short vowel segments is highly inaccurate; the resulting spectra contain strong side-lobes depending on the exact placement (in time) of the short time Fourier-transform windows.

To better understand how rigid and unreliable the Fourier-transform based frequency analysis is in the face of the rapidly changing resonances of the speech signal, look at figure 2.1! Here we can see the band-pass filtered channels of three consecutive pitch periods. During each pitch period of the higher frequency channels, short bursts of energy repeat. These bursts are reasonably close to and can be modeled as pure amplitude modulated and phase shifted sinusoid segments of the same frequency. If the phase shift between consecutive sinusoid bursts is close to half their period then their combined contribution to a con-

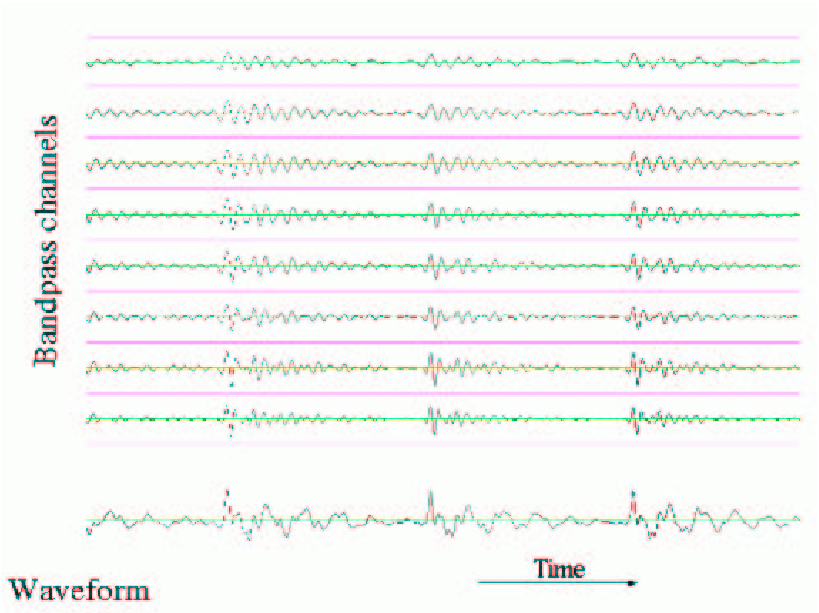


Figure 2.1: Time-aligned waveform and band-pass filtered channels. The center frequency of the bandpass channels is decreasing towards the top.

volution with any sinusoid at their frequency will cancel out (see figure 2.2). Consequently, if the window of the *discrete Fourier-transform (DFT)* covers the two sinusoid bursts in the signal their combined contribution to their detected frequency may cancel out. In other words, no matter how large the individual peaks of the energy bursts are, they may be not detectable by the DFT. Due to the *uncertainty principle* [40, chapter 8] we cannot arbitrarily shorten the window of the DFT without corresponding loss of frequency resolution. In fact, the shortest windows we can find applied in the literature are around 22 ms, whereas the average pitch period length is somewhere around 9 ms for male speakers. Thus the DFT window will not only cover neighboring sub-pitch-period bursts, but also whole neighboring pitch periods. This means that waveforms f neighboring pitch periods may interfere and cancel out due to phase shifts; that is, the exact frequency content detected by the DFT is dependent on the local pitch period length and the exact temporal placement of the DFT window. As both of the

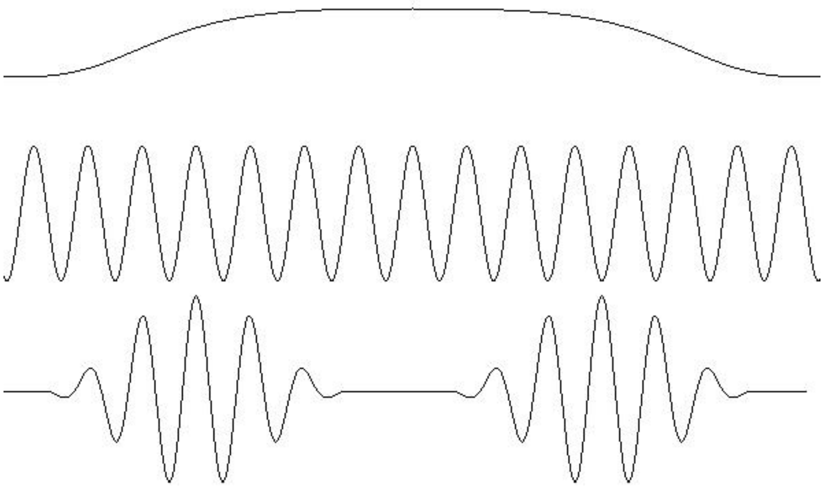


Figure 2.2: An example setup of how two sinusoid bursts can completely cancel out their contributions to the discrete Fourier transform at a certain ω frequency. Top: DFT window (Gaussian); middle: ω frequency sinusoid; bottom: two short ω frequency sinusoid bursts with half period phase shift. Note how the corresponding peaks in the bursts coincide with opposite sign values of the sinusoid in the middle. As the window function is symmetrical for the two sinusoid bursts, the convolutions of the sinusoid bursts with the windowed sinusoid will add to zero.

latter are independent of what vowel sound we hear when we listen to the signal, this is a highly undesirable dependence, and it shows that frequency may indeed not be the ideal primitive in speech analysis.

2.2 Do We Really Hear Phonemes?

The second problem with stationarity is that it enforces the view that the mid-level building blocks of speech perception are phonemes. Consider the vowels in utterances *lass* and *loss*. In both cases the articulators start from the *l* configuration. Then the tongue moves down and forward in *lass* and down and backward in *loss*, the lips round and the jaws open to different degrees, then the articulators go into the *s* position. If we pronounce these words extremely sloppily we usually just move our tongue down a little from the *l* position and open our jaws slightly during both vowels (the position not far from the position of the vowel in the word *but*). The preceding and the following positions are identical. Thus in case of very lazy articulations the trajectories that the articulators run through during the two vowels are virtually identical. This must mean that the acoustic signals that reach our ears are very close to each other as well. The puzzling part is, however, that even during these mumbled utterances of *lass* and *loss* we hear two distinctly different vowels – the reader is encouraged to try this very simple experiment with a partner! Such overlaps of the acoustic space point to the importance of context and make it impossible to uniquely define a mapping from the acoustic space to the set of phonemes. This severely undermines the role of phonemes as high-level primitives of speech perception.

Sound omissions, dialects and accents, non-time-aligned articulator changes – such as nasalization starts (i.e. the velum is lowered) before the closure for “m” – are all ubiquitous and hardly interfere with human speech understanding. They, however, further complicate the much sought-after mapping from the acoustic space to the set of phonemes. Thus the potential role of phonemes in perceptual organization is rather weak.

There is mounting evidence in the speech psychophysiology community that the ability to differentiate phonemes is learned with alphabetic reading and that the actual basic units of speech perception are *syllables*. Morais at. al. [26] showed that illiterate adults cannot segment utterances phonetically. In [55] Warren demonstrated that phonemes replaced by noise are indistinguishable from non-missing phonemes. Savin and Bever [18] showed that the perception of syllables and words takes place before listeners can identify phonemes and concluded that access to syllables occurred first and identification of the phonemes followed recognition of the syllable. In another set of experiments [42] native speakers of English perceived sequences of short vowels as valid syllabic units of English. These observations indicate that segmentation of speech into phonemes is preceded by recognition of syllables (see a much more detailed review in [57, pages 166-177]).

Our underlying assumption in this thesis will be that the human auditory pathway first maps the acoustic signal into articulator-configuration trajectories and then interprets the trajectories as a sequence of syllables. Recovering the articulator trajectories first not only solves all the problems listed at the beginning of this section but it also helps dealing with noise. The physical continuity of the articulators movements can be used to constrain the number of possible interpretations when parts of the signal are masked by noise.

2.3 Non-Linearity or A Priori Knowledge

Human perception breaks down the continuous and multi-dimensional acoustic space that utterances occupy into a finite number of discrete perceptual units. This phenomenon is called *category theory* in artificial intelligence or *digital abstraction* in engineering: we perceive firm boundaries where there is a continuous transition from one pattern to another. Putting an utterance into a category is a highly non-linear jump that we can think of as finding the template that is closest to some processed version of the observed signal. The templates represent expectations or a priori knowledge. They are another powerful tool that helps interpret the utterance in noise.

Such non-linear jumps to discrete interpretations are successfully applied in traditional speech recognizers in the higher level algorithms. They rely on context to prune the possible options in top-down dictionary searches using hidden Markov models and other elaborate statistical learning algorithms [27]. The popular pre-processing algorithms are, however, "dumb" in this sense: they apply either no or very little a priori knowledge early in the processing chain. This is a problem because by the time the powerful statistical techniques are applied, a lot of the fine-grained information may be unrecoverably lost through application of the discrete Fourier transform and other feature extraction steps, as we demonstrated in section 2.1. Another problem that is introduced by applying the "smart", non-linear statistical algorithm late in the processing is that by that time the useful part of the audio data is hopelessly mixed together with noise. Often in real life the noise happens to be a second speaker's voice. Any linear pre-processing algorithm will produce a set of features in which the values are weighted sums of the feature values that would result from processing the speech of the individuals separately. This makes it rather hard to imagine that an algorithm that applies non-linear steps relatively late could succeed in realistic situations.

A preprocessing algorithm that maps acoustic signals to articulator positions pitch period by pitch period in vowels would be highly beneficial on several grounds. We could establish pitch-period-based categories for the various articulator configurations, and non-linear template matching could be applied at a much earlier stage of processing than present practice; before much of the information in the signal is destroyed by feature extraction. It would also make a second stage of non-linear processing possible: one that uses the articulator trajectories to find syllables. This second stage would disambiguate the separation of the currently seemingly overlapping domain categories of the (*acoustic signal*) → *phoneme* mapping by embedding them in the much more sparse (*acoustic signal trajectories*) → *syllables* space. It would also, in effect, double the benefits of such non-linear mappings: we could rely on the continuity of the articulators' movements, the finite number of syllable articulator trajectory templates and the finite number of pitch-period templates corresponding to articulator positions to enhance speech recognition.

The second stage – mapping articulator movements into syllables – is certainly doable, as was shown by Sam Roweis [44]. In this thesis we will focus on the first stage and make one step in the direction of mapping pitch-period resonances to articulator configuration. First we will describe the mammalian auditory physiology in some detail to demonstrate that the measurements we use could be the processing steps applied by the auditory system. Then we will define an energy and a pattern concept and show observations that link the articulator configurations that follow the English *g*, *d* and *b* stop consonants as they are released into an *ae* sound (as in *bad*) to sub-pitch-period patterns of the vowel. We will go on to describe an algorithm to find pitch periods and extract certain attributes from the pitch-period resonances, and present statistical results that support our claimed link between the resonance patterns and the articulator configurations. Finally we will summarize our findings and hypothesis and give future directions.

Chapter 3

The Auditory System

We are going to give a dense and partial review of the mammalian peripheral auditory system. Except for some basics on the auditory nerve, we are going to skip the vast amount of knowledge about the auditory pathway in the brain. Hopefully we will still attain the single goal of this chapter: to show that the auditory system is capable of utilizing very short, fine frequency and time resolution patterns when it is dealing with vowel-like sounds.

Even though in this thesis we are trying to talk the reader out of using frequency, here we will have to resort to frequency analysis language in order to describe the auditory system simply because it has been the language of discourse in practically all of hearing research.

3.1 The Outer and the Middle Ear

Our ears have three parts, the outer ear, the middle ear and inner ear. The outer ear consists of the pinna, the ear canal and the eardrum or *tympanic membrane*. The outer ear has two roles in transmitting the sound to the eardrum. It helps sound localization by distorting frequencies in a source direction-dependent way and it amplifies the sound pressure at the eardrum by resonances.

Middle ear transmits vibration from the low impedance air to the higher impedance cochlear fluid. It achieves this through a lever mechanism of three small bones, the *ossicles* (also called *malleus*, *incus* and *stapes*). The ossicles are attached to the eardrum at one end and to the *oval window* of the cochlea at the other. This coupling mechanism reduces the energy of sound reflection that would be above 99% on the air–fluid border. It also acts as a mechanical lever through three

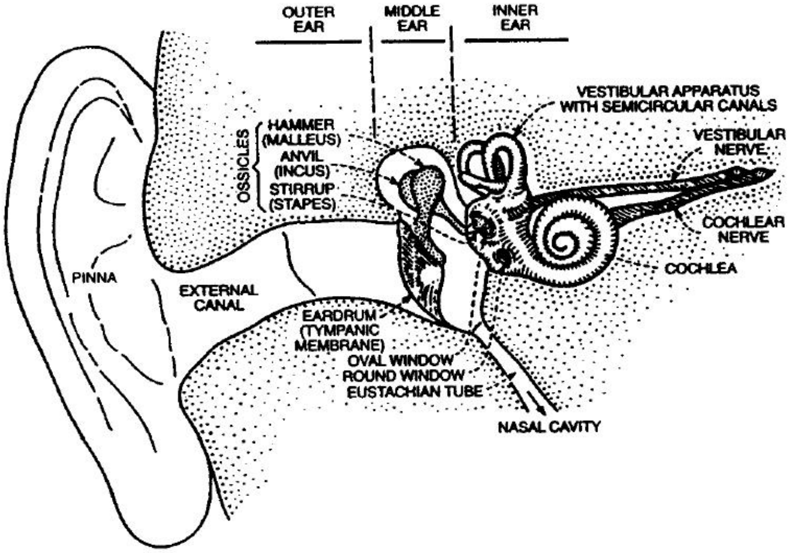


Figure 3.1: Outer, middle and inner ear. From Flanagan,1972 [11]

principles: the area of the eardrum is larger than the area of the oval window, thus the transmission of force from one to another will increase pressure at the oval window. The lever action of the ossicles acts to increase force and decrease velocity at the oval window, and likewise the curvature of the eardrum causes it to buckle under pressure, which increases the transmitted force and decreases speed.

Transmisson through the middle ear depends on frequency. It is also affected by the middle ear muscles that can constrain the motion of the ossicles. They may serve as automatic gain control of low frequency sounds in some restricted intensity range, protecting the cochlea from too loud stimuli and correcting the masking effects of low frequency resonances on high frequency sounds. If we ignore the gain control mechanism—which occurs only for either extremely intense sounds or in limited frequency range—the combined tranformation that the outer and middle ear perform on the signal can be modeled as a linear filter.

3.2 The Cochlea

The inner ear contains the *cochlea*, the organ of hearing. It is a coiled tube divided lengthways into three scalae: the *scala vestibuli*, the *scala media* and the *scala tympani*. The two outer scalae, the scala vestibuli and scala tympani are filled with intracellular fluid called *perilymph* and are separated along almost all the length of the cochlea by the scala media; they communicate at the apex of the spiral through an opening called the *helicotrema*. When the the oval window at the base end of the scala vestibuli is moved inward by the stapes the almost incompressible perilymph causes the round window at the base of the scala tympani to flex outward. The scala media or *cochlear duct* is seperated from the scala tympani by the *basilar membrane*, and from the scala vestibuli by the *Reissner's membrane*. Most of the volume between these two membranes is filled with positively charged intracellular fluid, *endolymph*. The scala media has a closed end at the helicotrema (at its apex). A third fluid called *cortilymph* is found in the *tunnel of Corti* inside the scala media. Some of these fluids have different electric potential which seems to play a role in activating the processes that send signals to the auditory nerve.

The auditory transducer, the *organ of Corti* sits on the basilar membrane within the scala media. It contains the transducing cells called *hair cells*. There are two types of hair cells, the *inner hair-cells* and the *outer hair-cells*. Each hair cell has many hairs or *stereocilia* protruding from its apical surface. There are about 12500 outer hair-cells arranged in parallel rows and about 3500 inner hair cells forming a single row, each having around 100-200 and 50 stereocilia respectively. The stereocilia of a single hair cell are linked so that they tend to move together as the their surrounding fluid moves. As sound waves send pressure waves up the scala vestibuli through the transmission mechanism of the middle ear, the pressure difference between scala tympani and scala vestibuli moves the basilar membrane. Deflection of the basilar membrane causes deflection of the fluids in the cochlear duct, which leads to a shearing motion of the stereocilia of the hair cells. This causes electromechanical changes within the hair cells, that lead to stimulation of the auditory nerve fibres that are attached to the receptor cells.

The flexibility and width of the basilar membrane changes as we advance up the spiral of the cochlea. It is stiffer (i.e. it is displaced less by unit force) at its base, near the oval and round windows, and grows gradually less stiff as we approach the apex. Its width is about 0.04 mm at the base and 0.5 mm at the apical and next to the helicotrema. The widening basilar membrane results in larger and larger mass being moved by vibrations. As we move apically along the spiral the interaction between mass and stiffness results in decreasing frequency value for which the magnitude of the mechanical admittance peaks locally. In other words, the further away we are from the base the lower the excitation frequency for which the particular area will respond with maximal amplitude fluctuations. For a particular frequency the place of maximum response on the basilar membrane is called the *resonance point.* The mechanical admittance for a particular frequency increases up to the resonance point and decreases beyond it. This frequency selectivity along the basilar membrane resembles coarse resolution Fourier analysis and serves as the basis of frequency selectivity of the rest of the auditory system.

When the cochlea is excited by a pure tone a *travelling wave* (figure 3.3) appears along a section around the resonance point of the tone. The travelling wave advances along a segment of the basilar membrane apically and moves the membrane points up and down with phase-shifted identical frequency. It has a constant-amplitude envelope that has a very sharp peak at the resonance point. The frequency response of a point on the basilar membrane looks like that of a low-pass filter with an incredibly steep fall-off at the cutoff frequency. Researchers have had a hard time reproducing the sharp peak of the travelling wave and the related steep cutoff frequencies as purely mechanical responses. When the blood supply of the cochlea is shut off temporarily, the sharpness of the response quickly diminishes. The damage is recoverable if the oxygen shortage is not very long, but after a while it

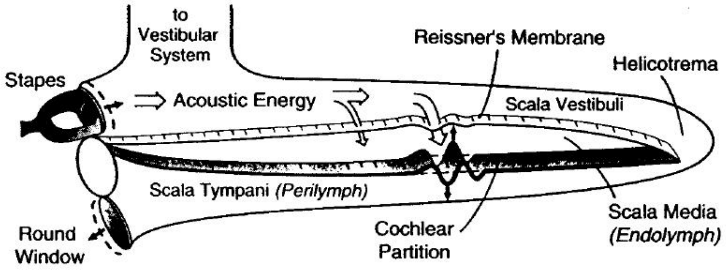


Figure 3.2: Straightened view of cochlea, showing the flow of acoustic energy. From Geisler 1998 [14]

becomes permanent and the peaks of frequency response of the basilar membrane become flattened [48]; identical with the broad low-pass filter responses found by Békésy [54] in dead animals. This suggests that there is active mechanical amplification that enhances the sharp peak responses of the travelling wave and frequency tuning at the resonance points. Although this process is not understood fully, it is widely believed that there is a non-linear feedback mechanism guiding the latter frequency selective amplification that acts through the motility of outer hair-cells. There are other phenomena that point to the existence of active mechanical processes in the cochlea. The ear can emit narrow band tone-like sounds. This "ringing of the ear" can be so loud that people nearby can hear it [24]. Another is called *two tone suppression*: when two tones whose frequencies are close to each other are presented simultanously, the louder tone will reduce the maximal response amplitude of the softer tone. This effect is usually attributed to lateral suppression through the mechanical constraints that arise because of the closeness of the resonance points on the basilar membrane (see [14] for more details).

There exist alternative theories about how frequency resolution is achieved in the cochlea. A particularly interesting one is that the stereocilia themselves act as fine tuned resonators in response to the sound waves that travel through the fluids of the organ of Corti. Although the inner hair-cells are uniform and thus cannot exhibit frequency selectivity, the size of the outer hair-cells gets gradually larger toward the apical end of the cochlea, yielding progressively lower resonant frequencies near the apex [32]. Dancer [9] presented evidence involving time

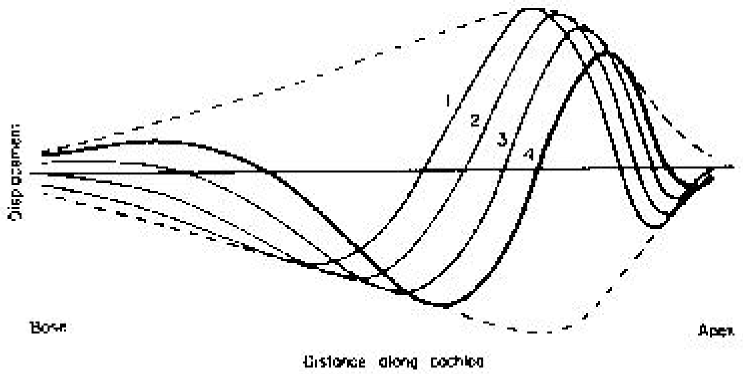


Figure 3.3: The travelling wave. The full lines show deflection of the cochlear partition at successive moments, as numbered. The waves are contained in an envelope which is static (dotted line). From Békésy, 1960 [54]

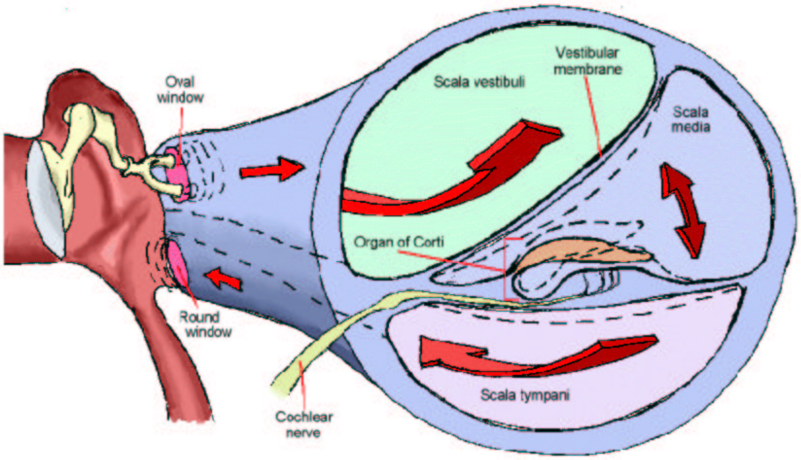


Figure 3.4: Cross section of the cochlea. The arrows show direction of fluid displacement. By Beth A. Hartwell, M.D., from website medic.med.uth.tmc.edu/Lecture/Main/ear.htm#inner.

delays and phase differences showing that the relatively slow motion of the travelling wave could not stimulate hair cells. Braun [7] suggested that the basilar membrane plays no role at low sound intensity levels, and at higher intensities its travelling wave motion dampens the displacements of the endolymph fluid and thus protects the rather fragile stereocilia from damage. He also concluded that at low intensities the outer hair-cells' motility could amplify the resonances, and the induced motion of the endolymph would excite the inner hair-cells, which in turn would pass it on to the auditory nerve.

3.3 The Auditory Nerve

There is a relatively large potential difference between the endolymph, that bathes the stereocilia, and inside the body of the stereocilia. When the stereocilia are deformed ionic channels open up in the membrane on its surface that cause a quick potential change in the receptor cell. This eventually leads to stimulation of the auditory nerve fibers that connect to the hair cell. The resulting neural discharges or *spikes* are transmitted to the brain (cochlear nucleus) through the auditory nerve fibers. There are approximately 30000 of these innervating each inner ear in humans. 90% of these end on inner hair cells. The remaining 10% of the fibers attach to outer hair cells. About 20 fibers innervate each inner hair cell, and 6 each outer hair cell. Each fiber that connects to an inner hair cell connects to one and only one, whereas those to outer hair cells typically branch and attach to about 10 outer hair cells. All of the fibers to inner hair cells are *afferent* fibers, that is they trasmit signals to the brain, while some 5% of the fibers to the outer hair cells are *efferent*, bringing signals from the brain. The latter, feedback-carrying fibers are of a type that is much smaller than the afferent cells. As a result they are much harder to observe and little is known about them.

When there is no stimulus individual auditory nerve fibers will fire spikes at their *spontanous firing rate*. This rate varies from a few per second to more than 100 per second. When the hair cell that the nerve fiber innervates is stimulated, the firing rate increases. The rise in the discharge rate depends on the intensity and frequency of the stimulus, but cannot exceed about 1400 spikes per second. The latter limit is related to the absolute *refractory period*, a period of about 700 microseconds of inactivity after every discharge that allows the haircell and the auditory nerve to recover the physical conditions needed for another discharge.

Auditory nerve fibers exhibit frequency selectivity closely related

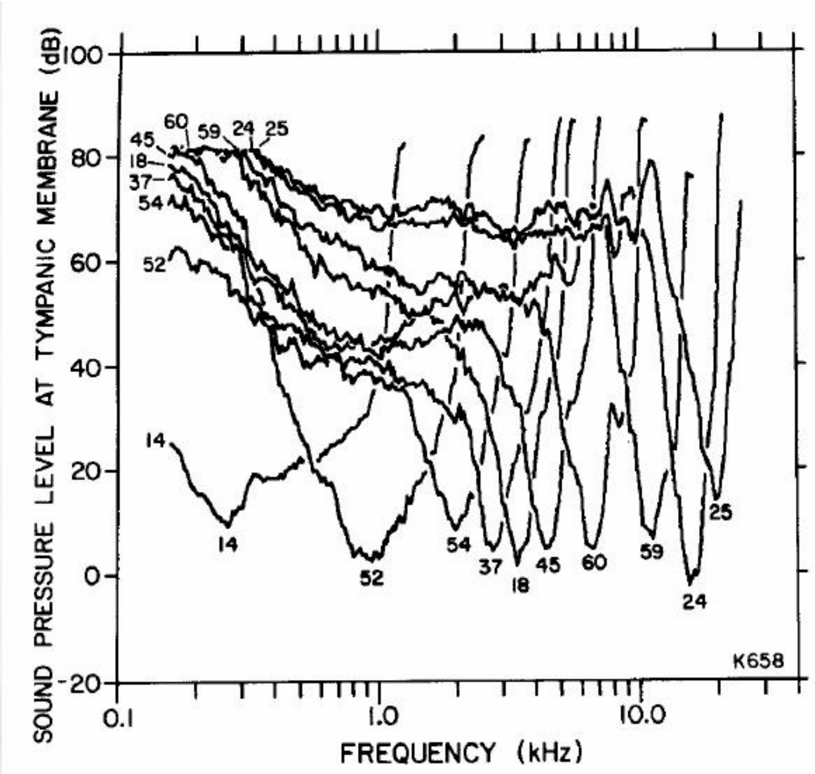


Figure 3.5: Auditory nerve tuning curves; from Kiang, 1984 [30]

to the place they innervate in the organ of Corti. The frequency that a fiber is most sensitive to (i.e. the one it responds to with firing rate increase above some threshold at the lowest stimulus intensity) is called the fiber's *best* or *characteristic frequency.* A fiber's frequency selectivity can be characterized by its *frequency tuning curve* that shows for each frequency the intensity required for a fixed increase in firing rate (see figure 3.5). Tuning curves greatly vary in shapes, but the vast majority of them have a sharp dip at the fiber's characteristic frequency with a steep rise at the high frequency side and a much gentler rising slope on the low frequency side. The typical tuning curve becomes more symmetrical and the dip less finely tuned as we approach lower frequency areas (towards the apex) in the organ of Corti [60]. Thus the auditory nerve fibers can be considered low-pass or asymmetrical band-pass filters. The filter characteristics of the nerve fibers coarsely resemble those of the basilar membrane at the place of innervation; the difference is that the nerve fibres are often far more finely tuned to their characteristic frequency than the basilar membrane. This may be caused by the earlier discussed filter characterstics of the stereocilia.

Another way to demonstrate frequency selectivity of the auditory nerve fibers is called the *isointensity contours.* These graph the number of spikes per second discharged by the nerve fiber in response to different tone stimuli at a fixed sound intensity. Rather interestingly, not only do these isointensity curves shift vertically (in the spike rate response) as we increase the intensity, but they also get deformed and change their maximum response frequency. Thus just by looking at the response rate of a single nerve fiber it is impossible to tell whether the response is to a low intensity tone near the characteristic frequency or to a high intensity tone further away from the best frequency of the fiber. However, as long as the frequency that is exciting the fiber is under 5 kHz, it is possible to identify it from the timing of the responses.

3.4 Temporal Attributes of Processing in the Auditory Nerve

At frequencies below 4-5 kHz the responses of the auditory nerve are locked to a particular phase of the cycle. A particular nerve does not fire in each cycle, but when it does its firing always falls in one particular part of the period. This *phase-locking* can be demonstrated by a *period histogram.* In a period histogram we bin the firings of a single nerve fiber, but every time we reach a given phase of the cycle we reset the time. The resulting histogram is a half-wave rectified version of the stimulating tone. The phase locking is usually explained by the fact that the nerve fibers are excited when the stereocilia is deflected only in one direction. Another version of period histogram is called *post stimulus histogram*, where we repeat the same stimulus over and over and record the response up to a length of time.

While the temporal information encoded by phase locking is present at all sound levels, the *spatial frequency selectivity* (i.e. fibers at which resonance point have maximal average firing rate response) quickly vanishes as intensity rises. Sachs and Young [34] investigated the response of the auditory nerve to steady state vowels. They observed that at low intesities the average firing rate for fibers with different characteristic frequencies showed clear, easily distinguishable peaks at the formant frequencies of the vowels. As the stimulus intensity was raised to relatively moderate intensity (68 dB SPL) or louder, the separate peaks disappeared leaving no basis for spectral differentiation. They explained the phenomenon as the combined effect of three factors: 1. the flattening of the isointensity response curves (and the tuning curves) of the nerve fibers as the intensity of the stimulus grows; 2. the non-linear flattening of the rate-intensity function at higher intensity, so that the firing-rate intensity is no longer responsive to small variations in stimulus intensity; 3. two-tone suppression. The loss of spectral discrimination in the average firing rate at normal sound levels points to the importance of the temporal information encoded in the firing pattern, and that possibly both spatial and temporal mechanisms are used by the auditory system.

Due to the long, slowly rising tails of the tuning curves of the nerve fibers on the low frequency side of their characteristic frequency, there are a lot of nerve fibers that will respond to a pure tone in the 50 Hz - 5 kHz frequency region. In response to stimuli that has a few well-separated (but possibly changing) main frequency components (such as vowels), the nerve fibers form clusters, with each cluster firing in phase with its frequency component f_i, as figures 3.6, 3.7, and 3.8 show. By finding the elapsed time between successive peaks of the temporal envelope of the pooled discharge pattern of the fibers in each phase-locked cluster, the current frequencies of the frequency components can be easily recovered. This is known as the *volley principle.* Let's ignore the non-linear nature of the processing up to the auditory nerves and consider them as simple linear low-pass filters with sharp cut-off frequencies! Then it can be shown that the length of time needed for a nerve fiber to respond to the frequency of a tone burst is long if the stimulus frequency is close to the cut-off frequency, and it gets shorter and shorter as the stimulus frequency descends further away from the

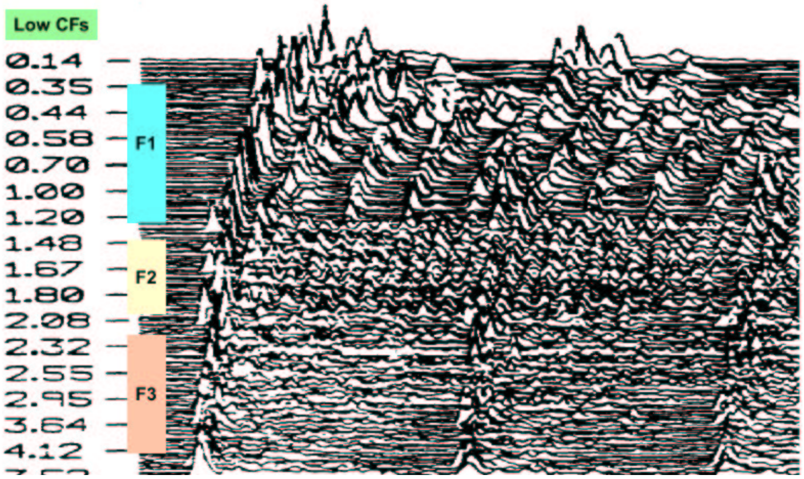


Figure 3.6: The responses (poststimulus time histograms) of a large group of auditory neurons evoked by two vowel pitch periods. Vertical axis represents different cochlear territories, horizontal axis stands for time. The areas F1, F2 and F3 span synchronized clusters. (Figure reproduced from MIT course HST.725 website, 2003.)

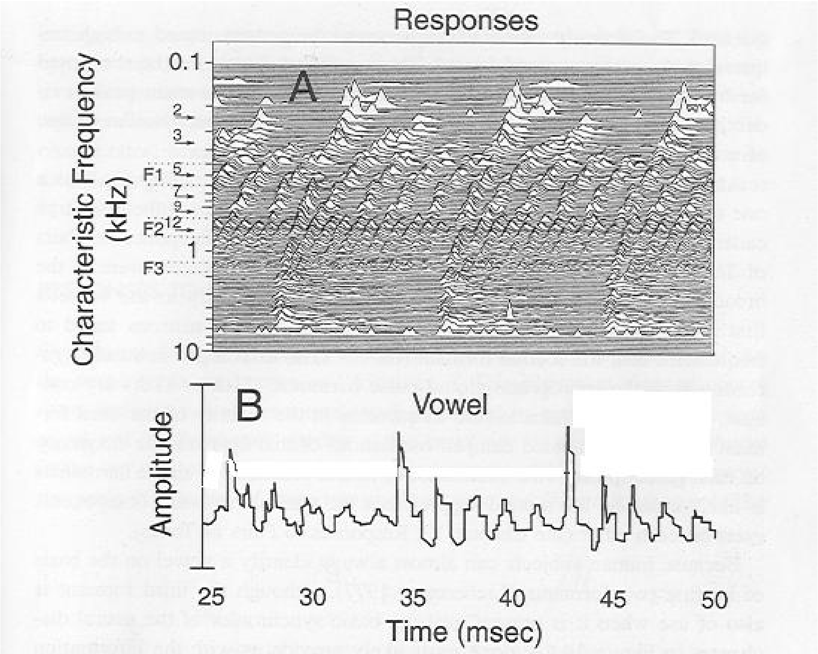


Figure 3.7: A: The responses (poststimulus time histograms) of a large group of auditory neurons evoked by three vowel pitch periods in syllable *da*. Vertical axis represents different cochlear territories with decreasing characteristic frequencies toward the top; horizontal axis stands for time. Various harmonics of the fundamental frequency are indicated on the ordinate, as are the frequency ranges of the vowel's three formants F1, F2 and F3. Note the synchronized clusters and the temporal structure within them (especially at the seventh harmonic)! B: Time-aligned acoustic waveform of the same vowel segment. (figure modified from [14, page 235], originally by Miller, Sachs and Shamma)

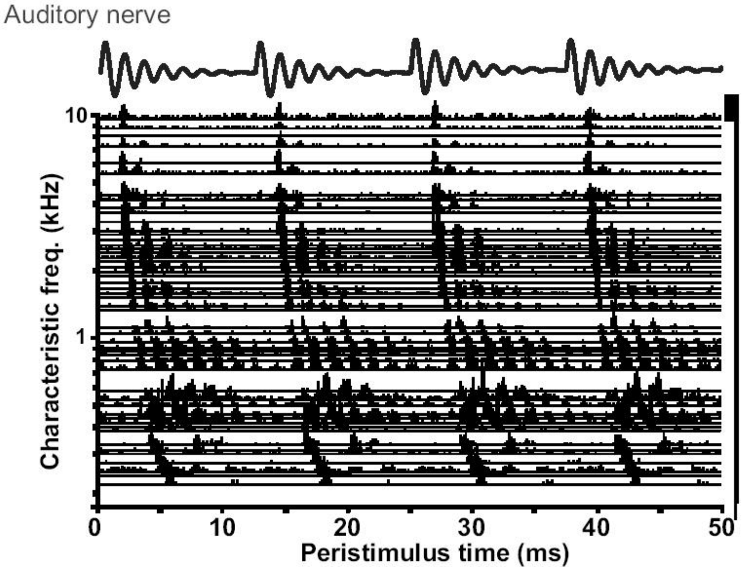


Figure 3.8: Auditory nerve post stimulus histogram in response to the stimulus shown by the waveform above the histogram. (Figure reproduced from MIT course HST.725 website, 2003.)

cutoff frequency (see Appendix A for proof).

We can combine the above ideas to get, for a special group of sounds, an interesting way around the frequency resolution versus time resolution tradeoff described by the uncertainity principle for linear systems. If the stimulus is a sum of a few, possibly quickly changing, well-seperated frequencies, the phase locked clusters of the auditory nerve can provide almost instantaneous fine resolution frequency information about the component frequencies. This is because the frequency components are well-separated: in each cluster there will be nerve fibers that are further away from the resonance point of stimulus frequency. These fibers will phase lock with the new, changed frequency component very rapidly. The period of their phase-locked temporal pattern will determine the new frequency one period after they have settled into their new discharge pattern.

It is reasonable to assume that the random processes that pick the cycles, in which nerves from the same frequency selective area respond, are identical and independent. In this case it does not matter wether we sample one nerve's firings over several cycles or many nerves from the same area during one cycle - with high probability they should yield the same distribution. This means that if the phase-locked cluster has many auditory nerve fibers, then by pooling the firing history of these fibers over one past period, the brain has almost immediate access to the *shape* (not just the timing!) of the same half-wave rectified version of the period hump of the frequency modulated signal that we see in the period histograms. This also means that there is no need for periodic repetition in the stimulus for it to be reconstructed. When the stimulus is a speech vowel-like signal with well-separated frequency components, the phase locked clusters are large. Also the signals exciting the stereocilia are band-limited frequency and amplitude modulated signals. Using the phase locked auditory nerve firings, the brain can reconstruct the half-wave rectified version of the hump shapes in the stimulus period-by-period! Thus it is conceivable that it can follow frequency and amplitude modulation (or equivalently, what we will later call temporal patterns) very closely, even at sub-pitch period time range. As there is some jitter in the exact firing time of each nerve fiber, phase locking becomes more and more smeared as the frequency of the stimulus increases, crossing over to uselessness above 5 kHz. We will only deal with vowels, which are perfectly recognizable from their frequency components under 3 kHz, so the latter limitation of the auditory system should not concern us.

As we hinted at it earlier, there have been two parallel theories of the auditory nerve representation of acoustic stimulus: 1. *Spatial* or *rate* *representation*: the auditory nerve fibers are tonotopically organized, and they can convey spectral contents of the stimulus by their average firing rate. 2. *Temporal representation*: the auditory nerve fibers are capable of locking, or synchronizing, to harmonics of the stimuli that correspond to formants of the speech signal. Many research papers suggest that the average firing rate is insufficient to represent speech information, and that the temporal information of the firing patterns should be included [34, 35, 3].

There is evidence that mammalian central auditory systems are sensitive to short amplitude and frequency modulated segments and spatio-temporal combinations [12, chapter 5]. It is also well established that envelope-based information with minimal frequency content retained in the signal is adequate for understanding speech. In fact Oxenham et al. [61] show that while our auditory system relies on the fine grained frequency content of the signal for sound localization and pitch perception, speech recognition relies more on information encoded in the envelope of band-pass channels of the signal. The latter provides a strong argument against using frequency as the sole primitive of speech processing. These findings, together with the temporal representation theory and the periodicity theory of pitch (see review in [57, page 66]), give some justification to our approach.

3.5 Time-Domain and Auditory Front Ends in Automatic Speech Recognition

There have been a number of attempts to process speech in the time domain. Baker examined statistical properties of zero crossings extracted directly from the waveform [4]. Seneff suggested a generalized synchrony detector (GSD) [49, 50] to identify formant peaks and periodicities of the speech signal. Hunt and Lefebvre showed noise-robustness of the GSD in recognition experiments using dynamic time warping recognizer [36]. Perceptual linear predictive analysis method of Junqua et. al. [21, 25, 28] is a perception-based technique in which the speech spectrum is transformed by several perceptually motivated relationships before performing linear prediction analysis. Itakura and Kajita in the subband-autocorrelation (SBCOR) analysis technique [45, 46] proposed to extract periodicities in speech signals by computing autocorrelation coefficients of sub-band signals at specific time-lags. SBCOR is shown to outperform the smoothed group delay spectrum for speech recognition tasks under noisy environments. The ensemble interval histograms of Ghitza [15] passes each element of an array of bandpass cochlear filters through an array of level-crossing detectors. Intervals between successive upward going level crossing points contribute to frequency information. Louder frequency component of the signal will cross more levels, and thus contribute more to its frequency. Kim created the zero crossings with peak amplitudes method [8] consisting of cochlear bandpass filters and nonlinear operations in which frequency information of the signal is obtained by zero-crossing intervals. Intensity information is incorporated by a peak detector and a compressive nonlinearity. This model also outperforms cepstral and other traditional front ends in noise.

Each of these techniques, one way or another, extracts frequency information and proceeds to perform statistical analysis. Thus, even though these methods start out in the time domain the output of their front end is still frequency, and thus they are subject to the same limitations as frequency-domain approaches.

Chapter 4

Patterns

We can see in figure 4.1 some sort of pattern repeating in the band-pass filtered channels of the signal during the pitch periods. Similarly, we can consider the pitch periods themselves as a repeating pattern that is slightly changing over time. In this chapter we will define patterns. Our goal is to create a concept that is better suited to describe pitch-period and other short resonances in speech than frequency.

4.1 Grouping Principles

From the observation above it is reasonable to postulate that pattern should be a recursive concept; that is, we should, at times, be able to consider a group of patterns as another pattern. Let's call the pattern that is atomic, that we cannot decompose into parts, a *primitive pattern* or *primitive.* We are facing 3 problems:

1. What are the primitives of the speech signal?
2. How are they combined?
3. How do we extract attributes from patterns that capture the vocal tract configurations across different utterances?

This is a very general set of questions that is highly relevant in almost any domain of perception. In essence we are asking how to form the representation or the vocabulary that is good for describing our domain. A good domain description would make it easy to separate members in the subsets of the domain that we want to differentiate (and only those). For example it could tell apples from pears, or in our case

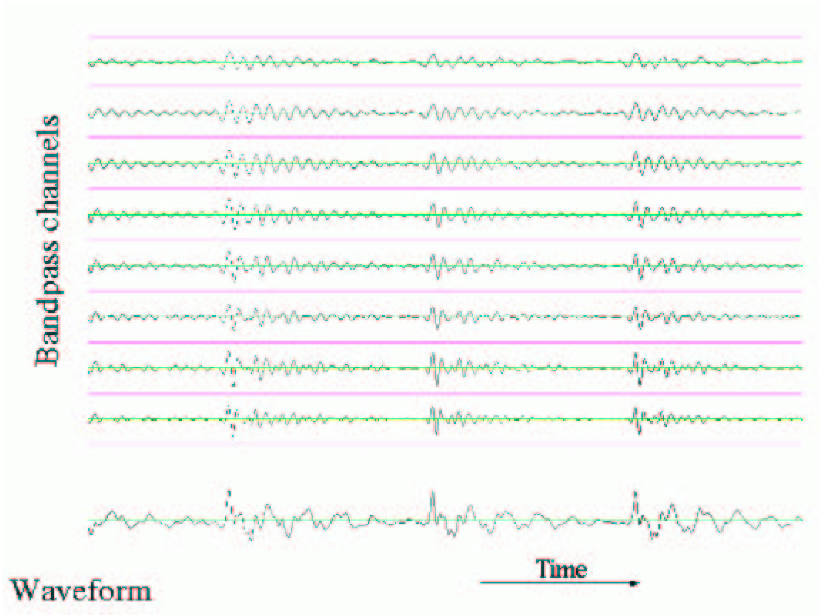


Figure 4.1: Pitch-period and sub–pitch-period patterns.

one articulatory configuration from another at mid-level, or syllables or words at higher levels of processing.

In speech processing the traditional and almost uniformly applied answer to question 1 is short frequency components that the discrete Fourier transform happens to respond to. Because of the reasons we spelled out in chapter 2, we consider the Fourier transform response too restrictive and brittle, and thus a suboptimal choice for a primitive. It would be fairly easy to find primitives and ways to combine the words of our representation that lead to a richer vocabulary that is better suited to our domain. The key problem to consider here is that we want a representation that is richer and has more descriptive power in our domain than the frequency based one, but at the same time restrictive enough to avoid an explosive growth of the possible combinations as we compile our subsequent layers of words.

4.1.1 The dot abstraction

We are going to demonstrate patterns via an abstraction that will hopefully help the reader visualize and better understand the pattern con-

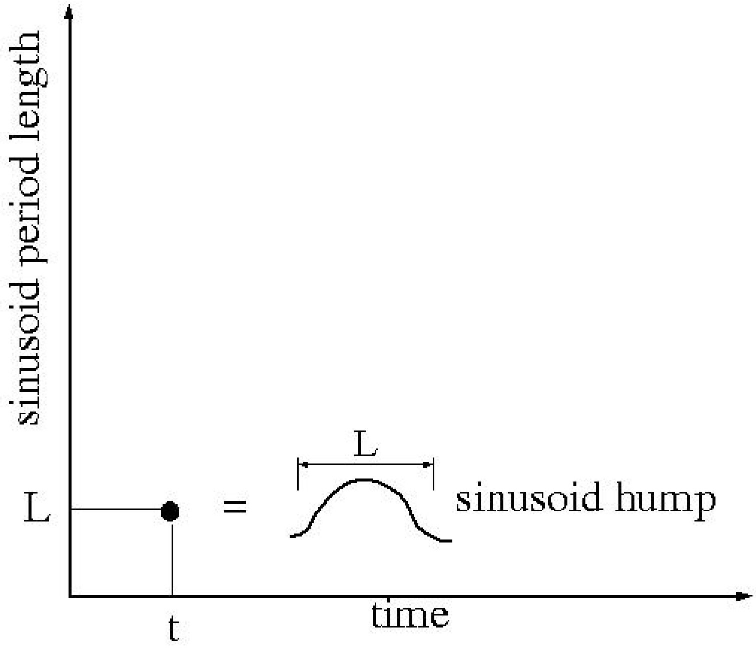


Figure 4.2: A sinusoid hump corresponds to a dot. The size of the dot indicates the amplitude of the hump.

cept. In this abstraction dots represent sinusoid humps in the time – sinusoid period-length space, as shown on figure 4.2. The abscissa represents time, the ordinate the period length of the sinusoid, and the size of the dot shows the amplitude of the sinusoid hump. Because this abstraction clearly does not deal with phase, we shall use it only for demonstration.

In this representation a sinusoid segment is a sequence of dots equally placed along a horizontal line (figue (4.3)). Higher frequencies would show up at a lower horizontal line and be placed more often.

If we try to represent what was going on in the smoothest channel (defined in the experimental section) of the $g \to ae$ transition, we get a configuration of dots depicted in figure 4.4. We shall soon define sequences like the one circled as a time pattern. For now we say that a time pattern is a group of varying sinusoid humps in a band-pass channel if they group around a maximal energy hump and they repeat

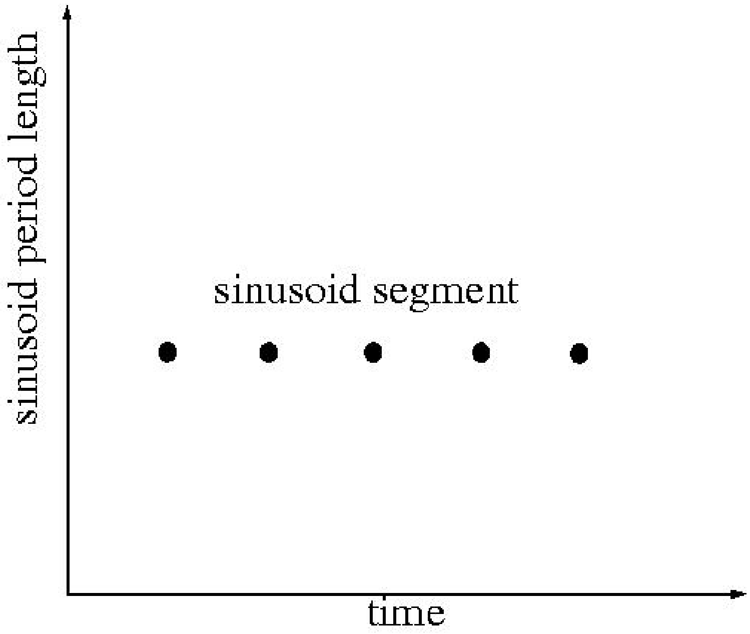


Figure 4.3: Dot represenation of a short sinusoid segment.

over time, as shown in figure 4.4. There are four things to remember about the elements of time patterns: 1. they are close in time 2. they are close in space, where the space dimension is defined as instantanous period (or equivalently, frequency) - here the period length of the sinusoid hump; 3. they group around a salient (maximum energy) element; 4. they, as a group, repeat over time.

If we look at multiple band-pass channels of the pitch period in the $g \rightarrow ae$ transition, the situation can be represented in our dot abstraction as shown in figure 4.5. There three time-patterns are grouped together to form a pattern. We will define a pattern as either a time pattern or a combination of patterns that are close by in time and space, group around a dominant pattern, and as a group repeat over time. Thus a pattern's components should

1. be close in time
2. be close in (frequency) space
3. group around a dominant (local maximum energy) component.

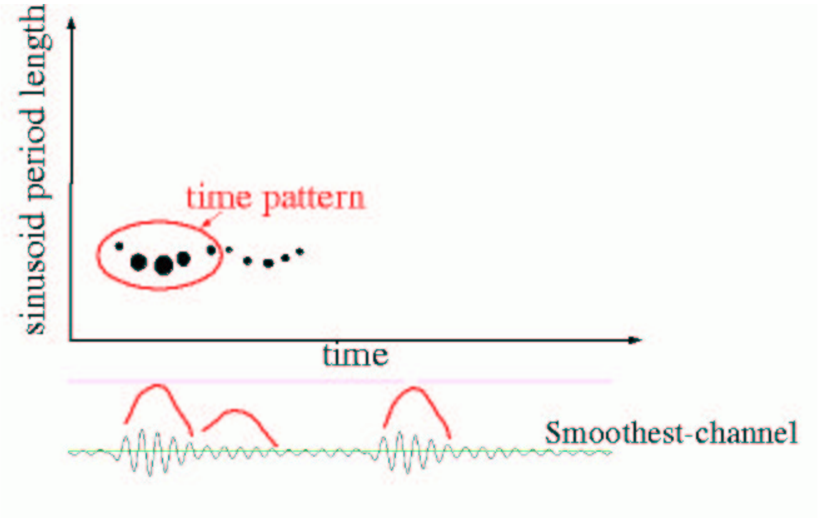


Figure 4.4: Illustration of the time pattern concept. The time axis in the dot abstraction and in the smoothest channel are aligned. (The smoothest channel will be defined in the experimental section.)

4. as a group, repeat

Why do we require all these constrainsts for a cluster of patterns to form a new pattern? We have started informally defining a language, whose primitives are time-patterns, and whose words are formed by recursively combining time patterns. If we just said that any local spatio-temporal combination as circled in figure 4.5 can form a pattern, than we would have an exponential explosion in the number of possible patterns as we combined newer and newer layers of words – practically every combination would qualify as a pattern. This situation is very similar to what happens when we have an overcomplete wavelet basis, which in practice often leads to computationally intractable NP-complete problems [38]. To avoid this problem we constrain the rules of combination by requiring that patterns always cluster around a dominant pattern, be close in time and space, and repeat. The least natural of these constraints, repetition, is ubiquitous in speech and turns out to be fairly restrictive in practice. Our constraints are rooted in the Gestalt theory of perception [31]. Before clarifying the pattern definition further, we need to define what we mean by energy.

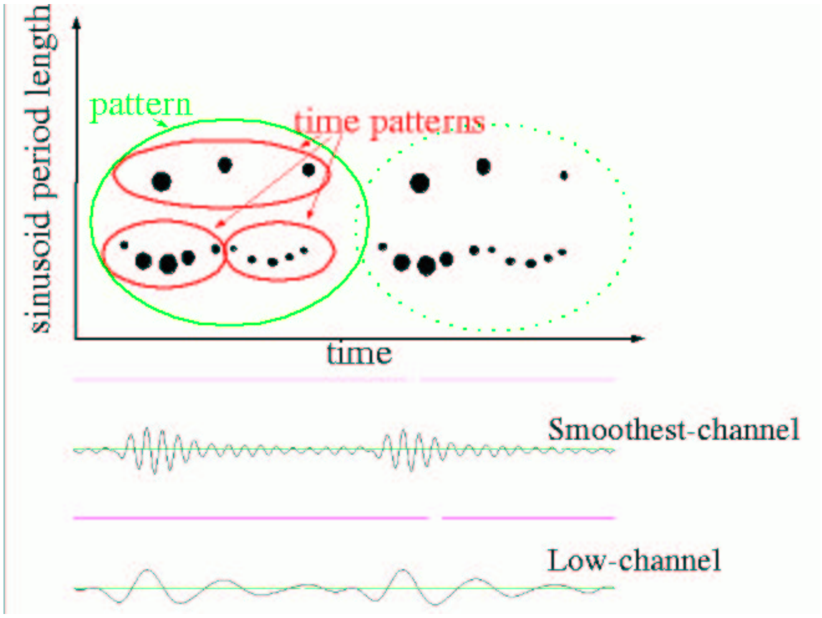


Figure 4.5: Illustration of the pattern concept. The time axis in the dot abstraction and in the smoothest and low channels are aligned. (The smoothest and low channels will be defined in the experimental section.)

4.2 Energy

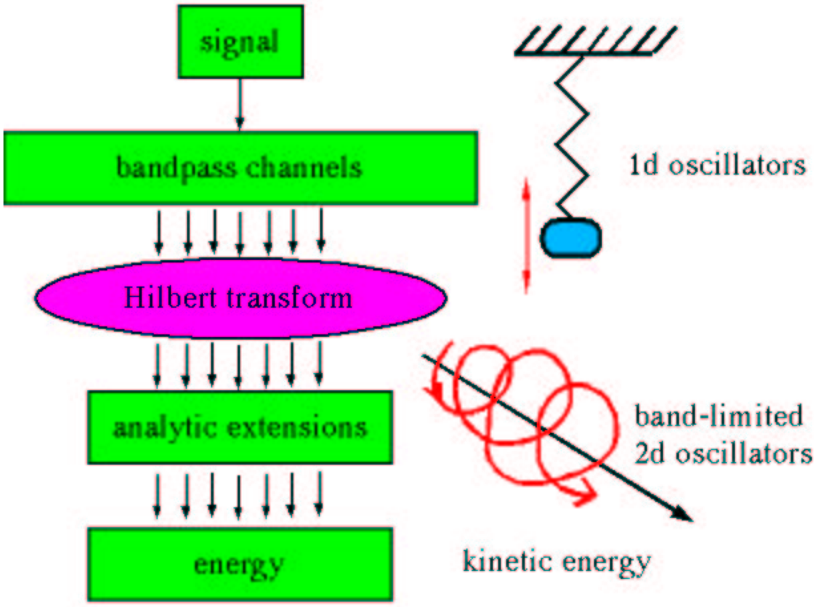
We can think of the sound signal as the sum of one-dimensional oscillators. This collection of various oscillators are analogous to a primitive cochlea model. Each oscillator corresponds to a section of the basilar membrane. The one-dimensional oscillations are induced by the movements of the endolymph (the liquid in the scala vestibuli), which are in turn generated by a sequence of vibration transmissions: from sound waves to the ear drum, the oscicles and the oval window. The combined effect of all of these is a linear mapping, and we simply take it to be the identity transformation. We further assume that there is hardly any liquid motion through the helicotrema. As the liquids are non-compressible, this means that the oscillations of the basilar membrane should add up to approximately the oscillations of the liquid in the scala vestibuli. Our last assumption is that the oscillators are band-limited; that is the frequency of each oscillation has an upper and a lower bound.

Equivalently, we can think of two-dimensional band-limited oscillators or *phasors* - unit weights that spin around a point in a plane, and whose one-dimensional projected motions add up to the signal. In a similar vein to how Teager defined energy [20], we would like a concept that measures the amount of energy stored in the oscillator at each moment. For example, the energy of a pure sinusoid should be constant as pure spinning motion (or harmonic oscillation in one-dimension) takes no external energy input to maintain. We define the energy of each oscillator as the kinetic energy of the unit-weight two-dimensional oscillator. Equivalently we could derive it as the sum of kinetic and potential energy of the one-dimensional oscillators.

What energies we get depends on the oscillators; that is, on the decomposition of the signal we choose. Which band-limited decomposition to choose is a rather hairy question, and we will not provide a computable answer here. When the system has no prior history of listening, our preferred decomposition is the one that minimizes the external energy input required to maintain the motions of the oscillators. When there is prior history – for example, a pattern that is repeating in the signal, or a sound pattern we have previously encountered many times before it is heard – the energy required to maintain the motion should decrease. We can imagine that it is as if prior history dug a hole under the pattern in a potential space that makes it easier for the system to recognize, or "slide into", that pattern. Since we do not know how to calculate either of these two cases in a reasonable amount of time, we shall abandon this idea. (We will return to the question of decomposition and feedback in the summary section.)

Instead we use a crude approximation: our one-dimensional oscillators will be defined by a fixed bandpass filterbank. We derive the two-dimensional oscillators by replacing the sinusoid components in the oscillation by complex exponentials. To do this we need to apply the Hilbert transform to each channel and add it to the original channel. This yields the *analytic part* or *analytic continuation* of the channels [1]. Then, channel by channel we can calculate the energy from the speed of the spinning point in the complex plane. Unfortunately the sum of the energies of these channels depends on the bandpass filterbank used – the total energy is not uniquely defined until we fix the oscillators. The analytic extensions of the channels are not necessarily orthogonal in the way complex exponentials are.

[1] The analytic part of a narrow-band band-pass channel is often called *phasor* in the literature.


signal
bandpass channels
Hilbert transform
analytic extensions
energy
1d oscillators
band-limited
2d oscillators
kinetic energy

4.3 Time-Patterns

Now we are ready to define our primitive: a *time-pattern* is the two-dimensional trajectory of an oscillator around an energy peak that repeats in time. Thus three things should hold for a time-pattern: (1) it should be a section of the motion of one of the oscillators; that is, it should be continuous in time and band-limited in frequency; (2) contain a local energy maximum; and (3) repeat over time. *Repetition* of a pattern means that either there is a similar trajectory soon after or before this one or that the *system* has encountered the pattern many times before. Finally two trajectories are *similar* if their normalized complex autocorrelation is high.

Note that this definition of time-patterns is phase-shift invariant; that is, a pattern may be repeating with phase shifts. Our similarity measure will yield the same result for phase-shifted patterns because the outcome of complex autocorrelation will rotate but its magnitude remains the same when the relative phases of the patterns change!

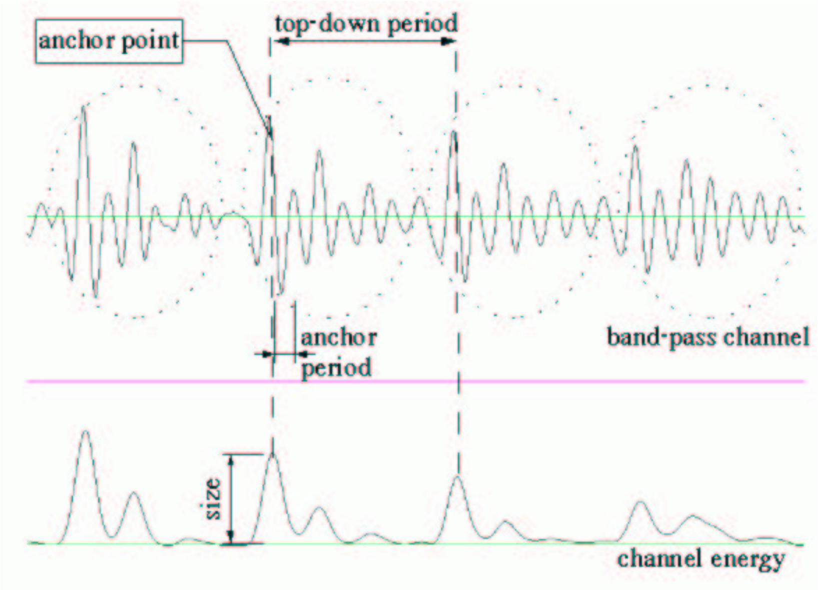


Figure 4.6: Illustration of time-pattern terms on a pitch-period time-pattern in a band-pass channel. The time patterns are circled with dotted line.

At times we will need to associate a point in time with when the time-pattern occurred. To this end we define the *anchor point* of a time-pattern as the maximal energy point across closest (in time) repetitions of the time-pattern. The anchor point is the point that maximizes the sum of the energies across the time aligned repeating time-patterns. Usually a good first approximation of the anchor point is the maximum energy point of the time-pattern itself. The *size* of a time-pattern is its energy at the anchor point. The *anchor frequency* of a time-pattern is the instantaneous frequency of spinning of its analytic part at the anchor point. We can find this by calculating the angular velocity of the time-pattern's two-dimensional trajectory. The *anchor period* is simply the inverse of the anchor frequency. The *top-down period* of a time-pattern is either zero if it has no neighboring patterns, or the time distance between its anchor point and the neighbor's anchor point to which it is most similar (see figure 4.3. Thus the top-down period is non-zero only if both time-patterns are neighbors in a sequence of locally repeating patterns. The top-down frequency of a time-pattern is the inverse of its top-down period.

4.4 Patterns

A *pattern* is either a time-pattern or a repeating combination of non-overlapping patterns around a maximal size pattern. Patterns are *non-overlapping* if the two one-dimensional convex hulls of their time-patterns in time are not intersecting; i.e. they occupy different time intervals. A pattern is *repeating* if either there is a similar pattern close by in time or the system has encountered similar patterns many times before. The *size* of a pattern is its energy maximum if the pattern is a time-pattern and the size of its maximal sub-pattern otherwise. If the pattern has a subpattern then its *anchor sub-pattern* is the sub-pattern that maximizes the sum of respective sub-pattern sizes across close (in time) repetitions of the pattern. The *anchor point* of a pattern is the anchor point of its anchor sub-pattern if it exists; otherwise, it is the anchor point of the time-pattern. The *anchor period* or *period* of a pattern is the top-down period of its anchor pattern if it has one; otherwise, (when the pattern is a time-pattern) it is equal to its anchor period. The *anchor frequency* is the inverse of the anchor period. The *top-down period* of a pattern is either zero if it has no neighboring (in time) patterns, or the time distance between its anchor point and the neighbor's anchor point (if there are two neighbors than the one to which it is more similar). A pattern is *similar* to another pattern if either both are time-patterns and they are similar or their sub-patterns are similar and the temporal and size distributions of their subpatterns are close.

Pattern is a recursively defined concept, and we would have to clarify three things to make the definition unambiguous; the exact meaning of similar time-pattern trajectories and sub-pattern similarity, and the meaning of encountering something "many times". Two time-patterns are *similar* if their normalized complex correlation is high (the exact usage will be explained in the experimental chapter). The second, sub-pattern similarity, we don not use in our algorithms so we leave it unspecified for now and return to the issue in the last chapter.

The inclusion of the system's history and recursion in our grouping principles are an interesting extension of both the Gestalt principles and memory-based parsing approaches (see for example [6]). It covers phenomena at all time-scales of our auditory experience starting at the sub-pitch period level all the way up to words of a language, or similarly from the short $(3 - 10\,ms)$ auditory percepts up to melody segments.

A tone burst is a time pattern, thus patterns can be considered a generalization of frequency. They are also a lot more general than wavelets [38] in that they are not tied to a fixed dyadic grid and they are phase-shift invariant. Additionally, patterns are recursive and can incorporate past experience, which give them a whole different level of expressive power. For example, they can explain how we can recognize the vowel in a syllable that has only one pitch period and would be ambiguous based on context; or how gradual perceptual shifts may occur during our adjustment to repeated stimuli [16, 43, 56]; or how we understand whisper.

Chapter 5

Observations

After laying out the rather general pattern framework we will probably disappoint the reader by using only a fragment of its descriptive power. We are going to present observations that link some vowel sub-pitch period patterns to certain articulatory configurations.

During each pitch period we choose band-pass channels in which we will look for time-patterns. The *smoothest channel* is the channel with the least variation in amplitude and frequency during the pitch period in a certain bandpass frequency range – intuitively the one that has the smoothest tone-like signal in it. The *low channel* is a bandpass channel close to the first maximal harmonic frequency. (We will describe how to find these channels in the experimental section in greater detail.)

5.1 Sub-Pitch-Period Time-Patterns Following Voiced Stops

The examples display the start of the vowel *ae* immediately after a voiced stop consonant: *g*, *d*, and *b* respectively. Figures 5.1 to 5.3 show the smoothest channel time aligned with the waveform below and the smoothest-channel energy above it. The hand-drawn humps in the smoothest channel indicate the locations of the vowel sub-pitch-period patterns.

Looking at a number of such examples, we have observed the following regularities following the voiced stop bursts in the pitch-periods of the transition to vowel *ae*:

If the stop consonant is a *g*:

1. the sub–pitch–period time-patterns in the smoothest channel have

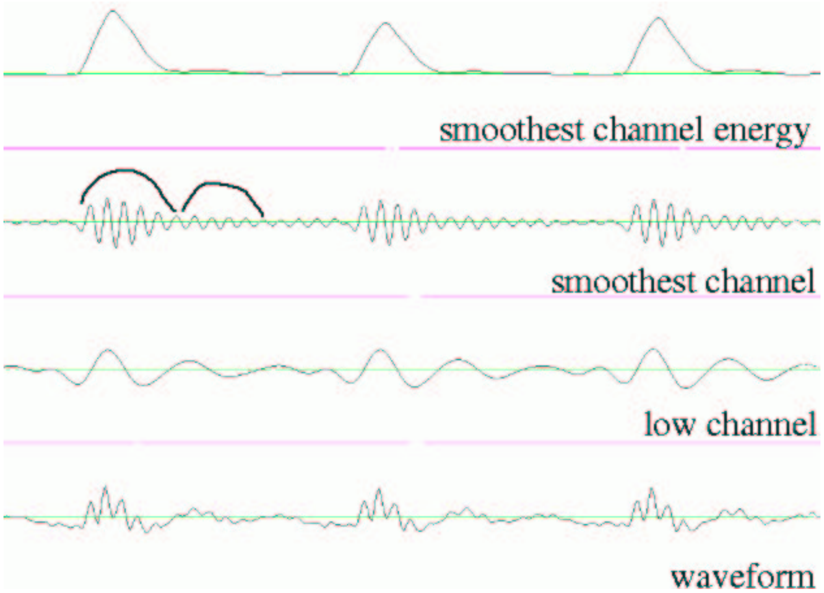


Figure 5.1: Repeating time-patterns during the first three pitch periods of the vowel in $g \rightarrow ae$ transition.

at least four complete periods (i.e. at least four humps) thus the ratio of the top-down pattern frequency and the pattern frequency is at least four.

2. the pattern period is at least as long as the the period of the low channel at the beginning of the pitch period; in other words, there are at most two patterns starting during the first two low-channel humps.

If the stop consonant is a *d*:

1. the sub-pitch-period time-patterns in the smoothest channel tends to have approximately three complete periods.
2. there are usually three patterns starting during the first two low-channel humps (i.e. from the start of the first low-channel hump to the end of the second hump), and their sizes are monotonically decreasing over time.

If the stop consonant is a *b*:

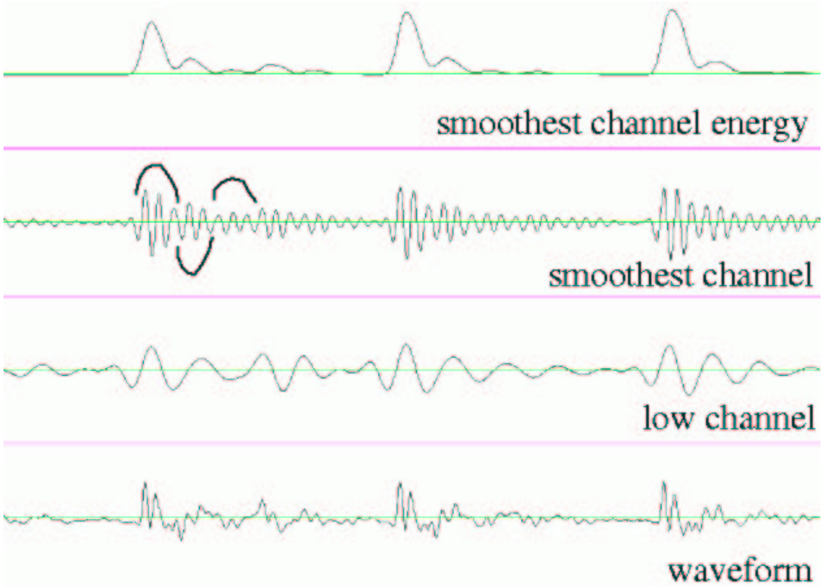


Figure 5.2: Repeating time-patterns during the first three pitch periods of the vowel in $d \rightarrow ae$ transition.

1. the sub-pitch-period time-patterns in the smoothest channel tends to have two complete periods.
2. the pattern period is about a third of the period of the low channel at the beginning of the pitch period. The sizes of the patterns decay as we get further away from the pitch period onset.

Note how the time-patterns are time-aligned with the energy peaks. In figure 5.2 and figure 5.3 depicting $d \rightarrow ae$ and $b \rightarrow ae$ transitions, we can see time-patterns that are repeating with half-period phase shifts. Perhaps this can be more clearly seen in the temporal fine structure for time aligned waveforms. Appendix B lists three further examples of sub-pitch-period time-patterns for each of the three configurations examined here.

5.1.1 Aeroacoustic flow in the vocal tract

The typically employed linear one-dimensional concatenated tube model of speech production [51] does not explain the presence of short quickly

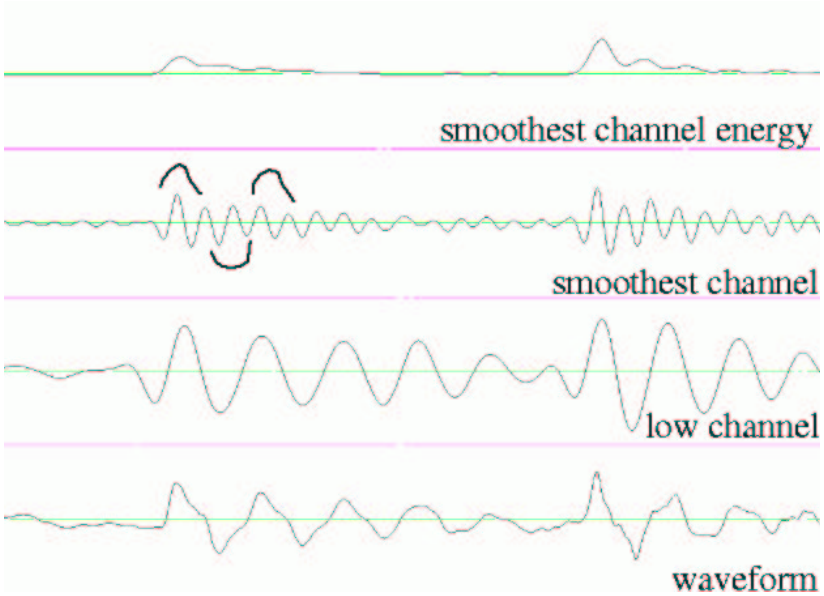


Figure 5.3: Repeating time-patterns during the first two pitch periods of the vowel in $b \to ae$ transition.

varying oscillations in either vowels or consonants. There is a growing amount of evidence that aeroacoustic fluid motion (figure 5.5) can significantly influence the sound field. Our understanding of this effect is limited; we are going to give only a glimpse of what has been done on the subject and a few references. Measurements done by the Teagers [20, 19, 53] show that during steady-state vowels there is a jet-flow in the vocal tract that travels much faster than the rest of the fluid. The slowly moving part of the fluid has a tendency to curl up in little swirls or *vortices* (see figure 5.4) - probably an effect of the viscous forces along the surface of the vocal tract. The Teagers studied several vowels and found that each of them had unique flow characterestics. Among the observed patterns were a jet flapping accross the walls of the oral cavity approximately at the pace of the first formant frequency, vortices stuck in various cavities of the vocal tract, and high-speed jets shedding vortices that keep moving slowly downstream towards the oral cavity. Kaiser [29] hypothesized that the temporal fine structure of speech can be explained by the interaction of the jet flow and the vortices. While vortices themselves are not efficient sound radiators, when they reach

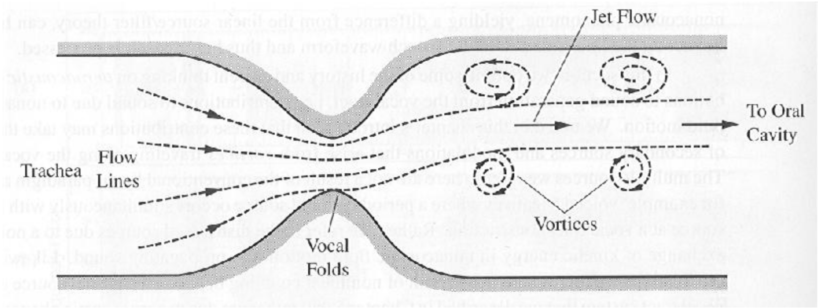


Figure 5.4: Vortices. Drawing from Quatieri [41]

changing solid boundaries they can be significant sources of sound energy. In addition to this, the flapping motion of the jet flow in certain vowels may act as excitation. Thus, depending on the vocal tract configuration, there can be multiple aeroacoustic sound sources that have different travel times and excite different cavities in the vocal tract. These phenomena may be responsible for the short oscillation patterns that we have observed. For a more complete review see [41, pages 562-572], and for a thorough treatment of vortices and flow-induced sound see Howe's books [23] and [22] respectively.

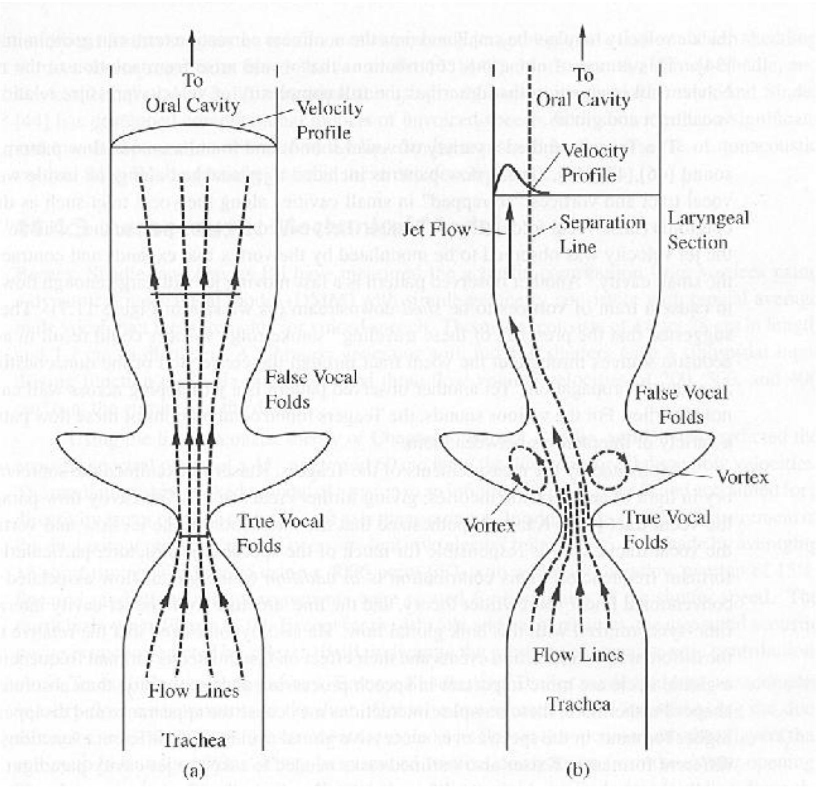


Figure 5.5: Comparison of (a) acoustic and (b) aeroacoustic models of speech production. Source: Kaiser [29]

Chapter 6

Experiments

In this chapter we describe the system we have built and the experiments we have run on it. First we review the algorithm, then the experimental setup, and finally the statistical results that support our earlier claims.

6.1 The algorithm

As shown in figure 6.1, the system starts with a linear pre-processing stage. This is followed by a non-linear time and channel-space segmentation algorithm that finds the vowel pitch period onsets, the vowel onset and certain special channels during each pitch period. The last two parts of the algorithm extracts three sub-pitch period pattern attributes and categorizes the samples as vowel release from alveolar closure (as in $d \rightarrow ae$), vowel release from velar closure (as in $g \rightarrow ae$), or vowel release from labial closure (as in $b \rightarrow ae$). Figure 6.2 shows a more detailed but still high level view of the four stages of the algorithm.

6.1.1 Linear Pre-Processing

We apply a simple linear preprocessing to the signal. First we upsample the 16000 samples/sec signal three-fold to 48000 samples/sec using linear interpolation. Then we run the signal through the filterbank whose structure is shown in figure 6.3. The high-pass filters (HPF) output is subtracted from the input, and the resulting difference is fed to the next high-pass filter. The cut-off frequency of the high-pass filters starts out at a high level and gradually decreases as we step through the filterbank. Due to the difference operation in the filterbank, we can consider

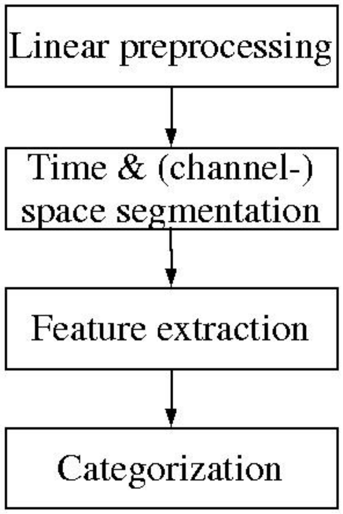


Figure 6.1: High level structure of the algorithm.

the input of each filter a low-pass filtered version of the original input. Consequently the output of the cascade of high-pass filters is a set of band-pass channels.

Figure 6.4 shows the structure of each high-pass filter with the frequency response magnitudes of the various stages of processing (for positive frequencies only). The input (1) is first convolved with the Gaussian e^{-x^2G} in the time domain yielding (2), then subtracted from the original input (3). (3) is first run through a simple averaging operation: we convolve it with a rectangular window of length $2G$ in the time domain. The resulting smoothed signal (4) is subtracted from (3) to yield the output (5).

The actual G parameter values used in the filters are shown in figure 6.5. We shall call the band-pass channels with $G = 5, 6, 6, 7, 8, 9, 10, 11$ *bands*. The lowest frequency ($G = 11$) channel of the bands will be referred to as *mid-channel*, the one with the highest frequency ($G = 5$) as *high-channel*. The channel with $G = 25$ will be called *low-channel*.

6.1.2 Time and channel-space segmentation

The second part of the algorithm finds the pitch period onsets and two special channels in each pitch period. Before we sink into the details,

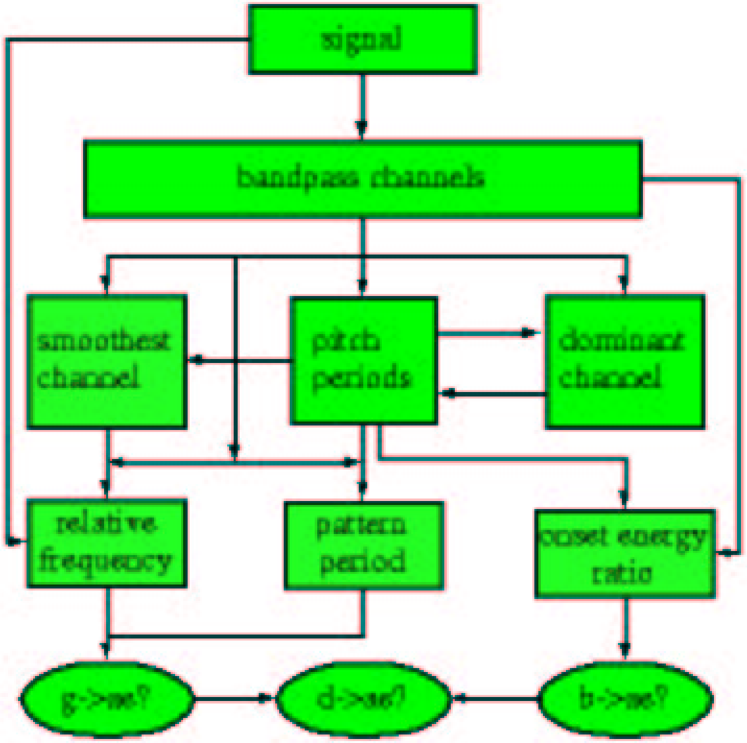


Figure 6.2: Flowchart of the main steps of the algorithm. Relative frequency, pattern period and onset energy ratio are the pattern attributes that the algorithm extracts from each pitch period.

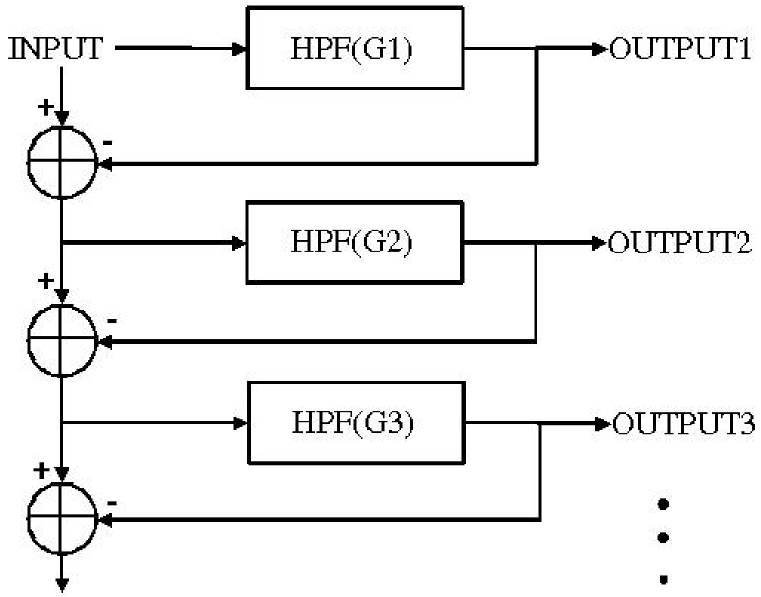


Figure 6.3: Structure of the filterbank.

let's take a closer look at the time segmentation problem: locating the pitch periods in vowels remains an unsolved, very complicated problem [39]. Figure 6.6 shows a few surprising counterexamples to what we may think vowel pitch periods should look like. Among the three shown waveforms we find examples of sudden pitch period length doubling and even tripling, pitch period onset time ambiguity, missed pitch period beats and very dominant high frequency energy peaks that are not synchronous with pitch period onsets.

Eventually we intend to use all this detail to differentiate between articulatory configurations found in vowels immediately after stop consonants. The articulators usually move fast during such releases, so when trying to say something about their position based on the vowel pitch periods, it's imperative that we find the first few vowel pitch periods right after the release of the closure while the tongue and the jaws are still close to their closure position. Incidentally, these first vowel pitch periods tend to be the most deformed and irregular, and thus the hardest to find. As a consequence, we need a really robust pitch period finding technique, not just one that finds most of them.

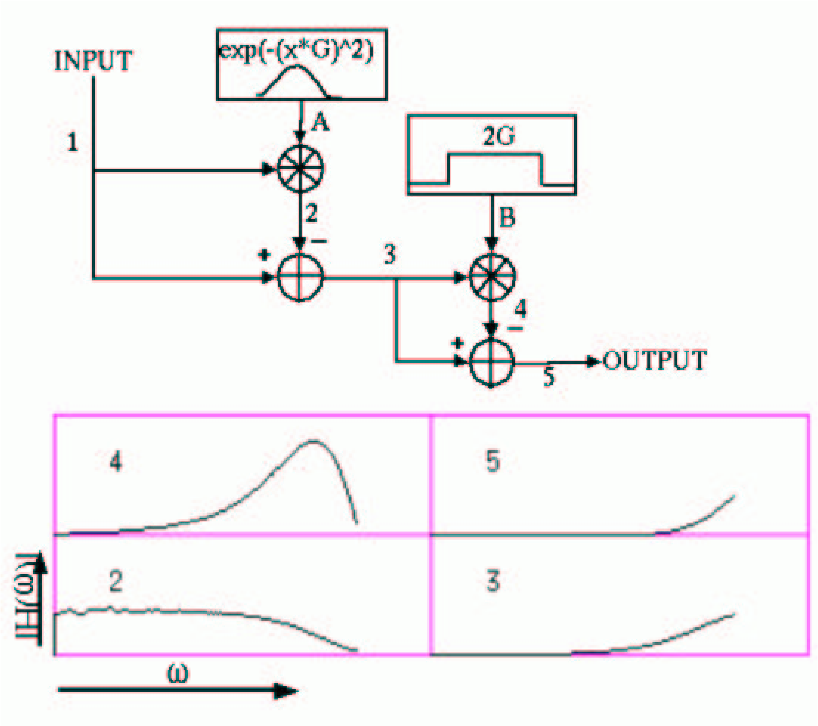


Figure 6.4: Top: structure of one high-pass filter. Bottom: frequency response magnitudes for positive frequencies.

The algorithm first calculates the analytic continuation of each band-pass channel using the following steps:

1. at each point in time find the Fourier transform using Hanning window of 2.5 ms duration [1] (120 units in our lengthened signal)
2. replace the negative frequency response values with zero and double the positive frequency response values
3. calculate the inverse Fourier transform of the output of 2.

The intuition here is that we calculate the analytic part for each sinusoid component separately: decompose the signal into its sinusoid components, then replace every sinusoid with the identical frequency complex exponential, and finally add the exponentials to get the ana-

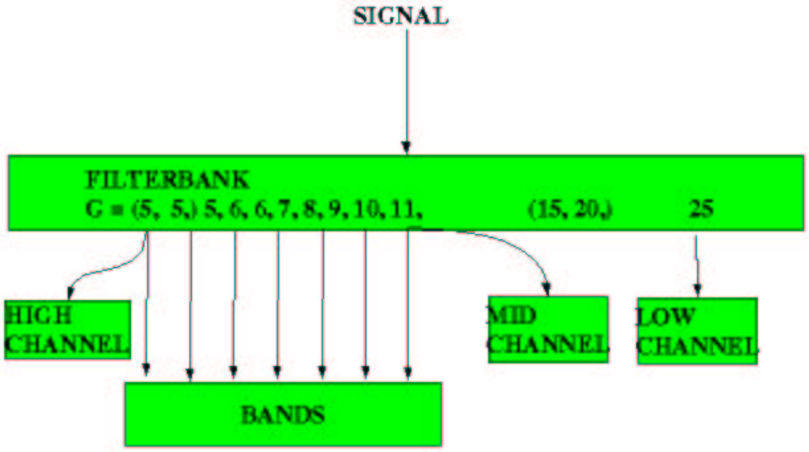


Figure 6.5: Parameter values used in the filter bank. The output of filters whose parameters are in brackets are not used in later processing.

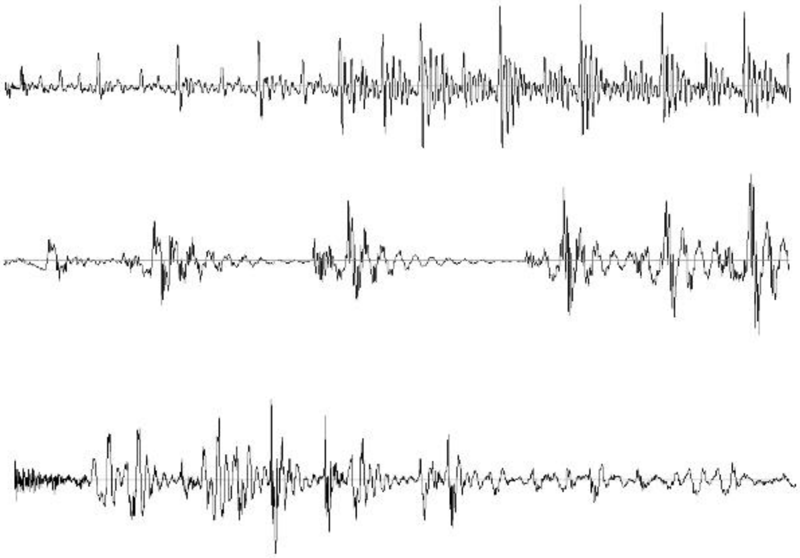


Figure 6.6: Three waveforms with irregularly changing pitch periods.

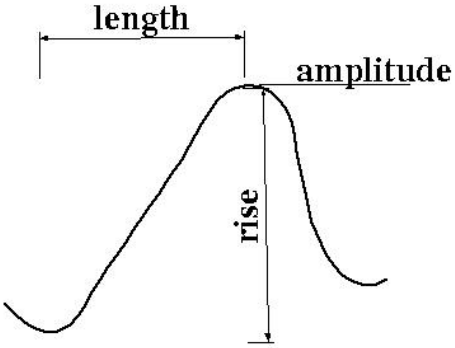


Figure 6.7: A hump.

lytic part of the original signal. The above steps indeed perform these steps: the transform

$Z_f(t) = 2F(\omega)U(\omega) = F(\omega)(1 + sgn\,\omega)$

is equivalent to our first two steps. If we take the inverse Fourier transform of its response to $f(t) = cos\omega_0 t$ equals $e^{j\omega_0 t}$, as $F(\omega) = \pi\delta(\omega - \omega_0) + \pi\delta(\omega + \omega_0)$, hence $Z_f(\omega) = 2\pi\delta(\omega - \omega_0)$. When we apply our third step, the inverse Fourier transform, we get $e^{j\omega_0 t}$.

Next, we use the analytic parts of the signals to find their energy and magnitude values. Let's denote the analytic part of signal s by $An(s)$; then the *energy* of s at time t is simply its squared velocity $|An(s)(t) - An(s)(t+1)|^2$, and the *magnitude* of $s(t)$ is the complex magnitude of $An(s)$ at t.

We discretize the various continuous signals by peak picking: find extrema, minimums and maximums (see figure 6.7). Each extremum has an *amplitude*, a *rise* equal to the difference of the present extremum's amplitude and the last extremum's amplitude, and a length denoting the elapsed time since the last extremum. A sequence of consecutive minimum-maximump-minimum forms a *hump*. The *humpamplitude* of a hump is the amplitude of its maximum, the *humpabsrise* value of a hump is the sum of the absolute values of the two rise values of the last two extrema of the hump.

Figure 6.8 shows the high-level structure of the pitch period finding algorithm. The process alternates between finding better and better candidates for pitch period starts and refining the dominant channel location for each pitch period segment. The shaded steps on the left are improving our pitch period temporal coordinate estimates, while the steps on the right stand for improvements in the spatial coordinate estimates. We start out using the mid-channel as our initial dominant-channel estimate.

Finding Pitch Periods I: Locating Dominant Low-Channel Peaks

In all of our pitch-period timing estimation steps our goal is to find the energy burst at the onset of the pitch period. First we calculate the integral of the mid-channel energy and low-channel energy during each low-channel hump. The sum of the latter two quantities during each low-channel hump (6.9) we will call *LowMidE*.

The difference of the present $LowMidE$ and the previous $LowMidE$ values is *LowMidEStep*. *LowMidE* tends to peak at the beginning of the pitch periods and smoothly decrease until the end of each pitch period, as the acoustic energy content of vowel pitch period gradually

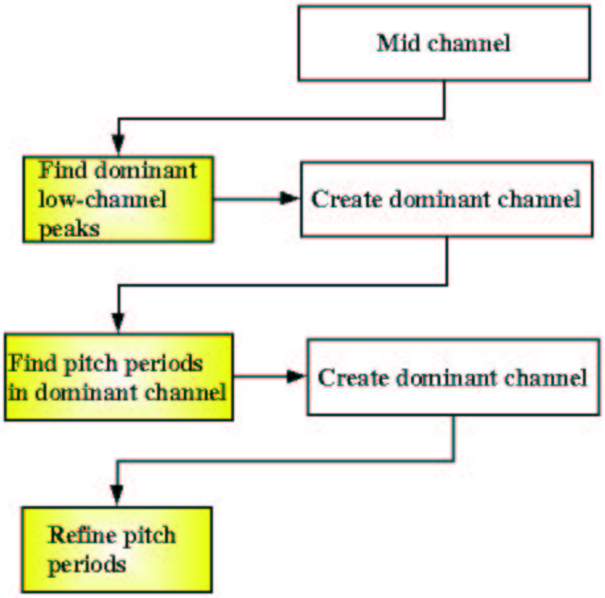


Figure 6.8: High level steps of finding pitch period onsets and the dominant-channel.

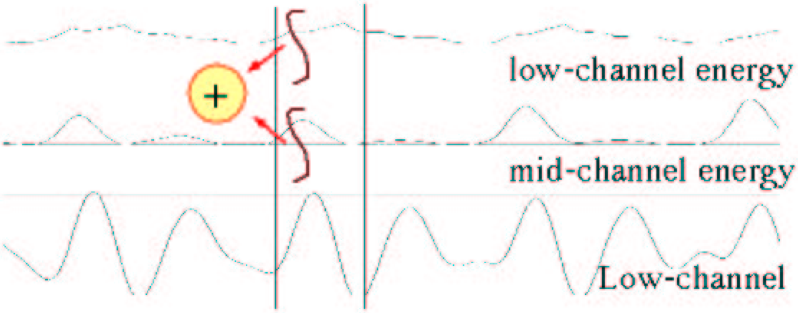


Figure 6.9: The sum of the integrals of low-channel and mid-channel energy between the vertical lines (from low-channel hump start to end) yields LowMidE.

falls off. As a result, *LowMidEStep* is negative for almost all its values that are not at the start of a pitch period, thus it tends to have a very strong local peak at the beginning of each pitch period. To get our first pitch period onset estimates, we form two running averages: the first, *AvgMax*, is the average of three neighboring local maximums of *LowMidEStep*; the second, *AvgNonMax*, is the average value of 6 neighboring non-maximum values of *LowMidEStep*, as shown in figure 6.10.

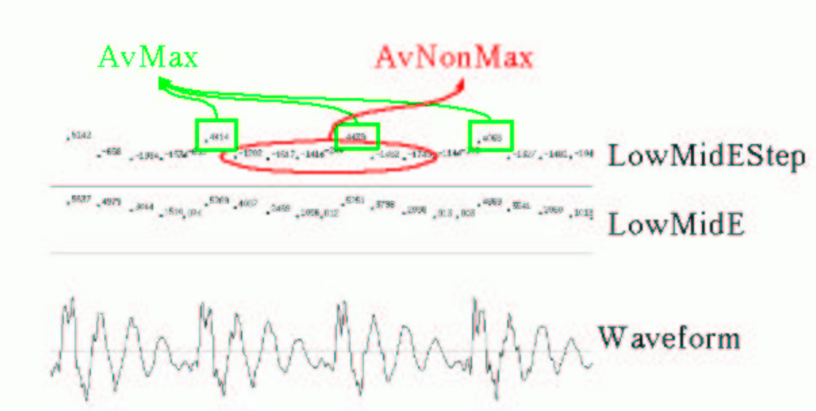


Figure 6.10: Averaging the three consecutive local maximum values in rectangles gives $AvMax$, and averaging the circled, non-maximum values yields $AvNonMax$.

We designate a low-channel hump as a dominant peak if three conditions hold: (1) it should be a local maximum of *LowMidEStep*, (2) $LowMidEStep > AvMax/5$ and (3) $LowMidEStep > AvNonMax + constant$. Next we refine the time segmentation by picking the mid-channel hump during the dominant low-channel humps that have the biggest increase in their humpabsrise value from the previous mid-channel hump (see figure 6.11).

The dominant peaks obtained this way serve as a crude first estimate of the pitch period starts.

Finding the Dominant Channel

The dominant peaks segment the signal into disjoint intervals. We proceed by picking the maximal channel out of the bands for each of these segments – this is the channel that has the maximum amplitude value

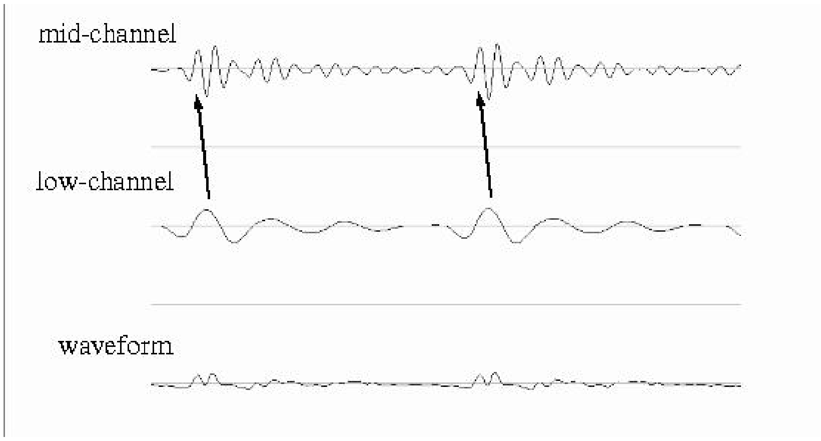


Figure 6.11: The arrows show how to refine pitch period onset timing from low-channel to mid-channel time resolution.

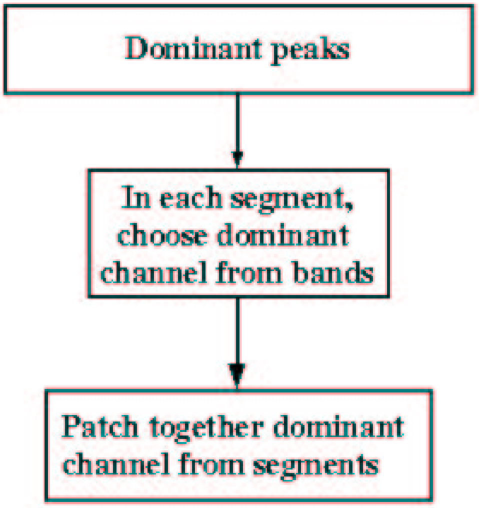

Dominant peaks
In each segment, choose dominant channel from bands
Patch together dominant channel from segments

Figure 6.12: Steps of creating dominant channel.

during the segment. Then we splice these channel segments together using a smooth linear transition between neighboring pitch periods.

Finding Pitch Periods II: Finding Pitch Periods in Dominant-Channel

In the next round of pitch period start time refinement, we attempt to capture the abrupt rise in dominant-channel energy and low-channel amplitude at the beginning of each pitch period (see figure 6.13).

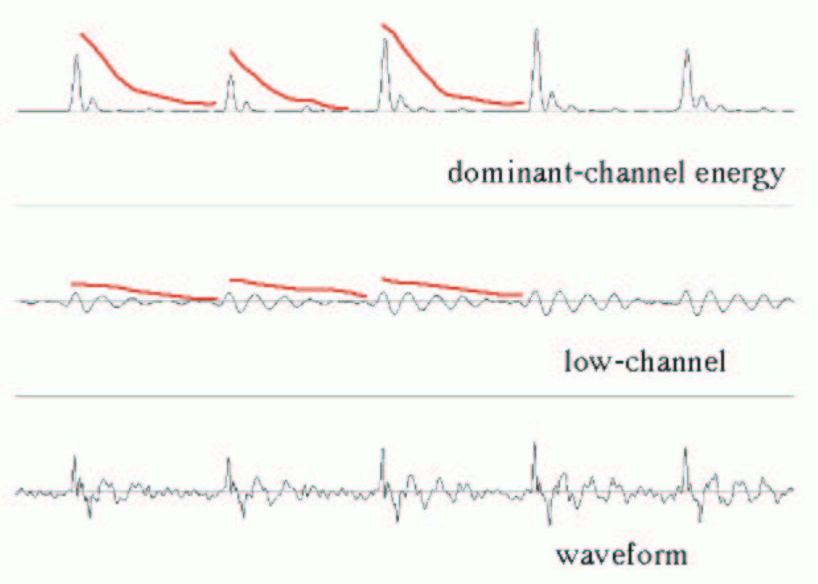


Figure 6.13: We intend to find the points where there is an abrupt rise in dominant-channel energy and low-channel amplitude.

To this end we define the *steprise operator* of a time series S: it is the leaky backward integral of S up to a point, divided by the leaky forward integral up to an earlier point in time as shown in figure 6.14. The steprise operator is parametrized by three values: the *forward halving distance* (the time distance by which the decay of the forward integral reaches 50%), the *backward halving distance*, and the *time-gap* between the start of the forward and backward leaky integrals. In our system the forward halving distance is 80 units, the backward halving distance is 50 units (one unit = $\frac{1}{3*16000}$ *seconds*), and the time gap equals the distance between the dominant low-channel hump and

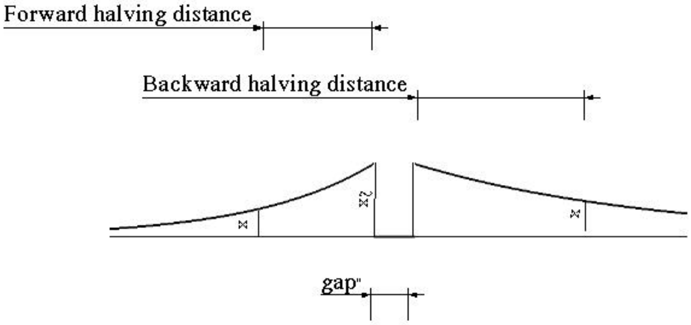


Figure 6.14: Illustration of the steprise operator parameters and decay.

the hump preceding it. The steprise operator of a time series – like the mid-channel energy that has an abrupt burst followed by gradual decay until the next burst – will be maximal at the burst points.

The flow-chart 6.15 shows the steps that determine whether a dominant peak can retain its pitch period candidacy. First we create the low-channel humpabsrise values, then apply the steprise operator to this time series, yielding the *low-channel hump absrise steprise* values. Second, we calculate the *dominant-channel energy hump amplitude* time series, and map it by the steprise operator to get the *dominant channel energy amplitude steprise* series. Third, we form the humpabsrise values of the dominant-channel energy and add them to the dominant channel hump amplitude steprise values. The resulting sum we shall call *EHAR*. For a dominant peak to become a pitch period candidate at this stage, two conditions should be met. First, either the dominant-channel energy humpamplitude steprise values should be larger than a threshold, or the low-channel hump absrise steprise values should exceed another threshold value. Second, the EHAR value of the dominant peak should be a local maximum in two ways: it should be a local maximum in its 1/480 *second* neighborhood (about 2/3 of the shortest pitch period length), and it should be a maximum over the period which is defined by the hills that the low-channel hump peak amplitudes form - as illustrated by figure 6.16. If both of these hold, then the dominant peak is selected as a pitch period start candidate.

Next, we go through the same pitch-period by pitch-period dominant channel selection and splice together steps previously demonstrated in constructing the dominant channel.

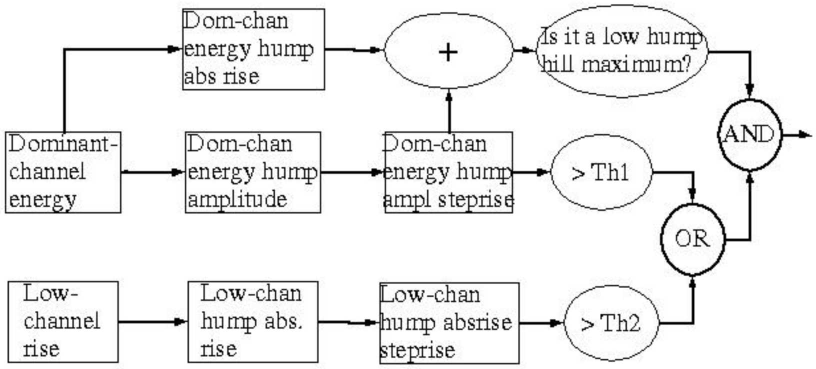

Dom-chan energy hump abs rise
+
Is it a low hump hill maximum?
Dominant-channel energy
Dom-chan energy hump amplitude
Dom-chan energy hump ampl steprise
> Th1
AND
OR
Low-channel rise
Low-chan hump abs. rise
Low-chan hump absrise steprise
> Th2

Figure 6.15: Flowchart of the second round of pitch period onset refinement.

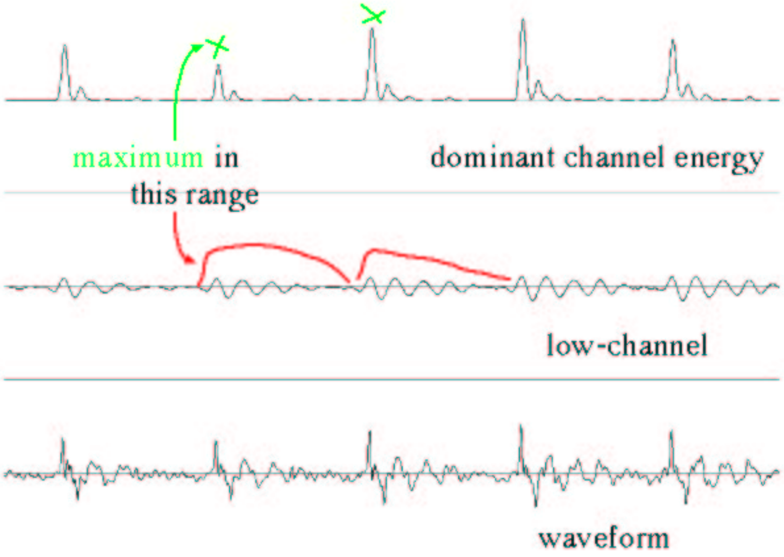


Figure 6.16: Humphill maximums.

Finding Pitch Periods III

We filter the pitch period candidates again,but use only values extracted at the onset of each pitch period candidate. First, we form the *average dominant channel pitch period onset energy amplitude* by averaging five consecutive pitch period onset values of the dominant-channel energy. Next, we find the *vowelcenter* by locating the maximum of the latter average (see figure 6.17). The vowelcenter is the location where the vowel is loudest. It also has the convenient quality that it is usually the moment having the least pitch period irregularity of the vowel. Starting from here and stepping backwards and forwards in time, we filter the pitch period onset candidates by three very loose smoothness constraints. The first constraint is that the next dominant-channel energy steprise value should be larger than 25% of the running average of the last five such values. The second constraint is that the low hump absrise value at the next pitch period onset should be at least 15% of the running average of the last five. The third constraint states that the next pitch period length should be within 2.5 and 0.4 times the running average of the last four pitch period lengths. The pitch period candidates that satisfy these constraints constitute the vowel pitch period onsets. Note that this method also finds the start of the vowel with high reliability.

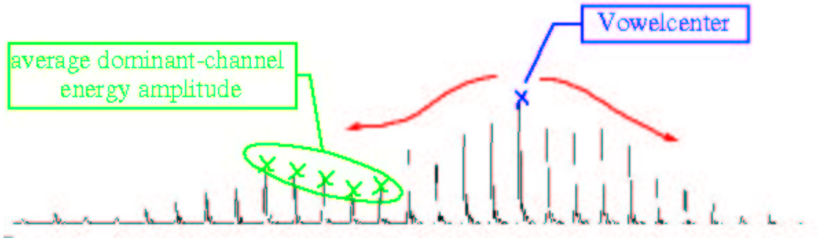


Figure 6.17: Average dominant-channel pitch period onset energy amplitude and the vowel center.

6.1.3 Feature extraction

We have used three sub-pitch period attributes to differentiate between articulatory configurations. Each utilizes some quality of the pitch period pattern that is characteristically different for the articulatory configurations that we are trying to differentiate. We shall describe first how to extract each of them from a pitch period, then describe who we combine them to make the actual decisions.

Onset Energy Ratio

Let us call the integral of the energy of a bandpass channel during the first low hump in a pitch period the *onset energy.* The *onset energy ratio* in a pitch period is simply the ratio of the sum of onset energies in the four lowest band channels and the sum of the onset energies in the two highest frequency band channels. This gives a measure of how dominant the mid frequency channels are relative to the high frequency channels. We will use this feature to differentiate the articulatory configuration at the $b \rightarrow ae$ transition from the $d \rightarrow ae$ and $g \rightarrow ae$ transitions. Looking at the energy distribution strictly at the vowel pitch period onset, the energy content is dominant, and thus has more discriminating power than the discrete Fourier-transform-based frequency band energy content measures because the latter blurs the the pitch period onset with the rest of the pitch period. This is bad because the energy content in each frequency band is falling over time, but usually at different rates, and thus the channels' relative energies may depend on the length of the pitch period.

Pattern Period Length

We will try to describe the intuition behind what we are trying to do and then discuss what we are actually doing to extract the feature.

We find the largest block of neighboring channels in the mid-frequency range whose humps are synchronized during the pitch period. Then find the smoothest channel, the one in this synchronized block whose analytic part has the least magnitude variation in this pitch period. Next, we try to find the pattern in the synchronized block of channels that is repeating over time during the pitch period. Finally, we find the top-down frequency of this pattern (or equivalently, the frequency of the pattern of the pitch period); that is, measure the time distance between the two largest neighboring patterns in the pitch period and express this distance by how many extrema it covers in the smoothest channel.

This is what we would like the algorithm to do; now consider what actual steps we carry out to extract the relative distance of the patterns! First of all, we are lazy and do not bother to find the synchronized block of channels. This is because among the band channels, there are at most two or three synchronized blocks, and the smoothest channel always seems to fall in channels 0-6 (where 0 is the highest frequency), and it also happens to be the smoothest (minimal) as measured by the magnitude variation of its complex part. Second, the pattern for the synchronized channels almost always seems to be identical with the pattern found in the second higher frequency neighbor of the smoothest channel. So this is what we do to find it: 1. Calculate the magnitude variation for channels 0-6 in bands up to 60 steps (1.25 ms) before the next pitch period onset; 2. Find the channel with the minimum value. Then we step down two channel indecies in bands to the *sub-smoothest channel*; that is, to the channel that is two indecies below the smoothest channel if the smoothest-channel index is at least 2, otherwise channel 0. Next, we try to find the patterns in the sub-smoothest channel.

To find the distance between repeating patterns in the sub-smoothest channel, we do two things: make an approximate hypothesis about where the first pattern is (i.e. where it starts and ends) and calculate the complex correlation of the analytic part of the sub-smooth channel between the segment defined by first pattern and its versions shifted to the subsequent extrema in the sub-smoothest channel. These complex correlation values will fluctuate; we find the maximal correlation value out of the set [second, third and the fourth peak] (the first peak, of course, corresponds to the pattern's correlation value with itself, hence we should leave it out). The shift that yields maximal peak value determines the d distance between the first two dominant patterns of the pitch period (which are invariably the dominant neighboring pair of the pitch period). Finally, we find the vowel onset in the sub-smoothest channel by finding the maximal HumpAbsRiseStep value in this channel during the first low channel hump of the pitch period, and simply calculate the number of extrema between the start of the pitch period onset in this channel and the point in time that is d distance from it.

We still need to explain how we try to pin down where the boundaries of the first pattern are. The situation is straightforward for the $b \rightarrow ae$ and $d \rightarrow ae$ transitions since the patterns in these highly correlate with the energy peaks; furthermore, the energy peaks are diminishing, at least through the first three peaks, as can be seen in figures 5.2 and 5.3. So for these configurations, we could just find the vowel onset in our channel and find the extremum sequence starting here that includes the maximum energy of the pitch period and the following energy minimum. The problem is that the energy maximums and minimums will not occur at the same time as the extrema of the sub-smoothest channel. To remedy this, we will use a smoothed version of the energy at each extremum value: the energy of the channel convolved with the e^{-x^2G} Gaussian where $G =$ dominant period of the channel during this pitch period. The pitch period *dominant period* of the channel is the reciprocal of the anchor frequency of the pitch period pattern in this channel. We use its simplified version which is obtained by calculating the dominant frequency and the dominant period from the angular velocity of the complex part of the channel at its energy peak point during the pitch period.

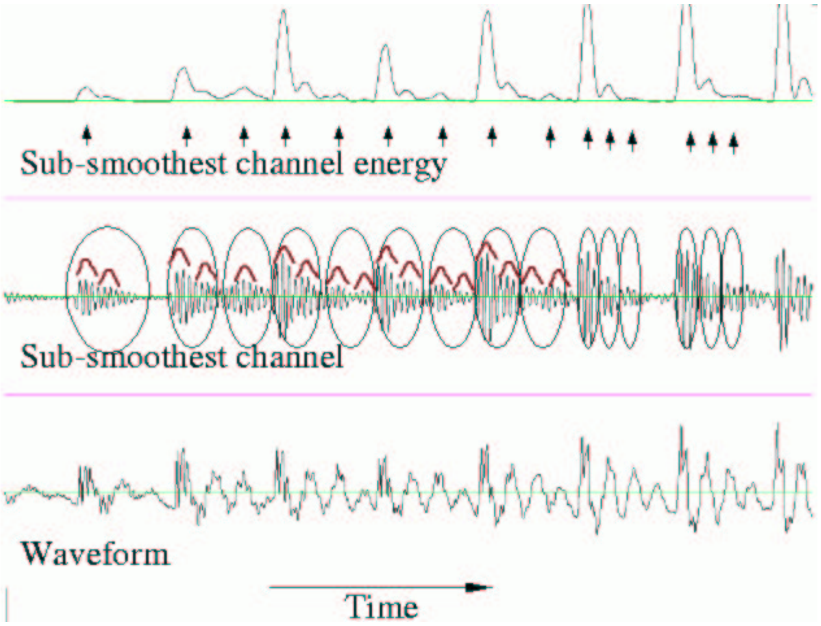


Figure 6.18: Time aligned sub-smoothest channel and its energy along with the waveform. The circled objects are sub-pitch period patterns, their anchors are indicated by the arrows pointing at the energy peaks above them. The little cups cover the subpatterns of the patterns.

We will, in fact, do almost exactly what we have described so far, with a slight modification to account for the much more complex situation at the $g \rightarrow ae$ transitions. According to our observations, figure 6.18 shows a typical sequence of pitch period pattern and energy landscape changes as the articulatory configuration goes from velar closure to an ideal *ae* utterance. The vertical arrows show the anchor points of the patterns that we would like to find. The patterns are circled by ellipses in the sub-smoothest channel. The little hat-like curves cover subpatterns within the circled patterns. The $g \rightarrow ae$ transition usually starts with a very long pattern that has a monotonically descending energy profile. This often takes place during the burst (and thus our system cannot see it). Next the long pattern breaks down into two subpatterns, and the length of the pattern gradually decreases while the prominence of its second sub pattern grows until we can no longer regard it a subpattern: it becomes the second of three identical, decaying energy patterns, which is the typical *ae* pitch period pattern. As can be observed in figure 6.18, the situation is complicated by the fact that often the pitch period is so short relative to the pattern that only part of the second pattern is observable. Other complications arise from the varying levels of subpattern prominence in different channels, which causes a lot of problems when our simple sub-smoothest channel picking method makes a mistake. Also, because the articulators' position at the onset of voicing (where we start inspecting the signal) depends on the speed of the articulators' movement and on exactly how long after the burst the voicing starts, we cannot really know for sure in which of these positions our pitch period sequence will start. Furthermore we cannot know how many pitch periods it will take for it to get to the *ae* position (and if it goes all the way there or not). These remarks should really make it clear for the reader that our primitive peak picking algorithm works only because we are using it solely for a binary decision. To actually follow the trajectory of the articulators, we would have to model how the patterns change from place to place with relatively fine resolution using a rich set of pattern templates.

Instead, we have exploited a characteristic difference between the typical *ae* pitch period pattern and the pitch period patterns closer to the velar release configuration: the first smoothed energy slump minimum after the peak (which we identify with the end of the first subpatterns) near the velar release tends to be very close to the energy peak of the subpattern. If the distance is less than that given by three sub-smoothest channel extrema, then we will take the second energy slump after the energy peak as the end of the first pattern. This method will make mistakes (one can see counterexamples even in figure 6.18), but since the final decision is made based on the complex autocorrelation values (which tend to give the correct answer even if we take autocorrelation of segments longer than the first pattern), they are usually not fatal. The second modification from the original peak finding algorithm is that if there is no second peak during the pitch period (i.e., the pattern is really long), we automatically return a preset, long distance between the first and the (not seen) second pattern that is characteristic of $g \rightarrow ae$ transitions.

There are still some details we skipped for clarity; now we will fill these in. The complex correlation we used is truncated if the shifted pattern gets too close to the next pitch period onset. The part that is closer than 30 steps is cut off. The autocorrelation value is normalized with the truncated length of the pattern segment. This step is important because if we allowed overlap with the very high energy first pattern of the next pitch period, it would very often interfere with the result.

Relative Frequency

Our next feature is much simpler to extract; it is the number of extrema from the onset of the vowel pitch period in the smoothest channel to the peak in the waveform during the second low channel hump. One can consider this a measure of relative frequency. It is more robust than frequency since it will not be affected by small local perturbations in the exact timing and shape of the sinusoid humps in the signal.

This and the previous feature (the relative peak period length and the relative anchor pattern period length) will be used to separate velar closure (i.e. *g*) to *ae* transitions from alveolar (*d*) and labial (*b*) to *ae* transitions.

6.1.4 Categorization

The algorithm is presented only with voiced *silence* → *stop burst* → *vowel* transitions from the signal, as they are segmented in the TIMIT database. At the beginning of the vowel of each of these releases the articulators tend to be in three distinct positions. After *b* they almost instantly assume the ideal configuration for the *ae* sound: jaws wide open and tongue in low frontal position. After *d* the tongue is high and frontal and the jaws are semi-open, and the tongue is rapidly moving down and the jaws are opening – everything is moving towards the *ae* position. After *g* the tongue is usually in the high back position moving forward and down, and the jaws are gradually opening as the configuration moves into the *ae* setup.

Because of the swift articulatory configuration change after the bursts in *d*-s and *g*-s, our algorithm needs to focus on the start of the vowel: it finds the first three pitch periods of the vowel that come after the vowel onset, and finds the above three feature values for each pitch period. It proceeds to calculate the average onset energy ratio, the maximum pattern period, the maximum of the relative frequency features over the first three vowel pitch periods. The latter are used in a simple binary decision tree that classifies the transition by making two decisions: 1. Is the input a labial release (b or not b)? 2. Is the input a velar release (g or not g)?

6.2 Experiments and Statistical Results

Data

We used the recordings in the TIMIT database [13] to run our tests. This database contains utterances from 630 speakers, covering 8 dialect groups of American English. The speakers are male and female adults with varied cultural backgrounds. They all say ten sentences; two of these are the same for everyone (the two *sa* or dialect sentences: “She had your dark suit in greasy wash water all year.”and “Don’t ask me to carry an oily rag like that.”), 5 and 3 of the other eight are randomly chosen from larger pools of 450 and 1890 sentences, respectively.

Our data set consisted of $(b - closure) \rightarrow b \rightarrow ae$; $(d - closure) \rightarrow d \rightarrow ae$ and $(g - closure) \rightarrow g \rightarrow ae$ sequences of TIMIT that do not contain word endings. Thus we excluded flaps (stop consonants that have no burst) or stop consonants that were in word ending position.

Results

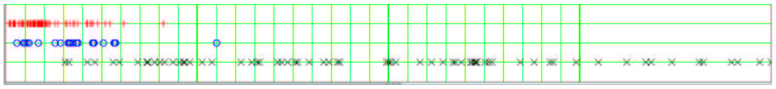


Figure 6.19: The abcissa shows the distribution of the *average onset energy ratio* feature values, the ordinate is 1 for $b \rightarrow ae$ (crosses: x), 2 for $d \rightarrow ae$ (circles: o), and 3 for $g \rightarrow ae$ (plus signs: +) transitions. The superimposed grid indicates integer values with (0,0) in the bottom left corner.

We have not used any statistical learning or other type of training algorithm. Instead we hand-picked the threshold values for the binary decision tree based on our observations. In particular we used the following decision thresholds for our features: if *average energy onset* > 6, then it is a $b \rightarrow ae$ release; else if *maximum pattern period* > 14 and $maximum\ relative\ frequency > 11$, then it is a $g \rightarrow ae$ release; otherwise a $d \rightarrow ae$ release. Table 6.1 shows the results for the three-way decision.

release type	recognized	false positive
$b \rightarrow ae$	93.5% (87 out of 93)	2.9% (3 out of 101)
$d \rightarrow ae$	81.8% (18 out of 22)	8.7% (15 out of 172)
$g \rightarrow ae$	86.1% (68 out of 79)	2.6% (3 out of 115)

Table 6.1: Recognition results for the three-way decision between voiced stop → *ae* transitions

release type	recognized
$b \rightarrow ae$	93.5% (87 out of 93)
d or $g \rightarrow ae$	97% (98 out of 101)

Table 6.2: Recognition results for binary decision between $b \rightarrow ae$ and d or $g \rightarrow ae$ transitions

We can also inspect the two binary decisions separately. Table 6.2 shows how the *average energy onset* feature fares in the binary decision between $b \rightarrow ae$ and d or $g \rightarrow ae$ transitions. Finally table 6.3 shows the accuracy of the binary decision between $g \rightarrow ae$ and d or $b \rightarrow ae$ transitions using features maximum pattern period and maximum relative frequency.

Analysis

Because we came up with the features and their threshold values while looking at the dataset ($g \rightarrow ae$, $d \rightarrow ae$ and $b \rightarrow ae$ transitions in TIMIT), it is unclear how much these results generalize to fresh, formerly unseen data. Also, the sizes of voiced stop → *ae* transition sets in TIMIT are 79 for g, 93 for b and 22 for b, which are too low to get overenthusiastic about our results. Nonetheless, the demonstrated clustering shows that there is some regularity worth further exploration. Here we will attempt to give an idea what goes wrong when the algorithms fail, and how the results compare to other current methods.

All three of the tested transitions have uncertainties that we have not modeled and thus these are a source of error for the system even if it works perfectly as planned in every respect. In a $b \rightarrow ae$ transition the tongue may not go all the way to the low position; if the speaker's

release type	recognized
$g \rightarrow ae$	87.3% (69 out of 79)
d or $b \rightarrow ae$	93.9% (108 out of 115)

Table 6.3: Recognition results for binary decision between $g \rightarrow ae$ and d or $b \rightarrow ae$ transitions

utterance is lazy, his tongue may stop somewhere halfway down in the *eh* position as described in the "lass" and "loss" example in the second chapter. In the $d \rightarrow ae$ and $g \rightarrow ae$ transitions the tongue and the jaws are already moving when the vowel starts. How much time passes between the release of the stop burst – when we know the articulatory configuration (at least the tongue and the jaw positions) with high certainty – and the start of the vowel varies and so does the speed at which the articulators move. Hence we have an inherent uncertainty about the articulatory configuration at the onset of the vowel. Even if our algorithm maps the sub-pitch period patterns to articulatory configuration attributes correctly, we may be making mistakes due to these phenomena.

We differentiate between three types of mistakes the system makes: 1. A *time-segmentation* mistake is when it gets either the vowel start or one of the pitch period start timings wrong; 2. A *space-segmentation* mistake is when it picks the wrong channels for inspection, and as a result does not find the pattern or compares the energies across wrong channels; 3. A *pattern segmentation* mistake is when it fails to find the main repeating pattern within the pitch period correctly. Out of the 21 misclassified samples (false positive plus false negative) in the three-way decision tests, 10 are due to time-segmentation mistakes - these are fairly easy to see by just looking at the individual samples. Pattern- and space-segmentation mistakes can be harder to spot for the naked eye, so we will not give an exact breakdown for the rest of the mistakes the system made. In our subjective opinion, about half of the remaining 11 mistakes are due to lack of proper modeling explained in the previous paragraph (i.e. in this case the expected patterns are simply not there), and the other half are pattern-segmentation and space-segmentation mistakes.

How do our results compare to prior experiments? The closest and best results we could find are in Suchato's recent PhD thesis [52] (see other work on stop consonants in the next section). He used knowledge about the human speech production system to determine place of articulation for stop consonants from the sound recording. In the experiment that is closest to ours [voiced stop consonants → frontal vowel transitions, testing data set is not identical with training data set (cross validation technique is used), only the data from the vowel following the stop burst is used] [52, table 4-5, page 111] he obtained 86.7%, 81.3% and 94.5% recognition accuracy for labial, alveolar and velar stops respectively. These are practically identical with our results after a permutation (93.5%, 81.8% and 86.1% for labial, alveolar and velar stops respectively). The comparison, however, is not completely fair because there are factors that make each experiment simpler than the other one in some respects. Suchato hand-picks the onsets of the vowels after the stop consonants and thus avoids time-segmentation type mistakes, which are responsible for half of our errors. He also uses human help in finding one of the formant frequencies knowing the transcription of the sample. An equivalent crutch for our system could be pointing to the patterns in the pitch periods (i.e. specifying pattern start times and channel coordinates). The latter would probably get rid of another 30% that is practically all the errors except the ones due to the lack of modeling the motion of the articulators during the burst. Suchato uses 5 features and a statistical learning algorithm; we use three features and simple hand picked thresholds. He uses a database that has only four speakers, which is likely to result in fairly uniform utterances. We use a subset of TIMIT database that has hundreds of speakers with diverse cultural backgrounds and dialects. The size of the subset is 177 speakers (21 different people say the 22 alveloar, 84 the 93 labial, and 72 the 79 velar utterances). On the other hand, our data set is much smaller, and one could argue that we use the training set as a test set, while Suchato employs cross validation testing method. Also he allowed all frontal vowels in his transitions, not just *ae*, as we did.

6.3 Prior work on stop consonants

Researchers have worked on stop consonant differentiation for decades. Delattre, Liberman, and Cooper [10] claimed that the place of articulation in stop consonants could be determined based on the second formant frequency transition but only showed such behaviour for *d*. Winitz, Scheib, and Reeds [58] looked at cues in bursts of the stop consonants to differentiate them. Zue [62] suggested the presence of context independent properties. He investigated a number of temporal characteristics of stops, voicing onset time (duration of frication and aspiration); and spectral characteristics such as frequency distribution in the burst spectrum. Blumstein and Stevens [5] showed that place of articulation could be determined based on a static snapshot of the frequency distribution taken shortly after the stop release. 80% recognition rate was achieved using short-time spectrum in the interval 10-20 ms after the release. Searle, Jacobson, and Rayment [47] obtained 77% classification accuracy by using features extracted from wide-band spectrum. Kewley-Port suggested that in some cases snapshots of the spectrum was sufficient, but in others it did not contain enough information. She used time varying attributes extracted from the beginning 20-40 ms interval of the stop-consonant – vowel transition, such as spectral tilt, existence of a sustained mid-frequency peak, and a delayed F1 onset value. Most of the later work used cues based on earlier publications – see a detailed and very recent review in [52].

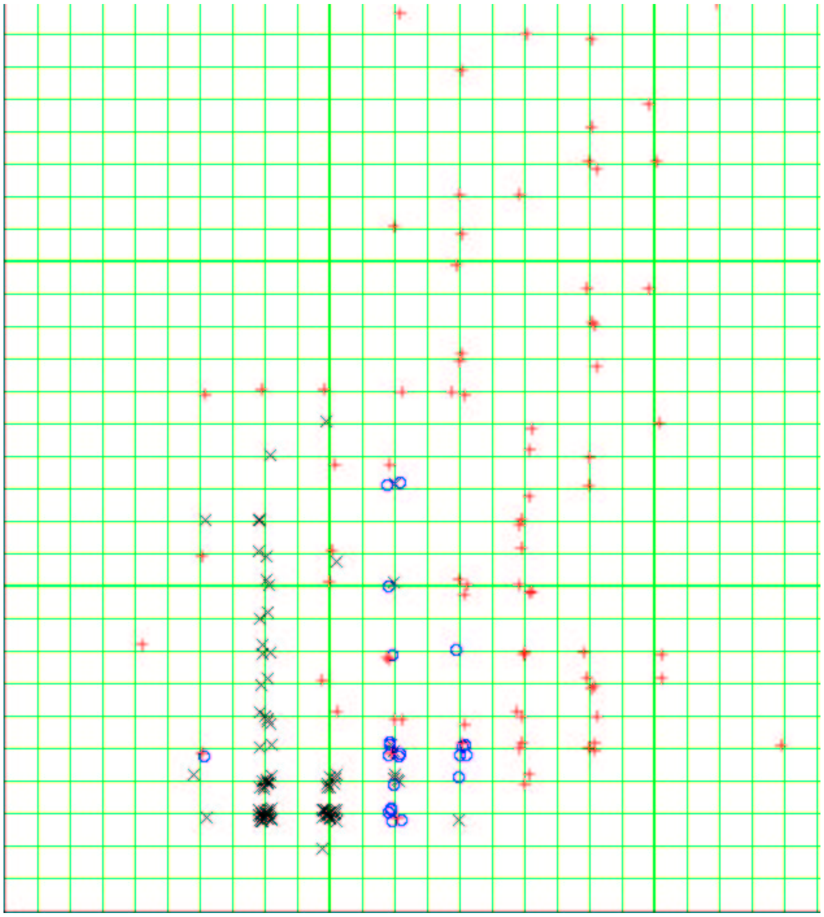


Figure 6.20: Scatter plot of features *maximum pattern period* (abcissa) and *maximum relative frequency* (ordinate). For better visibility of the individual samples, the figure shows slightly perturbed versions of the integer feature values. Crosses (x) correspond to $b \rightarrow ae$ transitions, circles (o) to $d \rightarrow ae$, and plus signs (+) to $g \rightarrow ae$ transitions. The superimposed grid indicates integer values with (0,0) in the bottom left corner.

Chapter 7

Summary

In this final chapter we will restate our views and elaborate on certain aspects of the pattern concept. We will also critique our approach on several accounts and thus give some idea about possible future directions of work.

We have argued that speech processing systems should rely on three layers of primitives: 1. patterns 2. articulatory configuration trajectories and 3. syllables. First, systems recover the articulatory configuration during vowels from sub-pitch period patterns. Then, they reconstruct the articulatory configuration trajectories using physical continuity constraints. Finally, they use the trajectories to find the most likely syllables that the speaker intended to utter.

Our pattern concept is a generalization of frequency. It is recursive and more flexible, and thus more expressive, than the frequency concept in several aspects. At the same time, patterns are constrained enough to avoid combinatorial explosion as we glue them together to form new patterns. By considering pitch periods as patterns, we treated vowels as a time division muliplexed signals with each frame containing a two dimensional image of the corresponding pitch period cropped at the end to fit the local pitch period length. We used the pattern concept to link the vowel pitch periods to articulatory configuration.

Now we would like to elaborate on one aspect of patterns that we have not defined yet. The similarity of subpattern time and energy distribution is a condition for pattern similarity. We have observed a number of times that the subpatterns that are in distant channels in pitch periods can show up with relative time shifts (the typical example is when the pitch period onset is at very different times in the high and low frequency channels). Thus we hypothesize that there should be flexibility in the relative timing and probably in the relative spatial locations of subpatterns. One may consider this quality the extension of the phase-shift invariance of temporal patterns. What is more, based on our observations, we have come to believe that this flexibility should extend to the similarity of subpatterns in the definition of pattern similarity. In other words, if there is one pattern with several subpatterns in the signal, then another pattern with similar spatio-temporal and size distribution of subpatterns, but deformed subpatterns, then these two should be recognized as similar. This means that, for example, in the simple case of a pure tone, it does not matter what the exact shape of the repeating humps are – as long as they are more or less following the beat of the original tone and resemble each other, they should be considered similar to the pure tone pattern. We, in fact, applied this principle in features *pattern period* and $relative\ frequency$ when we paid attention to only very rough indicators of shape, such as how many humps they covered in a certain channel.

We have still not defined pattern similarity (and thus, patterns) fully. What is missing is an exact definition of how we decide the similarity of the spatio-temporal and size distribution similarity of two patterns' subpattern structure. As we have used composite (i.e. not temporal) patterns only in the most rudimentary way and cannot justify any choice with experimental results, we will leave this measuring aspect ambiguous.

Although we have argued fiercely against using linear systems and brought nonlinear processing to a much earlier stage than is common, it is a valid observation that we are still using a linear pre-processing stage. We believe that this is wrong and has been a source of many of our problems. We do not know how to do this, but as we briefly explained in the chapter on patterns, we imagine that ultimately there should be nonlinear feedback even at the first stage of processing. The history (and perhaps a short segment of the future) of the system should affect how we break down the signal into channels. This decomposition should favor previously memorized and currently repeating patterns.

We have not come up with an algorithm that finds general temporal patterns or higher level patterns recursively. We have also failed to use the aspect of patterns that incorporate the history of the system. Nonetheless, we consider the recursive nature and the ability to form patterns based on non-local repetition crucial. These allow us to account for phenomena such as recognizing very short fragments of previously memorized patterns, recognizing non-locally repeating patterns like stop consonant bursts, or adjusting to a speaker's dialect as he is speaking.

We have pointed out several times how the repetitive nature of patterns and the continuity of articulatory configurations should help recognition in noise. In spite of this insight, we have failed to create a general pattern template that could be trained on the whole vowel space. If we had such a fine resolution $(vowel\ pitch\ period) \rightarrow (articulatory\ configuration)$ mapping template coverage, we could enforce continuity and test the system under various noisy conditions. Somewhat related is our lack of a model for non-vocalic sounds.

To demonstrate something about articulatory configuration, we would have been better off using the Wisconsin x-ray microbeam speech production database [59], which contains recorded speech and time-aligned data on some of the articulators' positions. Our decision to use voiced stop consonant releases from TIMIT is a quick fix and is a consequence of running out of funding.

Lastly, as reflected in our algorithm, we spent a lot of effort trying to find vowel pitch period onsets. It would be worthwhile to compare how well our algorithm works to other pitch period finding algorithms' performance.

Appendix A

Response time of low-pass filter with sharp cutoff frequency

We will prove informally that a linear time invariant (LTI) low-pass filter with sharp cutoff frequency takes long to respond if it is excited near its cutoff frequency in its pass-band, and that it responds quickly if it is excited further away in the pass-band.

First let's assume that there is an LTI low pass filter F that has a sharp cutoff frequency at ω_{c1} and it responds quickly to input frequencies ω, where $(\omega_{c1} - \omega)$ is small and positive. Then we can construct another such filter, $F2$, whose cutoff frequency, ω_{c2}, is slightly under ω_{c1}. $F2$ will not respond to inputs at frequencies between ω_{c1} and ω_{c2}. Thus by substracting $F2$'s output from $F1$'s, we can construct a bandpass filter whose bandwith is very small and that responds to its passband frequencies very quickly. This, however, contradicts the uncertainty principle, so our initial assumption must be wrong.

It remains to show that if the input frequency is further away from the cutoff frequency of the filter (but still in the pass-band), then the filter will respond quickly. If the low-pass LTI filter F has sharp cutoff frequency ω_c, then it can be decomposed into two components $F = F_{ideal}(\omega_c) + F_{bandpass}(\omega_c)$, where $F_{ideal}(\omega_c)$ is the rectangle-shaped ideal low-pass filter and $F_{bandpass}(\omega_c)$ is band-pass filter in a narrow frequency band around ω_c. When the input frequency is further away from the cutoff frequency, the sharp bandpass filter will not respond, so it is sufficient to consider the response of the ideal low-pass filter.

The latter's impulse response is $h(t) = \frac{\sin \omega_c t}{\pi t}$. Let the input be a causal sinusoid, $x(t) = sin(\omega_0 t)u(t)$, where $u(t)$ is the step-function. Then the output is [1]

$$\begin{aligned} y(t) &= (x * h)(t) \\ &= \int_{-\infty}^{\infty} \sin(\omega_0(t-\tau))u(t-\tau)\frac{\sin(\omega_c \tau)}{\pi\tau} d\tau \\ &= \int_{-\infty}^{t} \sin(\omega_0(t-\tau))\frac{\sin(\omega_c \tau)}{\pi\tau} d\tau \end{aligned}$$

Using the identity $\sin A \sin B = \frac{1}{2}(\cos(A-B) - \cos(A+B))$, we can rewrite the integrand as

$$\begin{aligned} & \sin(\omega_0(t-\tau))\sin(\omega_c\tau) \\ = & \tfrac{1}{2}\cos(\omega_0 t - \omega_0\tau - \omega_c\tau) - \cos(\omega_0 t - \omega_0\tau + \omega_c\tau) \\ = & \tfrac{1}{2}\cos(\omega_0 t - \omega_+\tau) - \cos(\omega_0 t + \omega_-\tau) \end{aligned}$$

where $\omega_+ = \omega_c + \omega_0$ and $\omega_- = \omega_c - \omega_0$. Expanding the cosine terms yields

$$\begin{aligned} \sin(\omega_0(t-\tau))\sin(\omega_c\tau) &= \\ \frac{1}{2}(\cos(\omega_0 t)\cos(\omega_+\tau) + \sin(\omega_0 t)\sin(\omega_+\tau) - \\ \cos(\omega_0 t)\cos(\omega_-\tau) + \sin(\omega_0 t)\sin(\omega_-\tau)) \end{aligned}$$

thus the output

$$\begin{aligned} y(t) &= \frac{1}{2}\sin(\omega_0 t)\int_{-\infty}^{t} \frac{\sin(\omega_+\tau) + \sin(\omega_-\tau)}{\pi\tau} d\tau + \\ & \frac{1}{2}\cos(\omega_0 t)\int_{-\infty}^{t} \frac{\cos(\omega_+\tau) - \cos(\omega_-\tau)}{\pi\tau} d\tau \end{aligned}$$

or

$$y(t) = \sin(\omega_0 t)S(t) + \cos(\omega_0 t)C(t) \tag{A.1}$$

The maximum amplitude of the output is limited by functions $|S(t)|$ and $|C(t)|$. We want to show that $|S(t)| + |C(t)|$ grows relatively fast

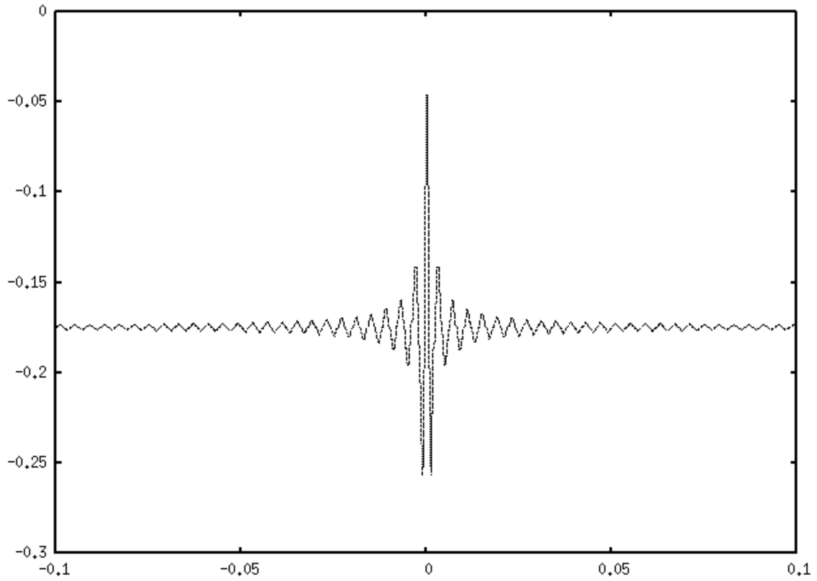


Figure A.1: Graph of $C(t)$ with $\omega_c = 500$ and $\omega_0 = 250$.

if ω_c is slightly above ω_0; that is, $\omega_- = \omega_c - \omega_0$ is a small positive number, and that it grows more slowly if ω_- is larger.

We will show first that $C(t)$ is small for $t > 0$.

$$\begin{aligned} C(t) &= \frac{1}{2}\int_{-\infty}^{t} \frac{\cos(\omega_+\tau) - \cos(\omega_-\tau)}{\pi\tau} d\tau \\ &= \frac{1}{2\pi}\int_{-\infty}^{t} \frac{\cos(\omega_+\tau) - 1}{\tau} - \frac{\cos(\omega_-\tau) - 1}{\tau} d\tau \end{aligned}$$

Using $\omega\tau = z$ substitution,

$$\begin{aligned} \int_{-\infty}^{t} \frac{\cos(\omega\tau) - 1}{\tau} d\tau &= \\ \int_{-\infty}^{0} \frac{\cos(\omega\tau) - 1}{\omega\tau}\omega d\tau + \int_{0}^{t} \frac{\cos(\omega\tau) - 1}{\omega\tau}\omega d\tau &= \\ \int_{-\infty}^{0} \frac{\cos(z) - 1}{z} dz + \int_{0}^{\omega t} \frac{\cos(z) - 1}{z} dz &= \end{aligned}$$

[1] The lion's share of the following derivation is from a handscript by Bertold K. Horn.

$$\gamma + C_+(\omega t)$$

where γ is a constant that is independent of ω.

- Case 1: if $1 > t >= 0$, then the Taylor series of cos converges and

$$\begin{aligned}
C_+(x) &= \int_0^x \frac{\cos(z) - 1}{z} dz \\
&= \int_0^x \frac{(1 - \frac{z^2}{2!} + \frac{z^4}{4!} - \ldots) - 1}{z} dz \\
&= -\int_0^x \frac{z}{2!} - \frac{z^3}{4!} + \ldots dz
\end{aligned}$$

  thus

$$|C_+(x)| < \int_0^x \frac{z}{2!} dz = \frac{x^2}{4} < \frac{1}{4}$$

  and

$$\begin{aligned}
|C(t)| &= |\frac{1}{2\pi}(\gamma + C_+(\omega_+ t) - \gamma - C_-(\omega_- t))| \\
&= |\frac{1}{2\pi}(C_+(\omega_+ t) - C_-(\omega_- t))| \\
&\leq \frac{1}{2\pi}(|C_+(\omega_+ t)| + |C_-(\omega_- t)|) \\
&\leq \frac{1}{2\pi}(1/4 + 1/4) = \frac{1}{4\pi} \qquad \text{(A.2)}
\end{aligned}$$

- Case 2: if $t \geq 1$, then

$$\begin{aligned}
C_+(x) &= \int_0^x \frac{\cos(z) - 1}{z} dz \\
&= \int_0^x \frac{\cos(z)}{z} dz - \int_0^x \frac{1}{z} dz \\
&= C_i(z) - [\log z]_{z=0}^{z=x}
\end{aligned}$$

  where $C_i(x) = \int_0^x \frac{\cos(z)}{z} dz$. Thus

$$\begin{aligned}
C(t) &= \frac{1}{2\pi}(\gamma + C_i(\omega_+ t) - [\log z]_{z=0}^{z=\omega_+ t} - \gamma - C_i(\omega_- t) + [\log z]_{z=0}^{z=\omega_- t}) \\
&= \frac{1}{2\pi}(C_i(\omega_+ t) - \log(\omega_+ t) + lim_{x\to 0}\log x - C_i(\omega_- t) + \log(\omega_- t) - lim_{x\to 0}\log x) \\
&= \frac{1}{2\pi}(C_i(\omega_+ t) - C_i(\omega_- t) + \log\frac{\omega_- t}{\omega_+ t})
\end{aligned}$$

therefore

$$\begin{aligned}
|C(t)| &= \frac{1}{2\pi}(|C_i(\omega_+ t)| + |C_i(\omega_- t)| + |\log\frac{\omega_- t}{\omega_+ t}|) \\
&\leq \frac{1}{\pi}\max_{x>=1}|C_i(x)| + \frac{1}{2\pi}|\log\frac{\omega_- t}{\omega_+ t}|
\end{aligned}$$

The absolute value of the log above will be large either if $\omega_- t$ is small or if $\omega_+ t$ is large; that is, either if the input frequency is very close to the cutoff frequency or if the cutoff is large. We are interested in signals with frequency under 2 kHz and in the case when the input frequency is not close to the cutoff frequency, thus we may say that the absolute value of the log is under $\pi/2$. To get an upper bound for the first part we refer to figure A.2. The maximum value $\max_{x>=1}|C_i(x)| < 0.5$, thus

$$|C(t)| \leq \frac{1}{4\pi} + \frac{1}{4} < 1/3 \tag{A.3}$$

Combining A.2 and A.3 yields for $t > 0$

$$|C(t)| \leq 1/3 \tag{A.4}$$

Next we will show that the peak of $S(t)$ is approximately three times larger than the above max for $C(t)$, and that $S(t)$ peaks close to $t = 0$ if the input frequency is not close to the cutoff frequency (i.e. if ω_- is not close to zero and positive).

Since

$$\begin{aligned}
\int_{-\infty}^{t}\frac{\sin(\alpha\tau)}{\tau}d\tau &= \\
\int_{-\infty}^{0}\frac{\sin(\alpha\tau)}{\tau}d\tau + \int_{0}^{t}\frac{\sin(\alpha\tau)}{\tau}d\tau &= \\
\frac{\pi}{2} + S_i(\alpha t)
\end{aligned}$$

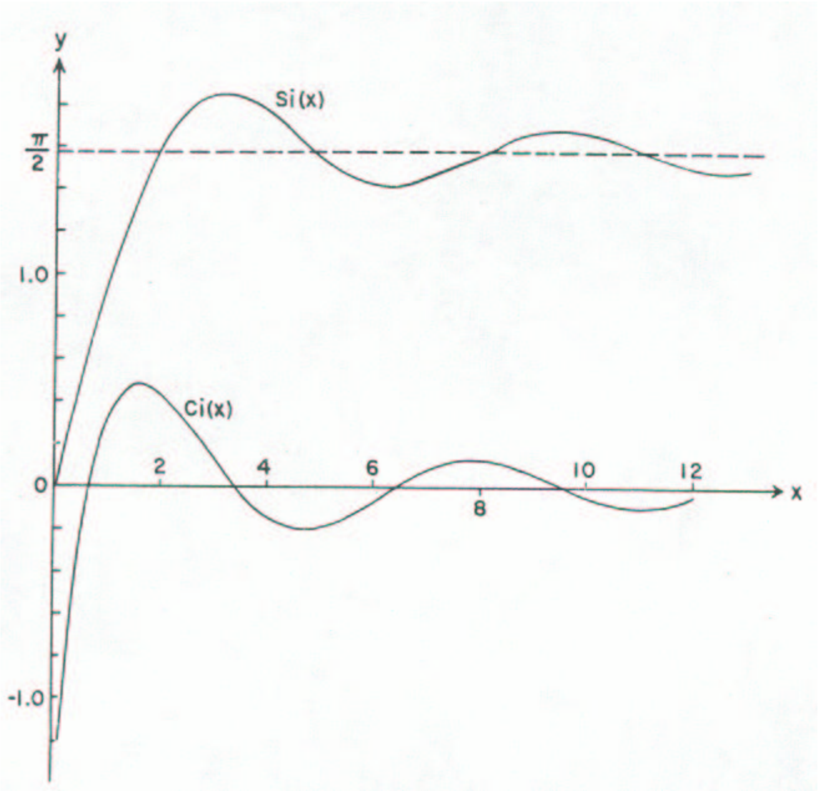


Figure A.2: Graphs of $S_i(x) = \int_0^x \frac{\sin(z)}{z}dz$ and $C_i(x) = \int_0^x \frac{\cos(z)}{z}dz$. (Figure 5.6 from [2, page 232])

we can write

$$\begin{aligned} S(t) &= \frac{1}{2}\int_{-\infty}^{t} \frac{\sin(\omega_+\tau) + \sin(\omega_-\tau)}{\pi\tau}d\tau \\ &= \frac{1}{2\pi}(\frac{\pi}{2} + S_i(\omega_+t) + \frac{\pi}{2} + S_i(\omega_-t)) \\ &= \frac{1}{2}(1 + \frac{1}{\pi}(S_i(\omega_+t) + S_i(\omega_-t))) \end{aligned}$$

Once again we refer to figure A.2 to get a feel for the shape of $S_i(x)$ (see figure A.2). From this it is easy to see that $S(t)$ is 0.5 at 0, it rises relatively abruptly from 0.5 to about 0.75 at $t = 1/\omega_+$, and is slowly

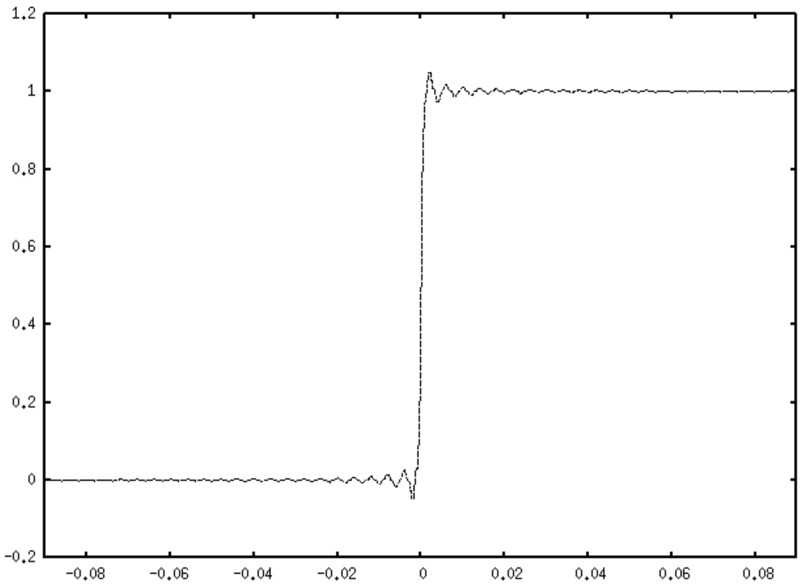


Figure A.3: Graph of $S(t)$ with $\omega_c = 500$ and $\omega_0 = 250$. The abscissa value where the function peaks is inversely proportional to $\omega_- = \omega_c - \omega_0$.

increasing afterwards to reach 1 at $t = 1/\omega_-$ (see figures A.3 A.4 A.5). This means that the peak response of the ideal bandpass filter is determined by the peak of $S(t)$ since this peak is much larger than the peak of $C(t)$. The timing of this peak response is exactly as we want it: the peak's distance from the start of the input is inversely proportional to $\omega_- = \omega_c - \omega$.

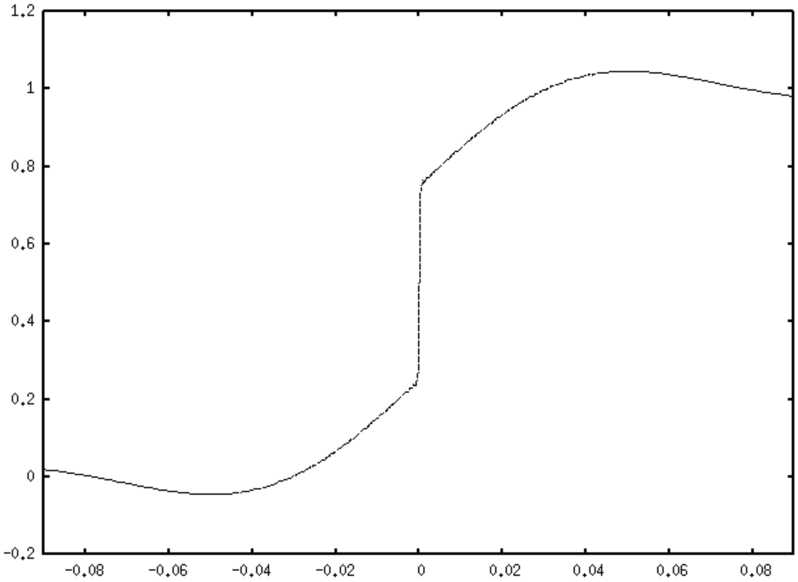


Figure A.4: Graph of $S(t)$ with $\omega_c = 500$ and $\omega_0 = 490$. The abscissa value of where the function peaks is inversely proportional to $\omega_- = \omega_c - \omega_0$; the abscissa value of the first abrupt rise after zero is inversely proportional to $\omega_+ = \omega_c + \omega_0$. See also the next figure.

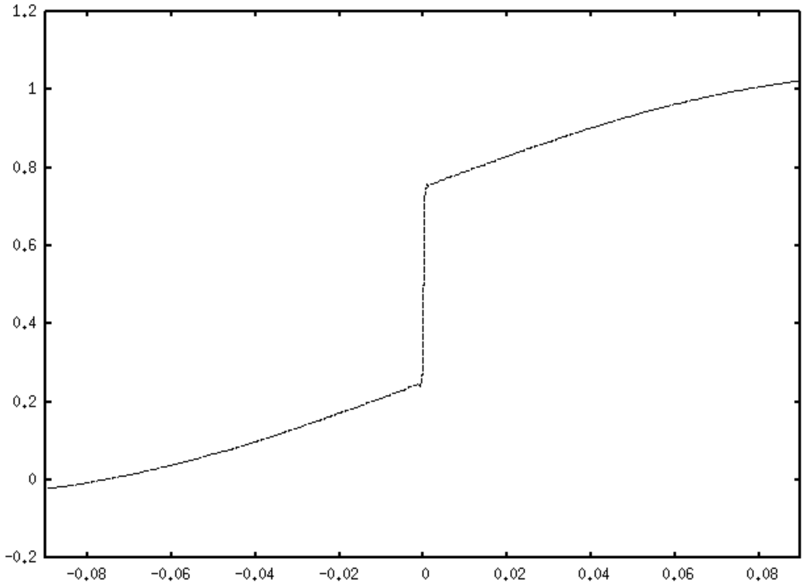


Figure A.5: Graph of $S(t)$ with $\omega_c = 500$ and $\omega_0 = 496$

Appendix B

Sub–pitch-period time-pattern examples

Each of the following nine figures show the waveform, the low channel, the smoothest channel and the smoothest channel energy time-aligned with one another. The hand-drawn humps in the smoothest channel indicate the locations of the vowel sub–pitch-period patterns. The examples display the first few (2-3) pitch periods of the vowel *ae* immediately after a voiced stop consonant: *g*, *d*, and *b* respectively. The low channel is the lowest frequency bandpass channel that we used. The smoothest channel is a channel that is spliced together from bandpass channels: during each vowel pitch-period we chose the channel that "smoothest" (see full definition in the Experiments chapter; see more detailed explanation of the figures in the Observations chapter).

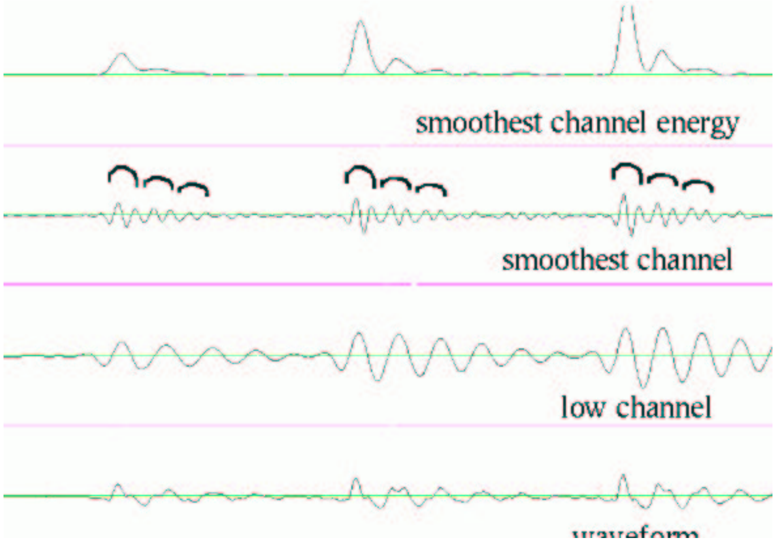


Figure B.1: $b \rightarrow ae$ transition.

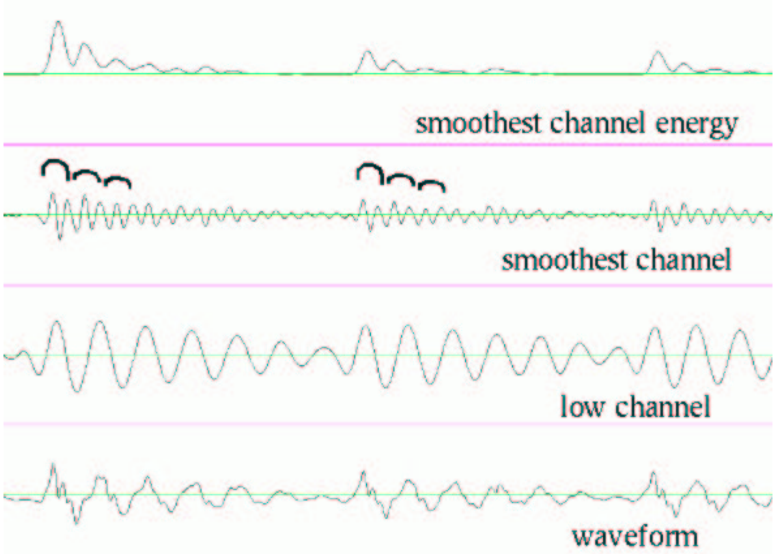


Figure B.2: $b \rightarrow ae$ transition.

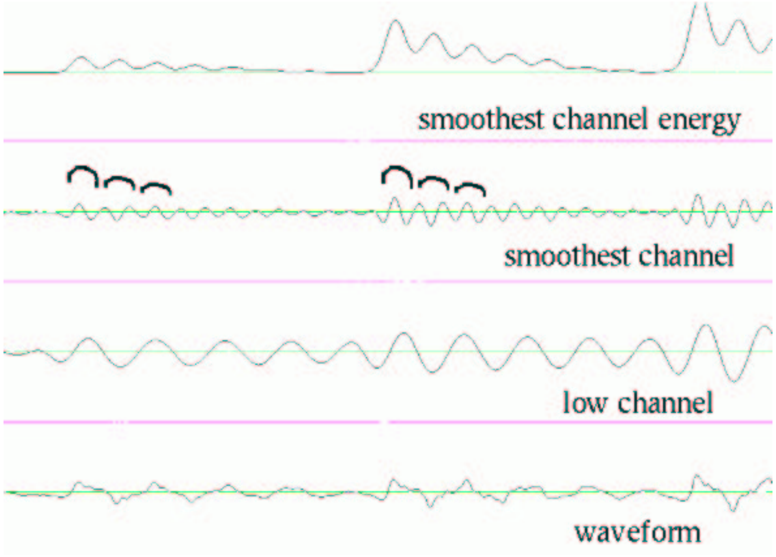


Figure B.3: $b \rightarrow ae$ transition.

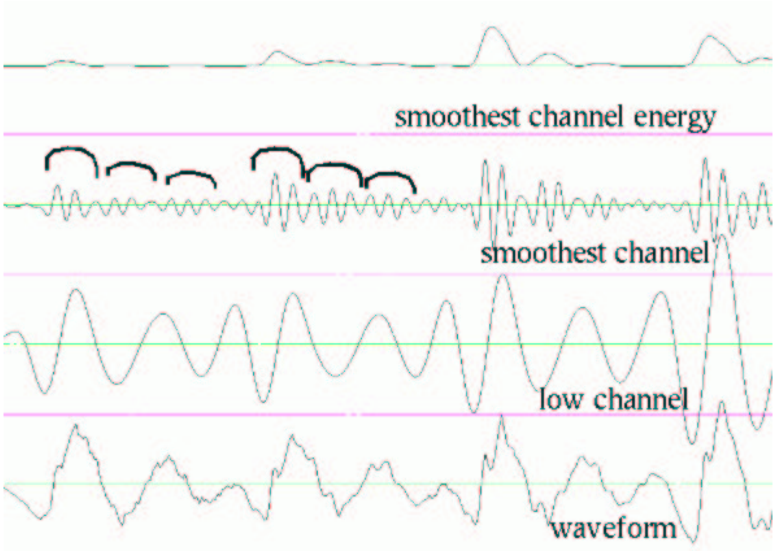


Figure B.4: $d \rightarrow ae$ transition.

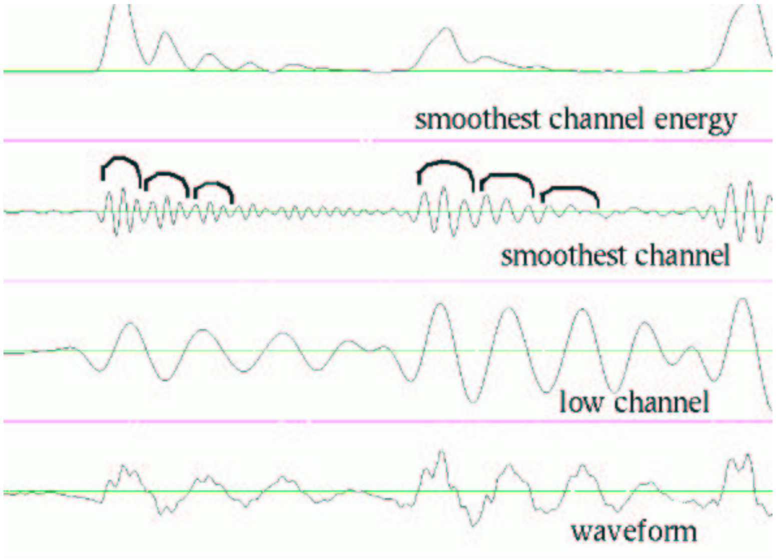


Figure B.5: $d \rightarrow ae$ transition.

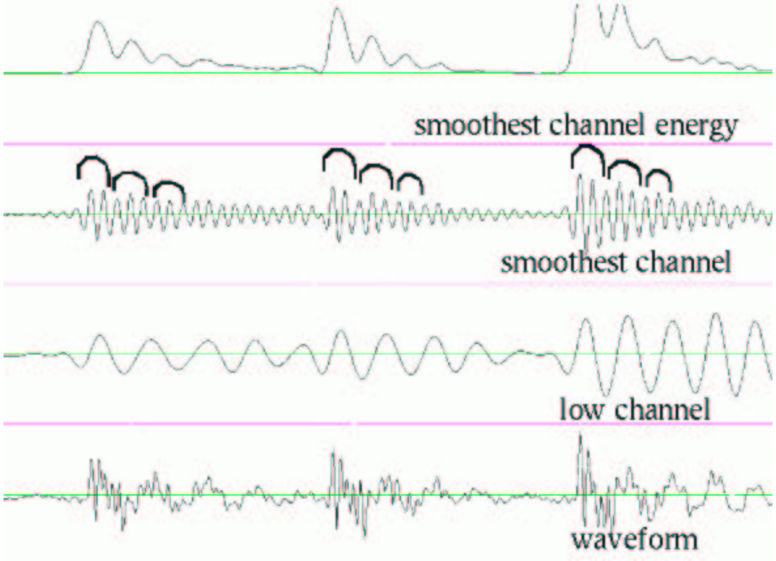


Figure B.6: $d \rightarrow ae$ transition.

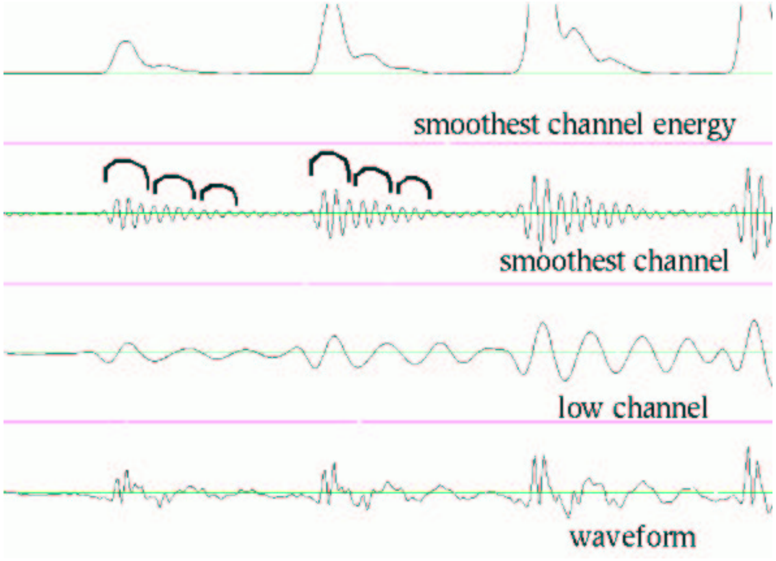


Figure B.7: $g \rightarrow ae$ transition.

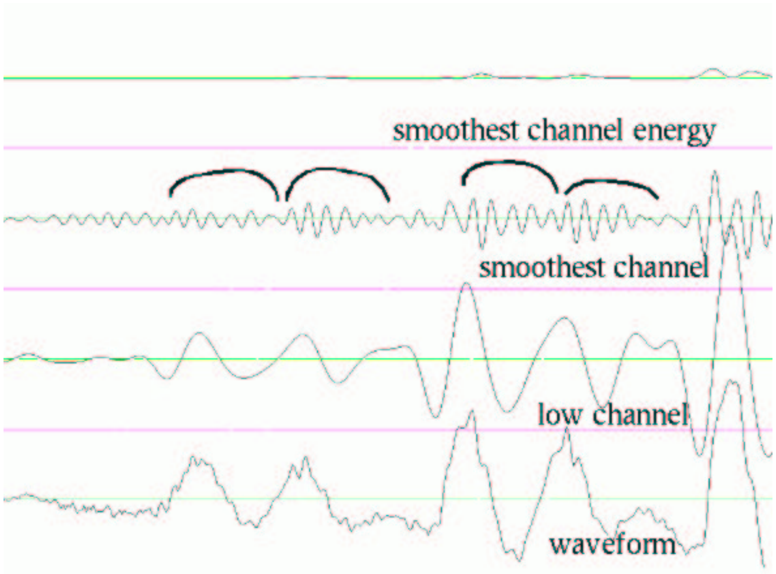


Figure B.8: $g \rightarrow ae$ transition.

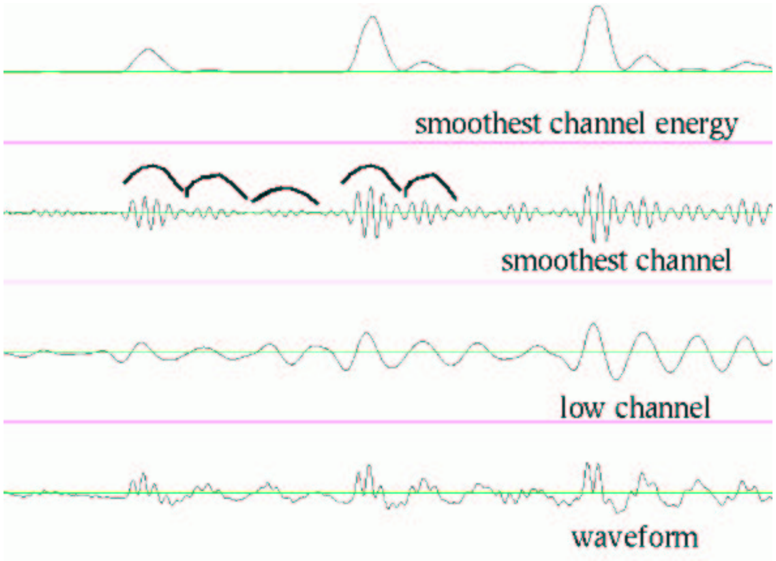


Figure B.9: $g \rightarrow ae$ transition.

Bibliography

[1] J. R. Buck A. V. Oppenheim, R. W. Schafer. *Discrete-Time Signal Processing.* Brown University, Academic Press Inc., 1998.

[2] M. Abramovitz and I. A. Stegun. *Handbook of mathematical functions.* United States Department of Commerce, National Bureau of Standards, 1964.

[3] N. Y. S. Kiang B. Delgutte. Speech coding in the auditory nerve. *J. Acoust. Soc. America*, 75(3):866–878, 1984.

[4] J. M. Baker. *A new time-domain analysis of human speech and other complex waveforms.* PhD thesis, Carnegie Melon University, 1975.

[5] S. Blumstein and K. N. Stevens. Acoustic invariance in speech production: Evidence from measurements of the spectral characteristic of stop consonants. *J. Acoust. Soc. Am.*, 66(4):1001–1017, 1979.

[6] Rens Bod. Memory-based models of melodic analysis: Challenging the gestalt principles. *Journal of New Music Research*, 31(1):27–37, 2002.

[7] M. Braun. Tuned hair cells for hearing, but tuned basilar membrane for overload protection: Evidence from dolphins, bats, and desert rodents. *Hearing Research*, 78:98–114, 1994.

[8] R. Kil D. Kim, S. Lee. Auditory processing of speech signals for robust speech recognition in real-world noisy enviroment. *IEEE Transaction on Speech and Audio Processing*, 7(1):55 – 69, 1999.

[9] A. Dancer. Experimental look at cochlear mechanics. *Audiology*, 31:301–312, 1992.

[10] P. C. Delattre, A. M. Lieberman, and F. S. Cooper. Acoustic loci and transitional cues for consonants. *J. Acoust. Soc. Am.*, 27:769–773, 1955.

[11] J. L. Flanagan. *Speech Analysis Synthesis and Perception.* Springer-Verlag, Berlin-Heidelberg-New York, 1972.

[12] R. Romand G. Ehret. *The Central Auditory System.* Oxford University Press, 1997.

[13] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S Pallett, and N. L. Dahlgreen. *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM.* US Dept. of Commerce, NIST, 1993.

[14] D. C. Geisler. *From Sound to Synapse.* Oxford University Press, 1998.

[15] O. Ghitza. Auditory models and human performance in tasks related to speech coding and speech recognition. *IEEE Trans. on Speech and Audio Processing*, 2:115–132, January 1994.

[16] J. J. Gibson. Adaptation with negative after-effect. *Psychological Review*, 44:222–244, 1937.

[17] G. W. Gray. Phonemic microtomy: The minimum duration of perceptible speech sounds. *Speech Monographs*, 9:75–90, 1942.

[18] T. G. Bever H. B. Savin. The nonperceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behaviour*, 9:295–302, 1970.

[19] S. M. Teager H. M. Teager. A phenomenological model for vowel production in the vocal tract. In *Speech Science: Recent advances*, pages 73–109, 1985.

[20] S. Teager H. M. Teager. Evidence for nonlinear sound production mechanisms in the vocal tract. In *Proc. NATO ASI on Speech Production and Speech Modelling*, pages 241–261, 1990.

[21] H. Hermansky. Perceptual linear predictive (plp) analyis of speech. *J. Acoust. Soc. America*, 87(4):1738–1752, 1990.

[22] M. S. Howe. *Acoustics of fluid-structure interactions.* Cambridge University Press, 1998.

[23] M. S. Howe. *Theory of vortex sound.* New York: Cambridge University Press, 2003.

[24] E. H. Huizing and A. Spoor. An unusual type of tinnitus. *Archives of Otolaryngology*, 98:134–136, 1973.

[25] H. Hermansky J. C. Junqua, H. Wakita. Evaluation and optimization of perceptually-based asr front-end. *IEEE Trans. Speech and Audio Processing*, 1(1):329–338, 1993.

[26] J. Algeria J. Morais, L. Cary and P. Bertelson. Does awareness of speech as a sequence of phones arise spontaneously? *Cognition*, 7:323–331, 1979.

[27] F. Jelinek. *Statistical Methods for Speech Recognition.* The MIT Press, 1997.

[28] J. C. Junqua. *Toward robustness in isolated-word automatic speech recognition.* PhD thesis, University of Nancy, 1989.

[29] J. F. Kaiser. Some observations on vocal tract operation from a fluid point of view. In *Vocal Fold Physiology*, pages 358–386, 1983.

[30] N. Y. S. Kiang. *Handbook of Physiology. The Nervous System. Sensory Processes.* Bethesda, MD: Am. Physiol. Soc., 1984.

[31] K. Koffka. *Principles of Gestalt Psychology.* New York: Harcourt Brace, 1935.

[32] D. J. Lim. Cochlear anatomy related to cochlear micromechanics. a review. *J. Acoust. Soc. America*, 67:1686–1695, 1980.

[33] R. P. Lippman. Recognition by humans and machines: Miles to go before we sleep. In *Speech Communication*, Vol. 18, p. 247 1996.

[34] E. D. Young M. B. Sachs. Encoding of steady state vowels in the auditory-nerve: representation in terms of discharge rate. *J. Acoust. Soc. America*, 66:470–479, 1979.

[35] E. D. Young M. B. Sachs. Encoding of steady state vowels in the temporal aspects of the discharge patterns of populations of auditory nerve fibers. *J. Acoust. Soc. America*, 66(5):1381–1403, 1979.

[36] C. Lefebvre M. Hunt. Speech recognition using cochlear model. In *Proc. Int. Conf. on Acoust. Speech, Signal Processing*, pages 37.7.1–37.7.4, 1986.

[37] J. R. Mundie M. W. Moore. Specification of the minimum number of glottal pulses necessary for reliable identification of selected speech sounds. *Aerospace Medical Research Laboratory Report*, TR-70-104.

[38] S. G. Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, 1999.

[39] E. Mousset, W. Ainsworth, and J. A. R. Fonollosa. A comparison of several recent methods of fundamental frequency and voicing decision estimation. In *Proc. ICSLP '96*, volume 2, pages 1273–1276, Philadelphia, PA, 1996.

[40] A. Papoulis. *Signal Analysis.* McGraw-Hill Book Company, 1977.

[41] T. F. Quatieri. *Speech signal processing: principles and practice.* Prentice Hall Signal Processing Series, 2002.

[42] Chalikia R. M. Warren, Healy. Auditory illusions and perceptual processing of speech. *Principles of Experimental Phonetics*, pages 435–466, 1996.

[43] R. E. Remez. Adaptation of the category boundary between speech and nonspeech. *Cognitive Psychology*, 11:38–57, 1979.

[44] S. Roweis. *Articulatory Speech Processing.* PhD thesis, California Institute of Technology, January 1999.

[45] F. Itakura S. Kajita. Speech analysis and speech recognition using subband-autocorrelation analysis. *J. Acoust. Soc. Japan*, 15 no 5:329–338, 1994.

[46] F. Itakura S. Kajita. Robust feature extraction using sbcor analysis. In *Proc. Int. Conf. on Acoust. Speech, Signal Processing*, pages 421–424, 1995.

[47] C. L. Searle, J. Z. Jacobson, and S. G. Rayment. Stop consonant discrimination based on human audition. *J. Acoust. Soc. Am.*, 65:799–809, 1979.

[48] P. M. Sellick, R. Patuzzi, and B. M. Johnstone. Measurement of basilar membrane motion in the guinea pig using the mossbauer technique. *J. Acoust. Soc. America*, 72:131–141, 1982.

[49] S. Seneff. *Pitch and spectral analysis of speech based on an auditory synchrony model.* PhD thesis, Massachusetts Institute of Technology, 1985.

[50] S. Seneff. A joint synchrony/mean-rate model of auditory processing. *J. Phonetics*, 16(1):55–76, 1988.

[51] K. N. Stevens. *Acoustic Phonetics.* The MIT Press, 1998.

[52] A. Suchato. *Classification of stop consonant place of articulation.* PhD thesis, Massachusetts Institute of Technology, June 2004.

[53] H. M. Teager. Some observations on oral air flow during phonation. *IEEE Trans. Acoustics, Speech, Signal Processing*, ASSP-28(5):599–601, 1980.

[54] G. von Békésy. *Experiments in Hearing.* McGraw-Hill, New York, 1960.

[55] R. M. Warren. Perceptual restoration of missing speech sounds. *Science*, 167:392–393, 1970.

[56] R. M. Warren. Criterion shift rule and perceptual homeostasis. *Psychological Review*, 92:574–584, 1985.

[57] R. M. Warren. *Auditory Perception.* Cambridge University Press, 1999.

[58] H. M. Winitz and J. Reeds. Identification of stops and vowels for the burst portion of /p,t,k/ isolated from conversational speech. *J. Acoust. Soc. Am.*, 51(4):1309–1217, 1972.

[59] J. R. Westbury with G. Turner and J. Dembowski. *X-ray microbeam speech production database user's handbook.* Waisman Center on Mental Retardation & Human Development, University of Wisconsin, 1994.

[60] N. Y. Kiang with the assistance of T. Watanabe, E. C. Thomas, and Louise F. Clark. *Discharge Patterns of Single Fibers in the Cat's Auditory Nerve.* MIT Press, 1965.

[61] A. J. Oxenham Z. M. Smith, B. Delgutte. Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, 416:87–90, March 2002.

[62] V. W. Zue. *Acoustic characteristic of stop consonants: A controlled study.* Massachusetts Institute of Technology, 1976.