ON THE NON-EQUIVALENCE OF FUNDAMENTAL FREQUENCY
AND PITCH IN TONAL DESCRIPTION

Philip John Rose

INTRODUCTION

This paper uses acoustical data from Chinese and Thai to examine the relationship between the acoustic and perceptual aspects of the linguistic category of Tone. Specifically, it questions the current assumption in linguistics that tonal pitch is a function of fundamental frequency (F0).1

In section 1 I characterise the way in which the relationship between tone and its acoustics is conceived in linguistics. Section 2 examines in some detail the relationship between F0 and linguistic pitch. Section 3 presents acoustic data from two tone languages which demonstrate that tonal pitch is not a straightforward function of F0.

1. Perceptual and acoustic correlates of tone

Tone is characterised in the modern linguistic literature as a contrastive use of the perceptual dimension of pitch, and tonal pitch is equated, exclusively and straightforwardly, with the acoustical dimension of fundamental frequency. This conception of tone can be illustrated by representative quotes from two well-known and widely cited reference works: Tone A Linguistic Survey (Fromkin 1978), and Suprasegmentals (Lehiste 1970):

In general (along with the other authors represented in this book) we will take tonal distinctions to be a subset of those corresponding to differences in pitch (a function, in turn, of fundamental frequency, or F0)


...pitch will be used to refer to the perceptual correlate of frequency, tone will be used to refer to the feature when it functions distinctively at word level...

(Lehiste 1970:54; see also Table 1.1, p.4)
The assumed straightforward relationship between tonal pitch and F0 (‘pitch is F0’) is reflected in the widespread use of F0 data to infer, evaluate, or justify pitch descriptions and, by extension, the phonological analyses that use them as input, e.g. Cheng (1973:270ff.; Zee (1980); Zee and Maddieson (1980); Chan and Ren (1986); Howie (1974); William S-Y Wang (1968). Other, more trivial, reflections of this assumption are the synonymous use of the terms F0 and pitch in the literature, e.g. Ohala (1978:6), and in the naming of acoustical instruments and their output (‘pitch meter’, ‘pitch curve’ for F0 meter and F0 curve).

The assumed exclusiveness of the relationship between linguistic category and acoustic dimension (‘tone is F0’) is arguably reflected in the fact that tonal acoustics, and in fact most prosody, are normally described in terms of F0 alone (Sorin 1981:359; Rose 1982b:18,19).

As well as the primary cue of pitch with its correlate of F0, the linguistic approach also recognises secondary tonal correlates like phonation type, phonation offset, vowel length and quality (Ohala 1973:4). These correlates are called secondary because they participate, albeit distinctively, in relatively few tonal contrasts (creak in Vietnamese; duration in Shanghai), or because they can be linguistically analysed as phonemically non-distinctive, (duration in Modern Standard Chinese; ‘constricted’ phonation in Standard Thai /high/ tone). Because of their secondary status, little attention is accorded them, and the question of their acoustic correlates is usually ignored.

The linguistic approach outlined above has been criticised (Coster and Kratochvil 1984), because it assumes tone has correlates in a single acoustic and perceptual dimension. It is very likely that such a conception, in its exclusive citation of F0, oversimplifies the relationship between linguistic unit and acoustic properties that encode it. The complexity of the relationship between speech acoustics and perception is well known (Studdert-Kennedy 1976:247-251). One aspect of this complexity ignored by the monodimensional approach is the fact that segmental or suprasegmental phonetic features like [voice] or [stress] - and by extension the phonetic and phonemic segments they make up - are typically signalled by a hierarchy of acoustical cues, not necessarily all from the same acoustic dimension (Studdert-Kennedy 1976:249; Abramson and Lisker 1985:25; Lehiste 1970:125-142). There is therefore no reason to assume the situation will be any different with the acoustic cues to tonal features like [Hi], [Fall], or phonetic or phonemic tones like [33] or /33/.

Kratochvil (1975:7, 8; 35-39) sees this oversimplification as symptomatic of a ‘phonological state of mind’, with its familiar emphasis on underlying simplicity to account for surface diversity. Doubtless, such an approach may be justifiable when trying to find an elegant description for the behavior and interrelationship of higher level phonological units like phonemes and morphophonemes. But the monodimensional assumption is indefensible as a heuristic in phonetic research. Rather than let phonological considerations inform phonetic research, many scholars (e.g. Ohala (1979); Linell (1979); Lass (1984)), now insist on the importance of empirical phonetic research for evaluation of phonological questions. Recent speculations on tonogenesis, which make reference to plurality of acoustic cues to a phonological feature, are one example of the usefulness of this approach.
It is important, therefore, to ascertain the nature of the oversimplification involved in equating tonal pitch with F0, and the aim of this paper is to clarify the extent to which F0 can be considered the acoustical correlate of pitch for the linguistic category of Tone. I shall do this by first looking at some general points concerning the relationship between F0 and pitch. Then I shall examine a particular instance of a tonal contrast in a variety of Chinese where pitch is demonstrably not a straightforward correlate of F0, to see what other factors need to be taken into account to relate F0 to pitch.

2. F0 and pitch

In order to be able to discuss the relationship between F0 and pitch precisely, two points must be clarified. Firstly, we need to be clear about what entities are being related. F0 is the basic rate of repetition of the quasi periodic part of the complex radiated speech wave, and is an easily quantifiable acoustical parameter. I shall use the term (tonal) pitch to refer to an auditory percept transcribed by linguists, ideally using (or convertible into) some kind of integer representation like the well known 5 point scale (Chao 1930). As such, it is considered the appropriate observation language (Lass 1984:7) for describing and classifying tone and tonal contrasts on both phonetic and phonological levels (Wang 1968; Anderson 1978:136).

The second point to be mentioned concerns the two senses in which F0 and pitch are said to be related. F0 is often described as the acoustical correlate of pitch; and tonal pitch is also described as a function of F0 (see e.g. the quotes above). The use of the term correlate usually implies nothing more than that changes in pitch will be accompanied by changes in F0. Although falsifiable, this claim does not emphasise the precise nature of the relationship. The notion of function, however, does: to say that pitch is a function of F0 means that there exists an expression which assigns a unique pitch value to any F0 value. It is therefore appropriate to ask to what extent it is true that there is a function which will permit the derivation of a tonal pitch specification like [41] from a set of F0 values. (I am of course assuming that linguists will agree in their pitch transcription of tones. Whether this is justified remains to be demonstrated: there is to my knowledge no data on pitch transcription comparable to Laver's (1965) quantification of variation in vowel transcription.)

There are several transformations that the F0 in a speech wave has to undergo - at least conceptually - before it can possibly be perceived as linguistic pitch, and the extent to which these transformations refer to parameters other than F0 indicates that pitch is not an exclusive function of F0.

A lot of psycho-acoustic research has been done on pitch perception of non speech stimuli (pure tones, periodic noise bursts etc.) and it is possibly the results of this research which were originally responsible for the equation of F0 and pitch in linguistics. The basics are well-known: for non-speech stimuli, pitch can effectively be considered a linear function of frequency below, and a logarithmic function above 1 KHz (Ladefoged 1962:77). Changes in amplitude and duration exert a negligible influence on the way frequency is perceived as pitch, at least for the ranges which are of interest here (Lehiste 1970:62-67; Ladefoged 1962:71). Psychoacoustic research
indicates, therefore, that F0 is going to be an independent variable, and possibly the most important term, in the function linking acoustics to tonal pitch.

However, it is difficult to say to what extent these findings hold for the perception of F0 as pitch in speech (Gandour 1978:42). The radiated time-pressure speech wave with its complex frequency content constitutes a very much more complex stimulus that a pure tone. It has mathematical properties (narrow-peaked spectrum; rising and falling time domain amplitude; changes in short term spectrum) which possibly trigger off a speech specific mode of perception (Stevens and Blumstein 1981:4,5) which interprets physical input differently depending on whether it is recognised as speech or non speech.

However, one processing stage probably common to all auditory input is the initial 'auditory' stage where acoustic energy is peripherally converted into patterns of neurological activity from which features like F0 can then be extracted, possibly at the cortex (Studdert-Kennedy 1976:245). These transformations will inevitably involve some loss of frequency information because of the operational constraints on the peripheral hearing mechanism and smoothing in higher neuronal networks, and some of these constraints are durationally related. Psychoacoustic studies have shown, for example, that various masking effects like sensitivity to variation in the frequency derivative are a function not only of size of frequency change but also stimulus duration (Collier 1984:243). Even at the initial stage of speech perception, then, pitch cannot be dependent on F0 alone.

The speech-specific mode of perception generally involves hearing stimuli as sounds produced by a human vocal tract, that is, interpreting the stimuli linguistically in terms of the anatomical and physiological constraints of the vocal tract that produced them (Studdert-Kennedy 1974:2351). As a part of this process, various influences on F0 have to be factored out by some kind of perceptual normalisation, before the linguistically relevant residue can be correctly interpreted.

The effect on F0 of the length and mass of the individual speaker's vocal cords is a well-known example. The resulting between-speaker differences in F0 are easily large enough to allow the possibility of the same F0 values to represent two phonologically different pitches, say a female's low level pitched tone and a male's high level pitched tone. However, effective between-speaker normalisation of F0 seems possible using just F0 parameters, for example mean and standard deviation, or range (Rose 1987a).

Within-speaker variance in F0 occurs as a result of differences in context, that is, where the F0 occurs relative to some other influence. One of the best known examples of this is found in the phenomenon of declination. F0 values tend to decrease throughout an utterance, possibly as the result of a gradual decrease in sub-glottal pressure (Ps) (but see Ladd (1984:62,63) for other explanations). This decrease is allowed for perceptually (Pierrehumbert 1979; Collier 1984:238), so that it is theoretically possible for the same F0 to signal a low pitched tone at the beginning of an utterance, and a high pitched tone at its end.

As a result of the same mechanism (perception of F0 contours in terms of the gestures involved in their production) it should be possible for the same F0 in the same position in two utterances to be perceived as a different pitch, depending on differences in context. For example, a stressed syllable might cause the same following F0 to be perceived as relatively higher than it would be
after an unstressed syllable. An effect of this kind is argued for by Lieberman (1967:48-61), in attempting to explain why listeners' judgements as to whether a terminal F0 contour rose or fell in pitch depended on the height of a preceding F0 peak. A higher preceding F0 peak (he argues) indicates a relatively greater local Ps increase, after which the Ps level is relatively lower. Because listeners are aware of the positive correlation between Ps and F0 (Ladefoged 1967:29-35), they accordingly lower their frame of reference for pitch after the local F0 peak.

There does not appear to be any reason why F0 data alone should not constitute sufficient information to allow correct perception of these effects, although it might be expected that the contribution of Ps to local stress-related F0 increases and global declination effects can possibly be better estimated by a combination of F0 and (radiated) amplitude (Ar) (Monsen et al. 1978).

One context effect which definitely does require information other than F0 for correct pitch interpretation is compensation for the so-called intrinsic 'pitch' of vowels. It is well known that F0 correlates positively with vowel height. The magnitude of the effect reported for some languages, e.g. English, Serbo-Croatian (Hombert 1978:97; Lehiste 1970:68-71) seems just about large enough to cancel some tonemically relevant differences in F0. (Lehiste (1970) gives a difference of 21 Hz between English [i] and [a], which can be compared with the mean maximum difference of 16 Hz between one Thai speaker's /mid/ and /high/ tones shown in Gandour (1978:43)).

It has been hypothesised (Hombert 1978:83) that intrinsic F0 differences associated with prevocalic consonants are minimised in true tone languages to avoid perceptual confusion. It would be interesting to see whether a similar effect occurs for intrinsic differences associated with vowel height. A glance at some of the available relevant data (Zee 1980; Han 1969; Rose 1987b) suggests that the size of such differences is a function of both tone category and speaker's F0 range, and that tonemically relevant differences in F0 will not be swamped by intrinsic differences due to vowel height. Nevertheless there is evidence that the perceptual mechanism makes a small (<5Hz) allowance for the intrinsic F0 of vowels: a low vowel will tend to be perceived higher in pitch than a high vowel of the same F0 (Hombert 1978:99). (I am not aware of any study which investigates whether the same perceptual effect can be induced by differences in manner on a prevocalic consonant.) Since the perceptual mechanism must presumably refer to spectral information (normalised F1 frequency) to determine the height of a vowel, this is a clear case of acoustic dimensions other than F0 contributing to pitch perception.

Whereas psychoacoustic studies show little effect of Ar on the perception of pitch within the ranges of interest, there is a limited amount of evidence from perceptual experiments on synthetic and resynthesised material that Ar does play a role in pitch perception in speech. Sorin (1981:364), for example, mentions an experiment (Pierrehumbert 1979) in which the same pitch percept was cued by different combinations of Ar and F0. Differences of 11 Hz and 3.6 dB between two stressed syllables, for example, were found to be perceptually equivalent to differences of 17 Hz and 0 dB. This kind of compensatory interaction between F0 and Ar points once again to a speech specific perception mediated by productional considerations. Here, the perceptual mechanism appears to be judging the Ar and F0 in terms of an evaluation of the Ps and vocal cord tension (VCT) involved in their production. However, as Pierrehumbert herself points
out (367), it is more likely that the combinations of Ar and F0 were being interpreted as cues for prominence rather than pitch.3

The significant role of Ar in pitch perception in speech has been demonstrated by Rossi (1978) with discrimination tasks on 20 csec synthesised vowels. These experiments show that Ar and F0 interact, but not entirely in the same compensatory way as reported by Pierrehumbert. For example, changing (increasing?) the Ar from a -8 dB drop to level meant that a concomitant F0 rise of 22 Hz had to be compensatorily decreased to 19 Hz for it to be perceived as the same rise in pitch. However, the same change (increase?) in Ar with a falling F0 of -23 Hz required an increase in F0 to a -19 Hz drop to cue the same pitch drop. This latter case is not consistent with productionally mediated perception, but a similar effect occurs with Zhenhai tones 1 and 5: tone 5, which has a considerably higher Ar than tone 1, has the same F0 as the onset of tone 1, but a higher pitch (Rose 1986:133).

It may be, of course, that the differences between these two experiments can be explained by assuming Ar and F0 are perceived differently, depending on whether the listener is making prominence or pitch judgements. However, since pitch is also an important perceptual parameter for stress, it is still quite clear that Ar does play a small but significant role in pitch perception.

Since the same pitch can be cued by different combinations of F0 and Ar it remains to be investigated to what extent different combinations of F0 and Ar are regularly used to signal the same pitch, either by different speakers, or by the same speaker under different circumstances.

A possibly more important finding of the Rossi study is that Ar contributes to contouricity. A combination of increasing Ar and F0 cues concave pitch; increasing Ar and falling F0 cue convex pitch (390, 392). This kind of perceptual interaction between F0 and Ar may be another factor in discrepancies between pitch percepts and F0 traces. For example, Zhenhai tone 1, which has a pronounced Ar shoulder, has a falling F0 but a pitch of [442] with an initial level component (Rose 1982b:158).

Finally, although we are primarily concerned here with perception of tonal pitch, there is evidence of the role of Ar in identification of tone from perceptual experiments with native speakers. Abramson (1975:5-8) found that the efficacy of F0 as a cue is enhanced by the presence of suitable Ar contours.

Indirect evidence for the importance of duration in the perception of F0 is offered by Kratochvil (1985), who has demonstrated that the shape of the F0 of tonal syllables in normal Peking dialect speech is a function of the duration of the tone-carrying part of the syllable. Kratochvil notes that duration in normal speech is primarily determined, not by segmental features, but by tempo and stress factors. This allows for the possibility that a tone with a low falling pitch in citation form or slow speech may have the same F0 contour as a tone with a high falling pitch said quickly in normal speech. Kratochvil hypothesises that duration has therefore to be taken into account before the F0 can be correctly interpreted.
2.1 Summary

The considerations discussed above indicate that tonal pitch is the perceptual result of both general auditory and speech specific processes operating on F0. The speech specific processes typically involve productionally mediated perception, and include speaker and context normalisation of F0. Both of these (but especially the latter) involve interaction, in ways mostly not yet clear, between F0 and all the other main acoustic dimensions of Ar, duration and spectrum. Thus tonal pitch cannot be an exclusive function of F0, although consideration of the magnitude of perceptual effects indicates that F0 constitutes the basic term in the function relating acoustics to pitch. In the second half of this paper, I shall present evidence to suggest that intrinsic effects on F0 are an additional factor which mediates the relationship between F0 and pitch.

3. Falling versus convex pitched tones - acoustical data

In the varieties of the Chinese dialect of Zhenhai, (a rural county NE of the municipality of Ningpo in Zhejiang province), there is a surface tonal contrast between falling and convex pitch after a low level pitch on a preceding syllable (i.e. [11 41] vs. [11 131]). Examples of minimal pairs are given below. (The phonotactic structure of the syllable is /C(G)V(η)/, where C = obstruent/sonorant; G = glide; V = syllabic vocoid).

1) with intervocalic obstruent:

[jə 'zy 11 41] 搭船 ‘rowboat’ vs. [jə 'zy 11 131] 搭船 ‘to row a boat’.
[ju 'dzəŋ 11 41] 油城 ‘oil town’ vs. [ju 'dzəŋ 11 131] 有城 ‘there is a town’.

2) with intervocalic sonorant:

[wəi 'lε 11 41] 回来 ‘to come back’ vs. [wəi 'lε 11 131] 会来 ‘able to come’.

3) with intervocalic glide and zero consonant (i.e. second syllable is GV(η)):

[dζəŋ 'jıŋ 11 41] 俗形 ‘circumstances’ vs. [dζəŋ 'jıŋ 11 131] 俗形 ‘to take shape’.

Note in the above examples that voice quality correlates with pitch onset height - syllables with high pitch onset [41] have modal voice; syllables with low pitch onset [11,131] have whispery voice [..,] (for this compound phonation type see Laver 1980). Pitch onset height also correlates with manner of syllable initial obstruent: voiceless lenis [dz z ʰ etc.] with low pitch; normal voiced [dz z b etc.] with high. In all examples, the second syllable is prominent.

The surface contrast between falling and convex pitches after a low level pitch is pertinent to this paper, because it involves real data from a tone language which present a problem for the conception of pitch as a straightforward function of F0. In disyllabic utterances with level-convex pitch we would reasonly expect the F0 to show a corresponding low level trace followed by a rise-fall contour. However, in disyllabic utterances with level-falling pitch and a voiced intervocalic segment (as in the examples above), the F0 must change from low level on the first syllable to high (falling) on the second syllable, but it must also be uninterrupted because of the voiced intervocalic segment. (Or almost uninterrupted - phonation often ceases for a short time during the hold phase of a voiced stop for well-known aerodynamic reasons.) Therefore the F0 time course in level-
falling utterances must incorporate a rise somewhere as it moves from the low level on the first syllable to the high fall on the second syllable. Yet no rising pitch is audible in these examples to correspond to this inevitable rise in F0. If it were, it might well cause confusion with the examples with level-convex pitch.

Below I shall present an acoustic analysis of this contrast, to try to establish what possible factors are causing the rise in F0 not to be perceived as a rise in pitch. I shall examine the acoustic correlates of the contrast under the three conditions illustrated above, namely on disyllables with intervocalic syllable-initial voiced/voiceless lenis obstruent, intervocalic syllable-initial sonorant, and intervocalic glide.

### 3.1 Procedure

Data from one male Shanghai-Zhenhai native speaker were used, recorded on professional equipment in the Manchester University Phonetics Laboratory. (The recording level was set manually for maximally undistorted amplitude registration.) To examine the contrast on disyllables with intervocalic consonant, I compiled a corpus of isolated tokens from previously recorded material, controlled for intrinsic F0 and Ar on the second syllable. The corpus comprised 15 examples of VCobsV(41) with 11 plosives and 4 affricates; 10 examples of VCobsV(131) with 5 plosives and 2 affricates; 6 examples of VCsonV(41), and 10 of VCsonV(131).

To investigate examples with the intervocalic glide was a problem. With these examples, it is important to have the same segmental sequence, at least on the second syllable, otherwise between-token formant values might tend to cancel each other out when means are calculated. Unfortunately, there were no minimal pairs, or near minimal pairs in my previously recorded data and the informant was also no longer available to record some. I therefore had to compromise by contrasting a set of 8 tokens of the same VGV word [pʰ̚ju 11 41] 'friend' with 6 tokens of a segmentally comparable monosyllabic GV word [jju 131] 'oil'. One of the /b̚jʊ/ tokens was recorded in isolation, and the remaining seven were edited from running speech, where they occurred in various stressed positions, e.g. [ŋɔyə b̚jʊ zəŋi ˈləχa] 'me' '格' '日来' '格' '格' 'My friend came yesterday'; [zəŋi ju ɪə b̚jʊ səŋ pəŋ ˈŋu i ˈmʊŋ cɪŋa cɪʒaʔ] '日' '日有一格朋友送' '我一本新' '格' '小说' 'Yesterday my friend gave me a new novel.' The six [jju] 'oil' tokens were all recorded in isolation.

The measurement procedure was designed to obtain detailed information on the mean time course of F0 and Ar over the two syllables, and on their relationship to the major acoustic boundaries (i.e. the boundaries separating the intervocalic consonant from the two vowels).

#### 3.1.1 Examples with intervocalic consonant

Tokens with intervocalic consonant were first segmented into first syllable vowel, intervocalic consonant, and second syllable vowel using wide band (300 Hz) spectrograms with their good time domain resolution. The duration of these three segments was measured, and then F0 and Ar sampled at various percentage points of the segments' duration with a sufficiently high sampling rate to resolve the details of their time course. F0 was measured from narrow band (45 Hz)
spectrograms (accuracy ± 4 Hz at 90% confidence level). Ar was measured from ‘average amplitude’ spectrograms (wide band, flat, effective full wave rectification; accuracy ± 0.5 dB at 90% confidence level). The accurate alignment of F0, Ar and segmental boundaries was achieved by reference to click transients, imposed for this purpose during the process of editing from the master tapes onto the analysis tape.

The duration of the first syllable vowel was taken to extend from onset of its phonation to the beginning of the hold phase of the intervocalic sonorant or stop as shown by the abrupt drop in amplitude of (higher) frequencies. (The beginning of the only two intervocalic alveolar fricative tokens was taken to be the drop in average amplitude and the onset of their typical high frequency aperiodicity.)

The duration of the second syllable vowel was taken to extend from the release of consonantal stricture, as shown by onset of higher amplitude periodicity with clear formant structure, to a point of irregularity in the increase of fundamental period which, for this speaker at least, served as a regular and reliable indication of phonation offset in syllables with falling F0 (Rose 1982b:105).

The duration of the intervocalic consonant, whether plosive, sonorant, affricate, or fricative, was taken to include everything from the offset of the first syllable to the onset of the vowel of the second syllable. In all cases, the points defining segmental boundaries involved clear acoustic discontinuities and could be located to within one or at most two glottal pulses.

Arithmetical means and standard deviations for F0, Ar, and duration in disyllables with intervocalic consonant are given in Table 1. Figure 1 shows the mean F0 and Ar values plotted as functions of absolute duration, as well as their relationship to major segment boundaries.

3.2 Examples with intervocalic glide

In the [bā ju] examples with the intervocalic glide, segmentation cannot be performed in the same way as in the disyllables with intervocalic consonant. This is because there are no spectral discontinuities corresponding to the boundaries of the auditory segments of the [ā] - [j] - [u] sequence. In these cases, then, the F0 time course was treated as a spline function consisting of three parts, and Ar and formant (F-) pattern (1st, 2nd and 3rd formants - the primary determinants of vocoid quality) sampled at 20 per cent points of the duration of these parts. The three parts were: onset of F0 to start of F0 rise; start of F0 rise to F0 peak; F0 peak to F0 offset. The F-pattern was measured from wide band contour spectrograms: their relatively large dynamic range (36 dB) was useful for resolving low level formants, like F3 in [u].

The perceptual boundaries of the first and second syllable in [bā ju] were determined with a truncation method similar to that used by Fujisaki (1977:360-361). Increasingly longer portions of the word were deleted from both ends until either only the [ā] or the [ju] was audible. Spectrograms of the remaining portions were made, and the perceptual boundaries of [ā] and [ju] measured from them.
Table 1

Mean and standard deviation values for Fundamental Frequency, Amplitude, and Duration in 10 tokens of [11 131] with intervocalic sonorant (A); 6 tokens of [11 41] with intervocalic sonorant (B); 10 tokens of [11 131] with intervocalic obstruent (C), and 15 tokens of [11 41] with intervocalic obstruent (D).

<table>
<thead>
<tr>
<th>Sampling point (% of duration)</th>
<th>F0(Hz)</th>
<th>Ar(dB)</th>
<th>F0(Hz)</th>
<th>Ar(dB)</th>
<th>Sampling point (% of duration)</th>
<th>F0(Hz)</th>
<th>Ar(dB)</th>
<th>F0(Hz)</th>
<th>Ar(dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>119 11</td>
<td>23.4</td>
<td>4.5</td>
<td>117 7</td>
<td>15.0</td>
<td>3.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>25.5</td>
<td>3.8</td>
<td></td>
<td>17.1</td>
<td>3.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>112 6</td>
<td>25.8</td>
<td>3.9</td>
<td>112 5</td>
<td>17.3</td>
<td>3.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(A) 40</td>
<td>110 4</td>
<td>24.8</td>
<td>2.9</td>
<td>(B)</td>
<td>112 4</td>
<td>17.2</td>
<td>3.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>110 3</td>
<td>24.7</td>
<td>2.6</td>
<td>114 5</td>
<td>17.5</td>
<td>3.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>110 3</td>
<td>23.9</td>
<td>2.7</td>
<td>114 5</td>
<td>18.0</td>
<td>3.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>95</td>
<td></td>
<td>23.1</td>
<td>2.9</td>
<td></td>
<td>18.4</td>
<td>3.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>109 4</td>
<td>22.4</td>
<td>3.7</td>
<td>118 7</td>
<td>19.0</td>
<td>3.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(C) 40</td>
<td>111 5</td>
<td>23.8</td>
<td>2.8</td>
<td>128 6</td>
<td>20.3</td>
<td>4.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>113 4</td>
<td>26.5</td>
<td>2.1</td>
<td>143 8</td>
<td>23.2</td>
<td>3.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>28.8</td>
<td>2.7</td>
<td></td>
<td>25.8</td>
<td>3.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>122 6</td>
<td>30.5</td>
<td>2.8</td>
<td>156 10</td>
<td>26.6</td>
<td>3.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td></td>
<td></td>
<td></td>
<td>159 11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>139 8</td>
<td>31.7</td>
<td>1.6</td>
<td>157 12</td>
<td>25.6</td>
<td>2.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>150 12</td>
<td>31.7</td>
<td>1.4</td>
<td>150 11</td>
<td>24.6</td>
<td>3.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>143 14</td>
<td>30.2</td>
<td>1.6</td>
<td>136 9</td>
<td>23.8</td>
<td>4.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>110 10</td>
<td>23.1</td>
<td>2.4</td>
<td>117 5</td>
<td>17.2</td>
<td>4.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration (csec)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First vowel</td>
<td>32.1</td>
<td>8.5</td>
<td></td>
<td></td>
<td>34.7</td>
<td>10.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intervocalic sonorant</td>
<td>9.1</td>
<td>1.3</td>
<td></td>
<td></td>
<td>8.5</td>
<td>2.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second vowel</td>
<td>26.8</td>
<td>3.4</td>
<td></td>
<td></td>
<td>22.3</td>
<td>3.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intervocalic obstruent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9.3</td>
<td>4.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration (csec)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>36.1</td>
<td>9.6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 1.
Mean acoustic characteristics of the level-falling vs. level-convex contrast on disyllables with intervocalic consonant. A = [- son \lya]; B = [- son \lyb]; C = [-obs \lyc]; D = [- obs \lyd]. For each example the top trace is amplitude, the bottom fundamental frequency. Interrupted lines indicate values during intervocalic consonant.
The F0, Ar and F-pattern of the [ju] syllables were sampled at 10 per cent points of the duration of the syllable.

Arithmetical means and standard deviations of the F0, Ar, F-pattern, and duration of the [bāju] and [Ju] tokens are given in Tables 2 and 3 respectively. (Table 2 also gives the value of the perceptual boundaries in [bājul].) Their mean F0, Ar and first three formants are plotted as functions of absolute duration in Figures 2 and 3.

Table 2

Mean and standard deviation values for Fundamental Frequency, Amplitude, Formant Pattern and Duration in 8 tokens of [bāju 11 41].

<table>
<thead>
<tr>
<th>sampling point (% of duration)</th>
<th>F0(Hz)</th>
<th>Ar(dB)</th>
<th>F1(KHz)</th>
<th>F2(KHz)</th>
<th>F3(KHz)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\bar{x}$</td>
<td>$s$</td>
<td>$\bar{x}$</td>
<td>$s$</td>
<td>$\bar{x}$</td>
</tr>
<tr>
<td>0</td>
<td>117</td>
<td>8</td>
<td>26.0</td>
<td>6.5</td>
<td>0.74</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>-</td>
<td>28.5</td>
<td>3.8</td>
<td>-</td>
</tr>
<tr>
<td>20</td>
<td>108</td>
<td>7</td>
<td>25.7</td>
<td>3.2</td>
<td>0.79</td>
</tr>
<tr>
<td>40</td>
<td>107</td>
<td>7</td>
<td>23.4</td>
<td>2.8</td>
<td>0.78</td>
</tr>
<tr>
<td>60</td>
<td>106</td>
<td>7</td>
<td>22.1</td>
<td>2.8</td>
<td>0.77</td>
</tr>
<tr>
<td>80</td>
<td>106</td>
<td>8</td>
<td>21.7</td>
<td>2.7</td>
<td>0.75</td>
</tr>
<tr>
<td>0/100</td>
<td>106</td>
<td>7</td>
<td>22.0</td>
<td>2.6</td>
<td>0.64</td>
</tr>
<tr>
<td>20</td>
<td>111</td>
<td>8</td>
<td>22.8</td>
<td>2.2</td>
<td>0.46</td>
</tr>
<tr>
<td>40</td>
<td>123</td>
<td>10</td>
<td>24.6</td>
<td>3.2</td>
<td>0.37</td>
</tr>
<tr>
<td>60</td>
<td>137</td>
<td>14</td>
<td>26.7</td>
<td>2.7</td>
<td>0.35</td>
</tr>
<tr>
<td>80</td>
<td>147</td>
<td>18</td>
<td>29.4</td>
<td>1.8</td>
<td>0.35</td>
</tr>
<tr>
<td>0/100</td>
<td>150</td>
<td>18</td>
<td>29.4</td>
<td>1.4</td>
<td>0.36</td>
</tr>
<tr>
<td>20</td>
<td>150</td>
<td>18</td>
<td>29.3</td>
<td>1.3</td>
<td>0.36</td>
</tr>
<tr>
<td>40</td>
<td>146</td>
<td>16</td>
<td>28.1</td>
<td>1.2</td>
<td>0.38</td>
</tr>
<tr>
<td>60</td>
<td>140</td>
<td>16</td>
<td>26.0</td>
<td>1.4</td>
<td>0.40</td>
</tr>
<tr>
<td>80</td>
<td>131</td>
<td>14</td>
<td>24.3</td>
<td>2.3</td>
<td>0.36</td>
</tr>
<tr>
<td>100</td>
<td>121</td>
<td>14</td>
<td>22.1</td>
<td>2.6</td>
<td>0.37</td>
</tr>
</tbody>
</table>

duration (csec) $\bar{x}$ $s$
from onset to start of F0 rise $= 23.3$ $8.6$
from start of F0 rise to F0 peak $= 17.1$ $2.0$
from F0 peak to offset $= 9.5$ $3.6$
total (first vowel plus second syllable) $= 49.8$ $11.4$

perceptual boundary (Csec from onset) $\bar{x}$ $s$
end of [ā] $= 22.2$ $6.1$
start of [ju] $= 28.8$ $7.7$
On the non-equivalence of fundamental frequency 67

**Fig. 2.**
Mean acoustic characteristics of [bäju – ∨] disyllables with intervocalic glide and level-falling pitch.

**Fig. 3**
Mean acoustic characteristics of [ju ∧] syllables with initial glide and falling pitch.
Table 3

Mean and standard deviation values for Fundamental Frequency, Amplitude, Formant Pattern and Duration in 6 tokens of [ju 131]

<table>
<thead>
<tr>
<th>sampling point (%) of duration</th>
<th>F0(Hz)</th>
<th>Ar(dB)</th>
<th>F1(KHz)</th>
<th>F2(KHz)</th>
<th>F3(KHz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>128</td>
<td>12</td>
<td>15.6</td>
<td>3.7</td>
<td>0.33</td>
</tr>
<tr>
<td>10</td>
<td>129</td>
<td>11</td>
<td>22.3</td>
<td>2.4</td>
<td>0.32</td>
</tr>
<tr>
<td>20</td>
<td>136</td>
<td>10</td>
<td>26.9</td>
<td>3.0</td>
<td>0.30</td>
</tr>
<tr>
<td>30</td>
<td>145</td>
<td>13</td>
<td>29.2</td>
<td>2.4</td>
<td>0.31</td>
</tr>
<tr>
<td>40</td>
<td>153</td>
<td>15</td>
<td>29.5</td>
<td>1.6</td>
<td>0.34</td>
</tr>
<tr>
<td>50</td>
<td>159</td>
<td>15</td>
<td>29.7</td>
<td>1.8</td>
<td>0.36</td>
</tr>
<tr>
<td>60</td>
<td>162</td>
<td>16</td>
<td>29.6</td>
<td>1.5</td>
<td>0.36</td>
</tr>
<tr>
<td>70</td>
<td>159</td>
<td>16</td>
<td>29.7</td>
<td>1.8</td>
<td>0.36</td>
</tr>
<tr>
<td>80</td>
<td>150</td>
<td>15</td>
<td>29.3</td>
<td>2.3</td>
<td>0.36</td>
</tr>
<tr>
<td>90</td>
<td>138</td>
<td>16</td>
<td>26.7</td>
<td>2.0</td>
<td>0.36</td>
</tr>
<tr>
<td>100</td>
<td>121</td>
<td>12</td>
<td>20.8</td>
<td>2.0</td>
<td>0.35</td>
</tr>
</tbody>
</table>

duration (csec) \( \bar{x} = 33.5 \quad s = 3.4 \)

3.3 Results and discussion

It can be seen from Figures 1 to 3 that all of the examples analysed show considerable rises in F0 at some point during the utterance. Before addressing the main question of why the F0 rises in level-falling examples are not perceived as a rise in pitch, two short observations are necessary.

With one exception to be discussed below, the acoustic data are homogeneous and unremarkable. Items with the same pitch (i.e. [11] [41] [131]) have statistically the same acoustics, once allowance is made for the effect on F0 and Ar of the intervocalic consonant, and the assimilatory effect of the second syllable on the first. (The gradually rising F0 on the first syllable in level-falling examples probably reflects anticipatory assimilation in longitudinal vocal cord tension and/or phonation type. Lack of control for intrinsic Ar on the first syllable has probably obscured this effect on the Ar profiles, though it is fairly clear on the example with the sonorant intervocalic consonant.)

The second observation is that the obstruents transcribed above as voiceless lenis show periodicity (i.e. F0) throughout their hold (Figure 1c). This is surprising, since periodicity normally signals the presence of the feature [voice]. One factor possibly contributing to the auditory percept of voicelessness for these obstruents is the large (-9 dB) concomitant drop in Ar (cf. the much smaller drop of -2 dB on the auditorily voiced obstruent in figure 1d). It should be noted that the auditory difference between these two sets of obstruents is an important junctural feature of Zhenhai, signalling the absence or presence of a preceding word boundary.
Figure 4 (above)
Mean fundamental frequency of level-falling (top) and level-convex (bottom) disyllables.

Figure 5 (top right).
Mean acoustic characteristics of the Pakphanang high convex tone.

Figure 6 (bottom right).
Normalised fundamental frequency contours of Pakphanang high convex tone (---) and Zhenhai falling tone after a voiced consonant (—).
The above discrepancy is worth commenting on for two reasons. Firstly, it is an example of the many-to-one relationship between acoustics and auditory features noted above: here, presumably, the perception of [-voice] is dependent on sufficient Ar level as well as F0. Secondly, it serves as a reminder of the inferential nature of the articulatory features with which we describe the auditory responses encoded in a phonetic transcription, and of the fallibility of the process: just because we hear [-voice] does not imply that the vocal cords are not vibrating.

To return now to the main problem: if pitch is a function of F0, why are the F0 rises in level-falling examples not perceived as pitch? A glance at Figure 4, which shows just the mean (sonorant and obstruent) F0 shapes of the two contrasting types, suggests that F0 information alone would probably not be enough to explain the differential perception (although the F0 time courses are contrastive enough, for example in maximum F0 height, to distinguish the two types). Results from experiments in the perception of F0 contours also indicate this. Both F0 rises should be easily perceivable because they are well above the 4 - 6.5 csec duration threshold for the perception of contouricity in rising F0 (Greenberg and Zee 1979:154), and perceived contouricity of both F0 rises will be enhanced by the presence of the preceding level F0 (Greenberg and Zee 1979:156). The only possible explanation for the differential perception in F0 terms alone lies in the difference in absolute rate of F0 rise, which is higher (2.8 Hz/csec) for the ‘non-perceived’ rise than for the perceived (1.9 Hz/csec). This difference is well above the 0.12 - 0.32 Hz/csec threshold for the perception of differences in the positive F0 derivative (Klatt 1973:12). But it is still so small that for it to function as the basis for the ‘non-perception’ of the greater slope as pitch would require a categorical perception effect, which is contentious for tone (Gandour 1978:58ff). Also the difference in F0 slope would be perceptually attenuated by the fact that perceived contouricity is directly proportional to the duration of the preceding level F0 (Greenberg and Zee 1979:156). In any case, these speculations are rendered unnecessary by the existence in the informant’s speech of rising pitched tones with F0 rise values in excess of a putative categorical boundary around, say, 2.5 Hz/csec. Mean values for his citation tone 6 for example (measured from tables in Rose 1982a:18) are 3.6 and 3.4 Hz/csec. Mean values for his tone 6 in a comparable syntactic position (following a short low tone in lexical sandhi) are even bigger: 5 - 6 Hz/csec (Rose 1985). It is clear then that we have to adduce data in addition to F0 values to explain the difference in the perception of the F0 rise.

One obvious candidate is Ar. It will be recalled that the pitch of speech has been shown to be a function of both F0 and Ar. Moreover, previous studies on Zhenhai have shown that Ar is differentially distributed with respect to F0 in some tones. In particular, the rising part of the convex tone is characterised by a high ratio of Ar to its F0. It is of interest, therefore, to see if Ar is in some way involved with the differential perception of F0. Do, for example, the different pitch percepts correspond to any large difference in the distribution of Ar with respect to F0, or is the F0 not perceived as a pitch rise associated with considerably less Ar than the F0 rise perceived as such?

There are several points to be noted in trying to assess the role of Ar to the perception of F0 rise in these data. Firstly, the raw Ar values are not directly comparable across examples, so it is not possible to demonstrate differences simply by regressing Ar on F0. Secondly, the Ar is difficult to evaluate because it is not possible to separate out all the factors involved in its production. (This is
essential because the speech specific perceptual mechanism judges Ar in terms of physiological effort required to produce it, in particular the Ps involved (Ladefoged 1967:35-41). Thus, although the effect of the supraglottal impedance on Ar can be discounted because it will be the same for items with the same segmental sequence, it is not possible to tell, for example, whether the large drop in Ar over the obstruent in level-convex examples reflects Ps or a change in glottal configuration, or both. Also, in examples with level-falling pitch, the Ar may very well incorporate the effect of an additional factor of change in glottal configuration from whispery to modal voicing. Because of these difficulties, observations will be confined to examples with intervocalic sonorant, which appear to show less drastic Ar and F0 perturbations. The inferences drawn from them are of course necessarily fairly tentative.

There seems to be a clear differential distribution of F0 with respect to Ar over stretches of F0 rise where the effect of supraglottal impedance on the Ar is constant. During the sonorant and up to about 5 csec after release (that is, for most of the F0 rise on the level-falling examples) both level-falling and level-convex examples have nearly identical Ar rise values (0.56 (convex) and 0.59 (falling) dB/csec), but the rate of F0 rise in the level-falling example (2.93 Hz/csec) is about 3 times that in the level-convex example (0.89 Hz/csec). Thus, as expected, there is about 3 times more Ar to F0 in the first part of the convex pitched syllable than the falling.8

However, it is unfortunately not possible to actually quantify this differential, and its perceptual effect, over the F0 rises of interest, because in these cases the supraglottal effect on Ar is no longer constant (we would be illegitimately comparing ratios of Ar rise to F0 rise during the hold phase in level-falling examples with those after release on the level-convex).

In sum, it looks as if there is a differential amount of Ar to F0 in the two examples, and this probably reflects differences between them in timing and degree of involvement of the mechanisms controlling F0 and Ar. In the level-convex examples, the rise in F0 and Ar (and the relatively high ratio of Ar to F0) is probably caused by a synchronous Ps burst; in the level-falling examples, the rise in F0 and Ar (and the relatively low ratio of Ar to F0) is probably the result of a combination of an increase in VCT and a slightly later, smaller Ps burst. It is doubtful, however, whether this difference in source control could be responsible for a difference in the perception of the F0 rise (though it probably serves to contrast the two types). Firstly, the difference is not astoundingly large: both Ar and F0 do, after all, rise concurrently and steeply in both examples. Secondly, for the perceptual mechanism to ignore the F0 rise on level-falling examples on the basis of Ar information would be equivalent to a specific instruction to ignore rises not produced by Ps increases, which seems highly unlikely.

The next, again fairly obvious approach is to allow for the possibility of pitch perception of F0 mediated by segmental information. It is possible that the timing relationship between laryngeal and supralaryngeal gestures is exploited perceptually by timing those F0 transitions not to be registered as pitch to occur at points defined with respect to the supralaryngeal gestures which provide cues to segmental structure. The perceptual mechanism could then be programmed not to register as pitch F0 changes that take place over, or in the vicinity of, certain segments in certain structural positions - e.g. F0 rises on syllable-initial sonorants or voiced obstruents. Note that the restriction to certain positions in the syllable is necessary because velar nasal sonorants constituting the syllable coda do quite clearly carry a tonally relevant part of the F0 contour (Rose 1982a: 33).
Voiced syllable initial consonants seem likely candidates for the role of segmental maskers. It has been hypothesised that in several Chinese dialects the F0 on a syllable initial sonorant is irrelevant for the tone. This is because much greater congruence exists between F0 contours of the same tone on syllables differing in the [± sonorant] feature of the initial consonant if the sonorant portion is ignored (Rose 1982a:47, 48). Another justification for disregarding the F0 on the sonorant lies in its lack of association with a pitch percept. In Zhenhai, for example, syllables with initial voiceless obstruent (e.g. [ȵʔ214] 擺 ‘lay out’) have the same pitch as syllables with initial sonorant in the same tone (e.g. [mq?214] 買 ‘buy’). If the F0 on the sonorant were tonally relevant, one might expect this to be reflected in a different pitch or even length percept.

Figure 1 shows that the F0 is, in fact, differentially distributed with respect to the main acoustic segments. Most of the F0 rise perceived as pitch occurs after the segment boundary, i.e. on the vowel, whereas some of the F0 rise not perceived as pitch takes place during the consonantal segment, i.e. before the vowel.

However, several considerations argue against the ‘consonantal masking hypothesis’. Firstly, the masking effect also occurs in level-falling examples without syllable initial consonants. In tokens with intervocalic glide, (like the [bá ju] examples to be examined below), or where the second syllable consists of a vowel only, e.g. [njɑ ·ǎ 11 51] 銀行 bank’, [gɑ ·i 11 51] 大衣 ‘overcoat’, no pitch rise is audible either.

Secondly, and more obviously, not all the F0 rise occurs on the consonant: over a third of the rise (40 per cent in sonorant examples; 36 per cent in obstruents) occurs on the following vowel, giving it a convex F0 contour. This convex shape differs from that on the speaker’s falling pitched syllables with sonorant initial consonant after a pause. (These show only a small (about 4 Hz) perturbatory rise for the first few (3) csec of the vowel.) It seems, therefore, that the large F0 onset perturbation on the second syllable vowel is caused by the transition rather than the voiced syllable initial consonant.

Finally, there are some indications that the convex F0 shape on the vowel should be enough to cue a convex, rather than a falling pitch. Figure 5 shows typical F0 and Ar values for the high convex pitch tone in the Southern Thai dialect of Pakphanang (Rose forthcoming). The speaker was female, hence the high F0 values; the measurements are arithmetical means of 3 x [ka] + 3 x [ku], and the tone has a clearly audible initial pitch rise before the fall. (This was corroborated by another professional phonetician.) The degree of similarity involved is shown in Figure 6, which plots the F0 values of the two tones (Pakphanang high convex; Zhenhai high fall) normalised with respect to the individuals’ F0 range for that particular context. It can be seen that the amount of F0 rise from onset to peak is in fact 10 per cent greater for the Zhenhai tone than for the Thai, and that there is therefore not enough difference between the two normalised F0 curves to explain the different pitches involved. The Chinese and Thai data seem to indicate the possibility that the same F0 contour on a vowel can cue two different pitches.

There are also several reasons why it is not possible simply to assume that the masking effect lasts until about 5 - 7 csec after the consonant release, (and therefore masks the whole of the ‘non perceived’ F0 rise). Firstly, as already noted, the same effect occurs even in the absence of consonants. Furthermore, one of the Zhenhai rising pitched tones is of typically very short
duration - for this speaker in citation form, between about 10 and 13 csec. Ignoring 5-7 csec of the F0 contour after release on this tone would leave very little to be perceived as pitch, especially when the duration threshold of 4-6 csec for perception of pitch control is also to be allowed for. Yet this tone has a clear rising pitch. Finally, there is the evidence from the Chinese and Thai data mentioned above. In some cases (as for example in Pakphanang convex tones with sonorant initial consonant) the 5-7 csec after release seems to include a pitch-relevant F0 rise which we would not want to be ignored; in others (as for example in Zhenhai falling tones with voiced initial consonant) the first 5-7 csec of F0 needs to be ignored from the point of view of the pitch percept.

We have now looked at several unsuccessful attempts to explain why the F0 rise in level-falling examples may not be registered as pitch. This inadequacy may have come from the assumption that a difference in perception must correspond to a difference in acoustics. However, as has already been amply illustrated in the first part of this paper, a lack of invariance between acoustics and perceptual units is another commonplace in the complexity of speech perception (Studdert-Kennedy 1974:2356ff). It seems therefore that a higher level, possibly language specific explanation is to be sought.

Such an explanation might involve distinguishing between intrinsic and extrinsic effects. The 'non-perceived' F0 rise in the level-falling example occurs because the F0 targets on the first and second syllables differ in height, and the F0 has to change from low to high. Also, since the intervocalic consonant is voiced, this change is continuous, giving an F0 rise. Because of an inertial effect, the F0 rise 'spills' over into the following vowel, and together with the falling F0 target on the second syllable produces a convex F0 on the second syllable vowel.

Thus it is clear that the F0 rise in level-falling examples is intrinsic: that is, not deliberately produced, but occurring as the result of some other, deliberate gestures (F0 targets; voicing on intervocalic consonant). It can be noted here that an intrinsic effect of the same magnitude appears to occur in some varieties of English. The resulting convex F0 shape on the falling pitched second syllable is very similar to those reported in Lea (1973) on English stressed vowels in haCV(C) utterances with sonorant or voiced obstruent consonant. (But cf. the very different F0 perturbations in Silverman (1984)). Thus the hypothesised minimising of intrinsic F0 effects in tone languages (see above) does not seem borne out by the Zhenhai data.

In contrast to the F0 rise in the level-falling examples, the F0 rise on the level-convex examples is not intrinsic, but part of the deliberately produced target F0 on the second syllable. It seems reasonable, and certainly in keeping with the idea of speech perception mediated by production, to assume that the perceptual mechanism recognises the F0 changes as intrinsic or extrinsic on the basis of their temporal relationship to segmental structure, and ignores the intrinsic changes as possible pitch cues. Thus the convex F0 on the Pakphanang high convex tone is perceived as convex because it is deliberately produced, whereas the convex onset to the F0 contour on the Zhenhai second syllable vowel is not so perceived because it is an intrinsic effect.

The hypothesis may be invoked to explain other examples of discrepancies between F0 and pitch. Rapid falls in F0, which are not perceived as pitch, occur as intrinsic perturbations caused by a syllable-final glottal stop in some varieties of Chinese (Zee and Maddieson 1980; Rose
1982b:206, 322, 360). In Thai Phake citation forms, considerable convex perturbations occur at the end of an F0 contour which are not registered as pitch. (In the case of the falling pitched tones the F0 sometimes actually rises higher at the end than its onset value.) Simultaneous air flow measurements indicate these perturbations are intrinsically caused by syllable final laryngeal adjustments (Rose MS 1986).

The ‘intrinsic’ hypothesis raises the problem of what happens in a tone language which contrasts high falling and high convex pitched tones after a low level. Would a voiced intervocalic consonant cause the high falling F0 to become convex and thus be indistinguishable from the high convex tone? In such a case, the contrastivity between types could be assured by a higher degree of coarticulation resistance (Nolan 1983:116-120) minimising the inertial effect, and resulting in all of the F0 rise in the hypothetical level-falling example occurring during the consonant. For this to be plausible, it would, of course, have to be demonstrated that the inertial effect was of the controllable-intrinsic type (Tatham 1971:143).

Although the Zhenhai data appear to confirm this - the rate of 5 - 6 Hz/csec quoted above for the speaker's tone six would produce the required rise of 40 - 50 Hz within the 8.5 csec consonant hold - it is possible that the F0 rise values in the level-falling and tone six examples are not strictly comparable. This is because they may be produced in different ways, with the high rate on tone 6 being caused partly by a Ps boost, as opposed to vocal cord tension on the level-falling examples (Rose 1984:164).

Possibly a more interesting consequence of this ‘intrinsic’ hypothesis concerns current speculations on tonogenesis. According to Hombert Ohala and Ewan (1979) tonogenesis necessarily involves reinterpretation of an intrinsically caused consonantal perturbation as extrinsic. Suppose, for example, the speaker of a non tone language intends to say a voiced consonant followed by a vowel. The voiced consonant intrinsically induces a low perturbation of the F0 on the vowel. The listener, however, who cannot read the speaker's mind (1979:37) cannot tell what is intrinsic in the signal, and what is deliberate. He reinterprets the F0 perturbation as extrinsically produced, and begins to produce it deliberately, and exaggerate it. With the devoicing of the initial voiced consonant, a low pitched vowel has now arisen which will contrast minimally with a high pitched vowel after an original voiceless consonant, and the transition to a tone language is complete.

The hypothesis suggested in this paper presents some problems for this scenario. Firstly, it is difficult to explain the phenomenon of differential F0 perception without assuming that listeners can in fact tell intrinsic from extrinsic effects by dint of their linguistic competence. Secondly, the hypothesis that intrinsic F0 effects have to be ignored as pitch is in direct conflict with the tonogenesis proposal that they are cued-into as pitch, unless one assumes that this is another way in which tone languages differ from non tone languages, or that it is possible to distinguish between intrinsic F0 perturbations caused by an initial consonant and those which are part of an intrinsic transition (see above). (Note that intrinsic F0 perturbations must be registered somehow, because they can act as secondary cues for the [voiced] and [± obstruent] feature on consonants (Abramson and Lisker 1985; Chistovich 1969).) It is not possible to resolve these problems here, but at least they suggest that tonogenesis is perhaps not so ‘rather well understood’ as claimed (Hombert 1977:9).
3.4 Examples with intervocalic glide

The analysis above has hypothesised that the temporal relationship of the F0 rise to the syllable-initial consonant is of importance in signalling how the F0 is to be perceived. What happens when there is no syllable-initial consonant, however? This is shown in the examples of the contrast on syllables with a glide but no initial consonant, the acoustic characteristics of which are given in Tables 2 and 3, and Figures 2 and 3.

Figures 2 and 3 show that, apart from small differences in overall level and offset value - which are probably functions of the different contexts from which they were taken - the glide examples have basically the same F0 shape as their counterparts with intervocalic consonant (henceforth VCV). Note in particular that the amount of F0 rise remains approximately constant: 44 (glide) and 42 Hz (VCV) in level-falling; 34 (glide) and 37 Hz (VCV) in (level)-convex. Also the rate of F0 rise is fairly constant for a given pitch: 2.9 Hz/csec for level-falling and 2.4 Hz/csec for convex (cf. 2.8 and 1.9 Hz/csec for the VCV examples). Thus, just as for the VCV examples, the rate of F0 rise is higher in the level-falling glide examples, but only very slightly so, and not enough to serve as a basis for a pitch difference on its own.

Although, as pointed out above, the time course of the F-pattern is continuous, the presence of simultaneous changes in the frequency derivative of two or three formants allows us to divide the utterance into three sections. The first portion (onset - csec 18.5) is characterised by static F1 and F2 values similar to those for the informant's [a] vowel, and F3 appears to rise.10 In the second portion (csec 18.5 - csec 30.0) the formants diverge to values typical of the speaker's [i]. The perceptual onset of [ju] is located at csec 29, which agrees very well with this acoustically defined point. The second syllable can therefore be considered to have begun, both acoustically and perceptually, by this '[j] onset point'. In the last part (csec 30 - offset) the formants move to values typical for his [u]. This last part also contains a clear landmark, at approximately csec 37, where F2 and F3 come maximally close. This partially reflects the intersection of the second and third resonances (as opposed to formants) which occurred in some examples.

The F-pattern on [ju] is simply a linearly expanded version of that part of the F-pattern in [bā ju] which starts at about 3.5 csec after the [j] onset. It looks therefore as if the supralaryngeal organs are programmed to realise the same acoustics, only quicker in the falling pitch example than in the convex.

Just as with the VCV examples, it is not possible with the glide examples to explain the difference in perception of F0 rise totally in terms of concurrent F-pattern features. The F0 rise is differentially distributed with respect to the F-pattern, to be sure, but only just over a third (39 per cent) of the 44 Hz F0 rise in the level-falling items is distributed over the F-pattern transition from [a] to [j] onset; the rest of the rise takes up no less than half of the remainder of the falling pitched syllable, resulting in a clearly convex F0 for the segmental [ju] syllable. Of this 27 Hz rise, the first 14 Hz might possibly be discounted because the concurrent F-pattern is still not congruent with that in the [ju] example, although the perceptual and acoustic onset to the second syllable suggest that this rise should be included. But the subsequent rise of 13 Hz does take place over comparable portions of the F-pattern, and in fact the remaining F0 shape resembles very closely
the convex F0 shape on the second syllable vowel of level-falling examples, and should theoretically still be perceived as convex on the basis of the comparison with the Pakphanang data.

It is interesting to note that, from the point of view of its temporal relationship to the F0, the [j] onset corresponds neither to second syllable onset in VCV examples, nor to second vowel onset, but to a point somewhere in the middle of the syllable-initial consonant. A segment of comparable duration to the intervocalic consonant in VCV examples centered around this [j] point would cover the same portion of F0 rise as covered by the consonant in VCV examples.

As with VCV examples, there is a clear differential distribution of F0 with respect to Ar over stretches of rising F0 where the supraglottal impedance can be considered constant. (As an indication of this last factor I used a stable F1, which contributes most to overall Ar (House 1967:59). The F0 rises over which this differential distribution can be demonstrated were over about the first 7 csec of [ju], and from csec 30.1 to csec 37 in [bā ju]). Over these stretches, the F0 rates are 3.5 Hz/csec for level-falling, and 1.2 Hz/csec for convex, and the Ar rates are 0.7 dB/csec for level-falling and 1.7 dB/csec for convex. So the ratio of F0 rise to Ar rise is again considerably greater for the convex than for the falling example.11

As with the VCV examples, this differential distribution of F0 and Ar can be assumed to reflect some difference in source control, but it is still not clear that this would contribute to a different perception of the F0 rise.

3.5 Summary

Allowing for differences due to the difference in contexts from which they were edited, the glide examples present the same problems as the VCV examples in accounting for the differential perception of F0 rises. Again, it seems that the acoustics alone are not enough to explain the perceptual difference. An additional interpretive principle concerning the temporal relationship between F0 and supralaryngeal patterns has to be adduced, according to which intrinsic F0 transitions and their associated F0 perturbations are ignored as pitch cues. These results highlight the importance of segmental information for the interpretation of F0.

For these Zhenhai data, it is possible to say that for an F0 change to be considered intrinsic, and therefore ignored as pitch, it has to start near in time (within less than 10 csec) to the onset of the articulatory transition between syllables. Thus in [bā ju], the F0 starts to change about 5 csec after the onset of the F-pattern transition to the [j] of the second syllable. The start of the F-pattern transition in VCV examples was not measured in this study, but casual inspection of wide band spectrograms of tokens with monophthongal first vowel suggests that the F-pattern transition starts about 7 csec before the onset of the intervocalic consonant hold, which means that it precedes the start of the F0 transition by about that amount.

For an F0 transition to be perceived as pitch, it must start at a point near the completion of the articulatory change between syllables. For VCV syllables, this would be about 5 - 10 csec after release of the consonant, which leaves most of the F0 rise in level-convex examples to be perceived as pitch.
On the non-equivalence of fundamental frequency 77

Figure 7.
Temporally aligned spectrograms of $A_t$ (A) $F_0$ (B) and F-pattern (c) in the utterance \[ tā jū — a ha jīng \] "you can also see 'buy oil'". Start of $F_0$ rise and \[ j ] onset are marked by 't'; 'p' indicates perceptual onset of \[ u \]. Scale 10 csec: 1.27 cms.
There was no example in this study of a disyllable with level-convex pitch and intervocalic glide, but Figure 7 shows the acoustics of such a combination - \([t\ddot{a} ju 33 131]\) "to buy oil" - as it occurred in the running speech of a female Zhenhai speaker. Despite the bad F-pattern resolution typical for a female voice it can be seen that the temporal relationship between F0 and F-pattern changes is as predicted: the F0 starts to rise at about the same time as the \([j]\) onset is reached. Note also that the relationship of the Ar to the F0 during the rise is not as expected: the Ar profile resembles more that on the level-falling examples examined, and the ratio of Ar to F0 is about the same (0.2) as in the level-falling examples too. This is a possible indication of the relative lack of importance of the concomitant Ar in cueing the differential perception of F0 as pitch.

4. Conclusion

In this paper I set out to examine the current assumption in linguistics that tonal pitch is an exclusive and straightforward function of F0. In the first part I looked at factors known to mediate the perception of F0 as pitch, and used these to demonstrate that F0 is a basic, but by no means exclusive term in the function relating tonal pitch to acoustics. In the second part I demonstrated the differential perception of an F0 rise in a Chinese dialect tonal contrast. This suggested an additional complicating factor in the perception of F0 as pitch, namely allowance for the non-registration of intrinsic F0 transitions and their associated F0 perturbations. The domain of the function relating F0 to tonal pitch is therefore discontinuous. Because the domain is determined by the acoustic consequences of supralaryngeal activity, perceived segmental structure is an important factor in the interpretation of F0 as pitch.

Taken all together, these considerations showed that the interpretation of F0 as pitch involves its complex interaction with all the major acoustic parameters of duration, Ar and spectrum, and so the idea of pitch as an exclusive and straightforward function of F0 does, as initially surmised, represent a considerable oversimplification.

This has several consequences for approaches to tonal description. Firstly, because acoustic parameters other than F0 contribute to pitch, attempts to deduce or evaluate pitch characteristics from raw F0 traces alone can only have limited validity. This is true even for the apparently simple case of tones in citation form, since no F0 contour is produced without concomitant intrinsic effects, the perceptual relevance of which it is difficult to assess.

Rather than use acoustics to infer tonal pitch, an alternative approach would be to shift the emphasis away from the perceptual side (which is associated with several significant problems) and base Linguistic Phonetic tonal research on acoustical data. (This is still not the appropriate observation language for tonology, but at least it is quantifiable, and can also provide valuable insights into the production of tone (Rose 1984)). This paper has shown that it is desirable that such studies should include other acoustic parameters besides F0, like Ar and duration. In particular, it has highlighted the importance of segmental information in the interpretation of F0, and more attention should be paid to the accompanying articulatory acoustics in tonal studies, so that more can be known about the principles governing the integration of segmental and suprasegmental strands (Nolan 1983:51,52; Fujisaki 1977). This last comment applies to non ton
languages too, of course: it would for example be interesting to see to what extent the English intonational contrast between falling and rise-fall nuclei after a low prehead (are you [əˈjuː ɪ 51] vs. are you [əˈjuː ɪ 1 1 3 1]) parallels the Zhenhai level-falling vs. level-convex contrast.

NOTES

1 This is a revised and expanded version of a paper given at the 1984 Australian Linguistic Society meeting in Alice Springs. I am grateful to Dr. Michael Wagner of the Australian Defence Force Academy for putting his synthesis facilities at my disposal; thanks also to Dr. Francis Nolan of Cambridge University for providing helpful comments and a second set of auditory transcriptions for the Thai data.

2 Arguably, since the more recent omission of other acoustic parameters could possibly be justified by the results of a few perceptual experiments which show F0 to be the most important acoustic parameter in tone identification (Gandour 1978:42)

3 This is also plausible from a consideration of the processes involved. The equivalence of pitch judgements of different F0 values in different parts of an utterance can be explained in terms of the sameness of a VCT parameter under conditions of declining Ps, or, possibly, by the sameness of a Ps parameter under conditions of delining VCT. But in this case, perceptual equivalence of some kind obtains between combinations of different F0 and Ar values. Hence different Ps and VCT values must be involved, and what precisely remains the same during the utterance is elusive, unless it is some kind of quantal energy constant defined as the product of Ps and VCT. This state of affairs is more reminiscent of stress than tone or pitch.

4 A detailed description of the acoustic characteristics of the informant's 6 citation tones can be found in Rose (1982a). Shanghai-Zhenhai is a common variety of Zhenhai dialect which shows to a greater or lesser extent lexical and segmental influence from the neighbouring prestigious dialect of Shanghai.

5 By 'effect' here I mean perturbatory effect, or apparent deviation from a smooth curve. Intervocalic sonorants appear to have no effect on the F0 contour, which shows a smooth transition from low F0 values on the first syllable to high values on the second. In other words there is a small, positive derivative of F0 with respect to time (DF0/Dt) during the transition. Intervocalic obstruents, on the other hand, do have discernible effects on the F0, bringing about a slight overall decrease in the rate of F0 change over the first part of the consonant, followed by a larger increase over the second half. This increase is large enough to ensure that the F0 values at the onset of the second syllable vowel after an obstruent do not differ significantly from those after a sonorant. The obstruent also causes a very slight decrease in the F0 derivative after its release. This results in an additional F0 perturbation after the obstruent in examples with convex pitch. The same difference between obstruent and sonorant in F0 perturbatory effect after release can be seen in citation form (Rose 1982a:33).
The effect of a sonorant on Ar is small, involving a slight overall decrease in the Ar derivative. During hold and transition onto the second syllable vowel the Ar derivative is small and positive, like the F0, giving a smooth rise in Ar. But the transition onto the sonorant is different, depending on the following tone.

Obstruents have a large effect on Ar, involving a considerable drop followed by a rapid increase to onset of the second syllable vowel. Syllables with lenis initial consonant have a much greater Ar fall than syllables with normally voiced obstruent. After obstruents, the onset value of Ar (as measured down from peak Ar) is about 3.6 dB lower than after sonorants. This is unlike the F0 values at the onset of the second syllable vowel, which are not affected by the difference between obstruent and sonorant. Comparison with the Ar on vowels after sonorants indicates that the lowering effect of the obstruent on Ar appears to last at least over the first half of the vowel in convex pitched syllables.

\[ \frac{\Delta Ar}{\Delta t} \frac{\Delta F0}{\Delta t} \]

(Rose 1982b:383ff.).

It was not possible to preserve the original Ar levels during the process of editing from the master tapes and spectrographic analysis, because some tokens had to be amplified to enhance resolution of other acoustic parameters.

The value is probably greater than this because the effect of source change from whispery voice to modal voice has to be taken into account. One would expect more Ar from modal than whispery voice.

This reasoning holds, of course, as long as the normalisation has not obscured some perceptually relevant aspect of the difference between the two F0 shapes. Besides its slope, the duration of an F0 ramp is an important factor in perceived contouricity (Greenberg and Zee 1979:154), and the difference in pitch between the two varieties might be because the F0 rise in the Pakphanang tone is deliberately produced longer. Durational differences are usually equalised in F0 normalisation (as in Figure 6), on the implicit basis that they are intrinsically determined by properties of the F0. Thus a possible explanation of why apparently the same F0 shapes are associated with different pitches would be normalised away.

The present data, from two speakers of different sex under different elicitation conditions, cannot of course validate the hypothesis, but there are some indications from informal tests I have done with resynthesised speech that the duration of the F0 ramp does indeed contribute to the perceived contouricity of the convex F0. A resynthesised [ma] syllable with F0 values similar to those on the second syllable in level-convex sonorant examples had a much less salient convexity to its pitch than the Pakphanang tone. This is an important result, because it means that durational differences should be taken into account in a perceptually correct normalisation of the F0 of different varieties. (Another, indirect, indication of the importance of the duration of the F0 ramp would be if the Pakphanang high convex tone turned out to have
a significantly longer duration than the Zhenhai falling tone under conditions which controlled for intrinsic and extrinsic effects on duration.)

10 The speaker's F3 in non-nasalised [a] vowels is often difficult to spot, but seems to lie between 2500 and 2800 Hz. Therefore I suspect that this rising frequency may in fact represent his F3 at onset, but a nasal resonance subsequently, the true F3 pole having been attenuated by a nasal zero.

11 The rather large difference in the factors between glide and VCV examples (the ratio of F0 rise to Ar rise is 7, compared to 3 in VCV examples) may have something to do with the difference between citation form [ju] and running speech [ba ju] contexts.

12 A token very similar to this, from Osaka Japanese, appears in Fujisaki (1977:359), who illustrates the acoustics of a two-mora [ai] vowel sequence with a rising pitch component. (The pitch is described as 'low-high-low' (p. 350) and '[+ ]' (p. 349), but it is not likely, to judge from the acoustics, that it sounds like the pitch on the monosyllabic Zhenhai [ju]. There looks to be a short level pitch component at the beginning which possibly corresponds to the first prominence mark.) In this token, as expected, the onset of the F0 rise occurs just before the F-pattern has reached its steady state for [i].

13 Beside the main linguistic assumption questioned in this paper of equivalence of F0 and pitch, I have reservations on another basic assumption in tonal description, namely the appropriateness of transcribed pitch as the observation language. This topic really needs a separate paper, but the following points should be noted. Firstly, perceptual descriptions or features like [55], [Rise] or [H] are clearly inappropriate to explain tonal processes that are productionally motivated (e.g. 53 → 55 / — \(\{\begin{array}{c}
55 \\
53
\end{array}\}\)) (Hyman 1975:218)). Secondly, we know little about the within- and between-transcriber variability in pitch transcription. Thirdly, a lot of uncertainty surrounds the perceptual mode of the transcriber. The native speaker's perceptual response to an acoustic speech stimulus is determined to a large extent by their native language. Speakers of tone languages, for example, pay attention to different features of the acoustic stimulus than speakers of non tone languages (Gandour 1978:62-71). Transcriptional performance also reflects this bias to a certain extent (Butcher 1982:62ff.). It is not clear to what extent the transcriber should try to - or indeed be able to - perceive in the same mode as a native speaker. It might seem desirable that the transcriber emulate native speaker response, but it is my experience that one's native perception of prosodic cues is very difficult to suppress. It should also be noted that some scholars (e.g. Anderson 1978:154) explicitly reject native speaker perceptual response as irrelevant in accounts of linguistic structure. On the other hand, there is some danger that a 'language-neutral' transcription might incorporate non linguistic (i.e. musical, non speech specific) modes of perception, especially in cases where the transcriber is concentrating on perceiving the details of a pitch contour (cf. Studdert-Kennedy 1974: 2361-2363). (Note that in those experiments that test native speaker response - see e.g. Kratochvil (1975) and other references...
in Gandour (1978) - a transcription is really superfluous, and only necessary as part of the method of establishing the linguistic contrasts to be tested.)

In both parts of this paper I have made the normal linguistic assumption, necessary for the evaluation of the main question, that the transcriber's mode is unproblematic. In the second part, for example, I have assumed that my perception and transcription of pitch are relevant and adequate for linguistic description. And it is implicit in the first part that the factors which relate a native speaker's perception of F0 to a linguistic unit are the same as those which relate F0 to a transcriber's pitch percept. Although it does not affect the validity of the demonstration in this paper that pitch and F0 are not equivalent, I hope my earlier comments show that these assumptions carry a very big caveat indeed.