1

INTRODUCING PROSODIC PHONETICS

Vincent J. van Heuven*

1. What is prosody?

Traditionally, phonetics is the study of speech sounds. It tries to characterize all and only the sounds that can be produced by the human vocal organs in the context of spoken language. Speech sounds, in turn, are defined as complex wave forms with relatively broad spectral energy bands that vary continuously as a function of time. This definition excludes singing (with relatively long portions during which the energy distribution hardly changes) and whistling (distribution of energy in narrow frequency bands only) from the domain of spoken language.

However, it is widely acknowledged today that there is more to phonetics than just studying the properties of the vowels and the consonants that make up a spoken sentence. In this context it is expedient to distinguish between segmental phonetics and prosody.

1.1 Segmental phonetics

Segmental phonetics studies the properties of utterances in so far as these can be understood from the properties of the individual segments (the vowels and consonants) in their linear sequence. Segmental phonetics addresses firstly the specification of individual vowels and consonants. This includes such articulatory features as manner, place, and voicing, as well

*I am indebted to Renée van Bezooijen and Toni Rietveld (Nijmegen University) as well as to all the contributors to this volume for their comments on an earlier version of this chapter.

as their acoustic and auditory correlates. Together these features define the phonetic quality (or timbre) rather than the quantity (length, duration) or pitch of speech sounds. Contrary to earlier treatments of the subject (e.g. Gimson 1969:55),[1] segmental phonetics also includes the inherent duration and inherent pitch of individual segments. It has been known for quite some time, for instance, that the jaw takes longer to reach its target position for the articulation of a low vowel [a] than for a high vowel [i] or [u], so that low vowels are generally longer than high vowels. This effect of vowel height on duration is segmentally conditioned, and therefore belongs to the realm of segmental phonetics. Similarly, high vowels are generally uttered on a higher pitch than low vowels; again this effect of inherent pitch is a segmental phenomenon.

In the production of speech the vocal organs change from one articulatory position to the next in a relatively slow and continuous fashion, so that the movement of the articulatory organs can be traced back from the continuously varying spectral energy bands in the acoustic signal. When a particular speech sound is heard in continuous speech, the listener is usually able to tell not only which sound is being uttered at that point in time, but also what the preceding and following sounds are. The mutual influence that sound segments have upon one another when they are uttered in continuous speech, is called coarticulation. Coarticulation, in our conception, still belongs to the domain of segmental phonetics.

1.2 Prosody

In contradistinction to the above, prosody comprises all properties of speech that cannot be understood directly from the linear sequence of segments. Whilst segmental properties serve to make the primary lexical

[1]The more traditional demarcation between segmental phonetics and prosody was in terms of quality (or spectral distribution of energy) for the former versus length (duration), pitch, and stress (loudness) for the latter. It was held that length, pitch and loudness *could* be properties of domains larger than the single phoneme. This suprasegmental (or prosodic) nature was implicitly denied in the case of phonetic quality. This latter view is demonstrably wrong given the existence of harmony phenomena (e.g. spreading of quality features between the vowels within a word but not across a word boundary).

distinctions, the linguistic function[1] of prosody is:

1. to mark off domains in time (e.g. paragraphs, sentences, phrases),
2. to qualify the information presented in a domain (e.g. as statement/terminal boundary, question/non-terminal boundary), and
3. to highlight certain constituents within these domains (accentuation).

The smallest domain that can be marked off is the syllable. When a vowel is pronounced at the end of a syllable (open syllable) it is longer than when it is pronounced – ceteris paribus – in a closed syllable, as in the pair *Grace eyed* vs. *grey side*.[2]

Prosody literally means 'accompaniment' (Gr. *pros odein* 'with the song'). This suggests that the segmental structure defines the verbal contents of the message (the words), while prosody provides the music, i.e. the melody and the rhythm. Indeed, prosody is often divided into two broad categories of phenomena: (1) temporal structure and (2) melodic structure. Let us briefly discuss these two classes of prosodic phenomena.

2. *Temporal structure*

I define the temporal structure of a language as the set of regularities that determine the duration of the speech sounds (or of the articulatory gestures underlying these sounds) and pauses in utterances spoken in that lan-

[1]Notice that we take the view here that the signalling of the speaker's attitude (e.g. approval, disgust, etc.) towards the verbal contents of the message or the expression of his emotion (happiness, fear, etc.) through prosody is not a linguistic matter. This is a rather arbitrary position, and I shall give it up as soon as I find convincing evidence that prosodic signalling of attitude and emotion is language specific and rule governed. Van Bezooijen 1984, in her intercultural study on the perception of emotion, presents evidence that allows us to argue both ways. Emotions expressed by Dutch speakers were recognised at more than three times better than chance (=10%) by Japanese and Taiwanese listeners (37 and 33% correct, respectively). This finding seems hard to reconcile with the view that the expression of attitude and emotion obeys language-specific rules. Yet, Dutch listeners identified the (Dutch) speaker's emotions much better still (66% correct), so that there must be a considerable language-specific component in the process. So far, no-one has been able to come up with rules for the synthesis of emotions in synthetic speech, not even for a single language (Carlson, Granström and Nord 1992).

[2]In this example, incidentally, the syllables make up meaningful units, be they words or morphemes. Generally, it seems, there is no need for marking off syllabic domains unless the syllable boundary coincides with a morpheme boundary.

guage. With the exception of intrinsic and co-intrinsic duration (see above) temporal structure depends on knowledge of the higher-order linguistic structure of the utterance. Temporal structure is typically used to signal cohesion between words on the one hand (usually by speeding up words within a prosodic constituent) and discontinuity on the other (by slowing down (parts of) words immediately before a prosodic boundary (pre-boundary lengthening).

2.1 Linguistic hierarchy and temporal coding

There is a wealth of evidence to suggest that speech rate is higher the longer the stretch of sounds the speaker intends to produce. Sounds are pronounced faster at the beginning of an utterance and speech rate slows down gradually towards the end of the sentence. Also, speaking rate is higher for longer utterances than for shorter utterances. It seems as though the words a speaker intends to utter are stored under pressure in a buffer. The more words are pushed into the buffer, the higher the pressure, and the faster the stream of sounds that is produced when the buffer is emptied. As the contents are released from the buffer, pressure diminishes, so that speech rate gradually slows down (further see Lindblom, Lyberg and Holmgren 1981).

The hierarchical organisation of linguistic structure is reflected to some extent in this temporal behaviour. Speaking rate is fastest at the beginning of a sentence, and even faster when the sentence is at the beginning of a paragraph. On a lower level, the beginnings of words are usually pronounced faster than the final syllables of words, and sounds are pronounced shorter in longer words than in short words. Fast rate and controlled relaxation of speaking rate are signs of cohesion (sounds and/or words belonging together).

On the other hand, breaks in the linguistic structure are temporally signalled by some degree of pre-boundary lengthening, the extent of which is controlled by the depth of the boundary at issue. There is a general tendency for the final syllable of a word to be longer as the depth of the boundary following it is deeper. Accordingly, there is only marginal pre-boundary lengthening in the middle of a constituent, but appreciable

lengthening in words at the end of an intonation phrase, sentence or paragraph. Moreover, when a boundary exceeds a certain depth, e.g., for boundaries marking off intonation phrases or even longer domains, the pre-boundary lengthening will almost certainly be accompanied by a pause, i.e. a physical silence.[1] Again, the duration of the pause increases with the depth of the linguistic boundary marked by it. Whenever pauses occur, any assimilation or coarticulation between sounds straddling the boundary will be blocked: they will be coarticulated to silence.

These aspects of temporal organisation indicate a great deal of pre-planning on the part of the speaker. Apparently the speaker has some notion of how much linguistic material he is going to utter before the next break, and what kind of a break this is going to be. Also, there are persistent claims that this planning strategy is stronger during premeditated speech (oral reading, rehearsed lines) than during spontaneous speech production (conversation, improvised lecturing).

2.2 Between-language differences in temporal structure

I am not aware of any differences between languages in terms of their macro-temporal organisation. All languages seem to reflect their higher-order syntactic/prosodic structure by the same temporal means, i.e., stronger pre-boundary lengthening and longer pauses accompanying deeper structural breaks. However, it should be pointed out that no comparative studies have ever been done on macro-temporal organisation.

Differences between languages in lower-level temporal organisation have been researched more extensively. I shall shortly dwell on lower-level temporal phenomena, even if these are excluded from the realm of

[1]Berkovitz (1993), however, claims that preboundary lengthening before a gapped position is implemented by a qualitatively different timing pattern than before an other deep boundary. Before ordinary boundaries segments are progressively decelerated, so that the final segment has the longest duration (relative to its inherent duration). In pre-gap position the final segment is relatively short whilst the preceding vowel is stretched much more. The observations have been made for Hebrew. A methodological weakness in the research is that ordinary preboundary words and pre-gap words were collected in separate experiments, using different speakers and different lexical materials. More controlled research is needed to substantiate Berkovitz' claim.

prosody (see above), as they present a fascinating challenge to comparative phonetics.

For quite a number of languages data have been published on the inherent duration of the sounds in their phoneme inventories. Even superficial inspection of the available data suffices to conclude that systematic generalisations are hard to make. For instance, the general claim (see above) that vowels are inherently longer as they are more open, is extremely hard to substantiate in any single language. Even in a language such as Indonesian, with a simple six-vowel inventory and lacking a short-long contrast, we found that the mid vowels were longer than either the low or the high vowels (Van Zanten and Van Heuven 1983; Van Zanten 1989). What would be needed here is a general phonetic theory that precisely predicts the inherent duration of an arbitrary vowel in a language given the spectral properties (i.e. the phonetic quality) of all the vowels in the phoneme system in that language. One could foresee two forces operating on the duration structure of the vowel system:

1. a general force reflecting the effect of mouth opening, and
2. a system contrast force compensating a lack of spectral distinction between neighbouring vowels.

Thus, we would predict that Dutch open /a:/ is rather long anyway because it is an open vowel; it would be longer than Indonesian or Spanish /a/ because it also has to be differentiated from short /a/, but not as long as a long /a:/ in, e.g., Hungarian. In contrast to Dutch, where short /a/ is also more back than long /a:/, cognate short-long vowels in Hungarian have exactly the same phonetic quality, so that in Hungarian the duration difference should be longer in order to make up for the lack of spectral distinctivity. The theory would also have to take into consideration the (lack of) perceptual contrast between non-cognate vowels. So far, there is an elegant theory that predicts the phonetic quality for the vowels in an arbitrary language given the number of vowels in the inventory (Liljencrants and Lindblom 1972; Ten Bosch [n.d.]); however, this theory does not (yet) address any of the duration issues.

Going up to the level of the syllable, comparative studies seem to be clustered around the phonetic implementation of the voicing contrast. In onset position the research has focused on the phenomenon of voice onset time (VOT). The relevant parameter here is the time difference between

the onset of voicing and the moment of consonant release. When voicing begins before the mouth opens, VOT has a negative value (voice lead, expressed in ms); when the onset of voicing follows after the consonant release, we have positive VOT (voice lag). The voice-lag period is typically filled with a voiceless vowel sound (traditionally called aspiration). Two and even three member contrasts are made along the VOT-axis. Dutch implements a two-member contrast where voiced stops have negative VOT and voiceless stops have zero VOT. English also makes a binary contrast, but positions the voiced stop at zero VOT and the voiceless cognates at positive VOT-values. Thai is an example of a ternary opposition, with a [lax, voiced] member at negative VOT values, a [tense, voiceless] stop at zero VOT and a [lax, voiceless] stop at positive VOT values.[1]

In medial and final positions the research has concentrated on the duration ratio of consonant and preceding vowel. Typically, when the consonant is long, the vowel is short, and vice versa. Long consonant with short vowel codes a voiceless obstruent, whilst the reversal of these cues (short consonant preceded by longer vowel) is the phonetic correlate of a voiced obstruent. In some languages the ratio differs only a little for voiced and voiceless (medial) obstruents (e.g. Spanish); in other languages the ratio difference may be much larger (Delattre 1965). The largest difference is found in English, a language which maintains a clear voicing contrast even in final position, where the contrast is neutralised in most other languages.[2]

Finally, turning back to prosody, there is a persistent claim that languages can be ordered in terms of their rhythmic behaviour along a scale

[1]There are strong indications that the voice-lead portion as such is perceptually irrelevant. It has low intensity and contains low frequency components only. This type of sound is typically masked by the ambient noise or even by forward masking from the preceding vowel. In Dutch and English, the voiced-voiceless opposition is effectively communicated even in whispered speech, where any contribution of voicing is cancelled, indicating that other correlates of the contrast are effective and more robust than the mere presence vs. absence of vocal cord vibration. One would like to know to what extent the Thai three-member contrast is held up in whispered speech.

[2]Interestingly, there is quite a body of research to show that even languages where an underlying voiced-voiceless contrast is neutralised in word-final position, the difference can still be measured acoustically (in terms of the vowel/consonant duration ratio) on the phonetic surface. Such differences are largely subliminal and have little or no perceptual relevance (cf. Port and O'Dell 1985).

that runs from syllable timed on one end to stress timed on the other (Abercrombie 1967). In an unadulterated syllable-timed language all the syllables have equal duration (or: syllable isochrony), regardless of such factors as stress, yielding a staccato-like rhythm (e.g. Spanish). At the other extreme we find languages such as English, which have foot[1] isochrony, i.e., where the time interval between successive stressed syllables is constant regardless of the number of unstressed syllables intervening between two stresses. In stress-timed languages the duration of the syllables, including the stressed syllable, is shorter as more syllables are squeezed in between two stresses.[2] It remains to be shown, however, if there is more to syllable timing vs. stress timing than just a conspiracy of lexical properties. It appears that syllable-timed languages typically have no vowel length contrast, have open syllables, do not allow complex consonant clusters, and do not reduce vowels in unstressed position. Stress-timed languages, on the other hand, allow complex (and closed) syllables, often have a vowel length contrast, and reduce unstressed vowels to schwa (Dauer 1983). Consequently, when speakers of a stress-timed language such as Dutch pronounce words of Italian origin like *macaroni, spaghetti*, or *salami*, the timing is the same as that of Italian speakers, representing a syllable-timed language (Den Os 1985).

This excursion on stress timing versus syllable timing shows that it is important in comparative research to sharply distinguish differences in linguistic structure from phonetic differences.

[1]The Abercrombian foot is the time interval beginning with a stressed syllable and extending to the next stressed syllable, and includes all intervening unstressed syllables. It has no internal (binary constituent) structure; in metrical phonology it would be dubbed an unbounded left-dominant foot.

[2]In a way, this so-called anticipatory shortening of the stressed syllable (as a function of the number of unstressed syllables following within the foot), and the tendency to squeeze in more unstressed syllables without increasing total foot duration, are a manifestation of the tendency noted earlier in this chapter for speaking rate to increase at the beginning of a new prosodic constituent. One may therefore seriously question whether stress-timing should be viewed as an independent linguistic/typological parameter.

3. Melodic structure

Melodic structure can be defined as the set of rules that characterize the variation of pitch over the course of utterances spoken in a given language, excluding micro-variations due to intrinsic and co-intrinsic properties of segments from the discussion. We know with near-certainty, that no two languages have the same melodic properties.

3.1 Linguistic structure of speech melody

In terms of linguistic structure the melody of a language is defined by the sequence of discrete pitches (typically one level pitch per syllable), which can assume only a limited number of values (never more than four (high, mid-high, mid-low, low), and typically not more than two or three: high, (mid,) low. In some languages syllables may carry two successive pitches, sometimes accounted for by assuming a sub-syllable timing unit (called mora), such that each mora carries its own pitch. Whichever the case may be, successions of two different pitches define contour (i.e. non-level) tones (rises or falls) on a syllable. The phonological component of the grammar of the language should specify the inventory of tones (i.e. number of levels) and contain rules that define legal successions of pitches making up tonal configurations and intonation patterns.[1] Obviously, languages may differ both in terms of the tone inventory and in the combination rules.

In so-called tone languages the pitch, or sequence of pitches, within a word is lexically determined, i.e., functions to distinguish between words in the lexicon roughly the same way segments do. In intonation languages the sequence of pitches does not serve lexical distinctions; it may have other linguistic functions, such as highlighting (focusing) important

[1]Here I basically follow the tenets of autosegmental tonology (cf. Gussenhoven 1988; Gussenhoven and Rietveld 1991 and references given there). I believe that a tonal representation in terms of discrete levels underlies the pitch movements that can be observed on the phonetic surface. The theory of intonation developed at the Institute for Perception Research at Eindhoven ('t Hart, Collier and Cohen 1990), which has had an enormous influence on prosodic studies in the Netherlands, is predominantly a theory of phonetic implementation of this underlying structure.

words in the utterance, marking break positions in the syntactic/ prosodic structure, and qualifying such breaks as either terminal or non-terminal.

These two uses of pitch (marking lexical distinctions vs. highlighting important information in sentences) seem hard to combine. Many tone languages use particles (i.e. separate words or morphemes) in order to express focus. Still, there are languages that exploit the pitch parameter to code both lexical and sentence-level distinctions, and one would like to learn how this is done. Preliminary results of research that we embarked upon to shed light on this matter, indicate that focus in Mandarin Chinese is marked by expanding the pitch range within which the four lexical tones are realised. Thus, the high level tone (tone 1) assumes a higher pitch in an important word than it would have had in an unimportant word (Van den Hoek 1993). By the same token the three Mandarin contour tones (tone 2: rising, tone 3: dipping, tone 4: falling) are given larger excursions in focused position than in non-focused positions.[1]

In the phonetic manifestation of the sequence of pitches over the course of an utterance, the discrete character of the individual pitches gets largely lost. The pitches are strung together through tonal coarticulation and what we observe on the phonetic surface are pitch movements only.[2] Languages differ systematically in the way these pitch movements are made. Movements may differ along a restricted set of melodic parameters, such as the direction (rise, fall), size (large, small), abruptness (steep, gradual), and segmental alignment (early, late in syllable). It was found, for instance, that English pitch movements are steeper than their closest Dutch counterparts.

[1]One wonders what happens when a language has lexical tones, intonational focus as well as lexical stress. Since all three the distinctions are coded melodically (as well as temporally) some complex arrangement will have to be found. Papiamentu is claimed to be a language where this rare combination of prosodic structures occurs.

[2]Typically, the pitches in tone languages keep much more of their underlying discrete character than those in intonation languages. The melodies of utterances in tone languages are often described as being akin to singing, i.e., a note or level tone per syllable.

3.2 Phonetic correlates of pitch

During phonation the vocal cords open and close rapidly. During each cycle of opening and closing of the vocal cords, a puff of air is released from the larynx into the throat. Given that the vocal cords of a male speaker open and close between some 70 and 200 times per second, the larynx functions as a machine gun, shooting some 70 to 200 air bullets into the throat, generating a complex harmonic sound with a fundamental frequency of 70 to 200 hertz (Hz). It is this series of rapid sharp taps that gives human speech its voice, its carrying power. When the rate of vocal cord vibration is low, the pitch is low, when the firing rate increases, the pitch goes up accordingly.

There is more to pitch than just this. During phonation the larynx is not stationary. In order to produce a low pitch the speaker pulls his larynx and tongue root down, thereby increasing the length of the vocal tract (especially that of the mouth cavity). Conversely, during the production of high-pitched sounds the larynx and the tongue root are raised, thereby effectively shortening the vocal tract, particularly the mouth cavity. The variation in length of the vocal tract is reflected in the resonances that give the various speech sounds their phonetic quality. Thus there are indications that especially the second lowest resonance (called second formant or F_2, which predominantly reflects the length of the mouth cavity) goes up and down with the movements of the larynx and tongue root, and therefore mimics the fundamental frequency (i.e. pitch).[1] This is one reason why whispered speech, where the vocal cords do not vibrate, still conveys some sense of melody.[2]

The rate of vocal cord vibration roughly depends on two factors. One is

[1]The reverse also holds: when a high vowel is produced, the tongue root and the larynx which is attached to it, are pulled up, so that the vocal cords start vibrating faster (Ohala 1978). This is currently the most plausible explanation for the inherent pitch phenomenon discussed at the beginning of this chapter.

[2]As a case in point, Mayer-Eppler (1957) showed that the difference between German declarative and question intonation could still be heard in whispered speech. Several studies followed showing that lexical tone differences were effectively communicated in whispered speech (Kloster-Jensen 1958; Miller 1961; Wise and Chong 1957). More anecdotally, a Jesuit priest claimed he had no problems when Chinese christians whispered their sins to him during confession, even though they depended on pitch to make lexical tone distinctions.

the pressure difference across the glottis. Simplifying somewhat, the more air pressure there is below the vocal cords, the faster they vibrate. During the production of an utterance the air trapped in the lungs is gradually expended, so that subglottal air pressure, which is high immediately after inhalation, gradually decreases towards the end of the utterance. This explains, to some extent, why spoken sentences usually start at a higher pitch than they end. This gradual downtrend of the pitch over the course of an utterance has come to be called declination. Generally, when the speaker intends to utter a long sentence he inhales more air than when he plans to produce only a short sentence (see above for a similar claim with respect to temporal organisation). As a consequence, longer utterances start at a higher pitch than short utterances, and their pitch goes down more slowly.[1] In fact, the speaker takes the deepest breath of air at the beginning of a paragraph, progressively shallower breaths of air prior to each successive sentence within the paragraph, and still shallower breaths of air at boundaries within the sentence. As a result, the vocal pitch is reset to a higher level at each prosodic break, with larger resets at the deeper boundaries.

The second factor determining the rate of vocal cord vibration is the tension of the vocal muscles themselves. Fast pitch movements, whether rises or falls, are typically caused by rapid changes in the tension of the vocal cords through muscular adjustments.

There is ongoing debate about the division of labour between voluntary and involuntary processes that underlie this encoding of linguistic structure in downtrend and pitch resets. Obviously, inhalation must be under the voluntary control of the speaker, since the volume of air inhaled is commensurate with the speaker's estimation of how much air he needs to produce the next utterance. However, though there is a general tendency for the pitch to go up and down with changes in subglottal air pressure (about 7 Hz per cm water, cf. Ladefoged, 1967:14), the actual magnitudes of the declination effects and resets across boundaries are larger than can

[1] A useful formula for calculating the declination (D, expressed in semitones per second) as a function of the duration (t, in seconds) of an utterance is given by 't Hart, Collier and Cohen (1990:128): D=(-11)/(t+1.5). For utterances longer than 5 seconds the declination interval is limited to 8.5 semitones maximum. This formula yields perceptually adequate results in artificial speech. It is often difficult to observe declination in human speech production, especially so in unpremeditated speech.

be explained by the influence of subglottal pressure alone. We must assume, therefore, that part of the linguistic encoding is brought about by voluntary adjustments of the laryngeal muscles ('t Hart, Collier and Cohen 1990:140). Also, it is possible (though not often observed) for a speaker to reset the pitch baseline after a linguistic break without having to inhale. Moreover, in Danish, differences in global downtrend over the course of an utterance are used to signal different sentence types: normal declination signals statement, shallower declination represents continuation, and the absence of declination (or even slight inclination) is characteristic of questions (Thorsen 1980). Speculating to some extent, I suggest that intonational downdrift in unmarked sentences passively reflects the decrease in subglottal air pressure, but additional effects, caused by voluntary actions of the laryngeal and/or respiratory muscles, may amplify or counteract the passive effects in order to signal structural breaks and marked sentence types.

3.3 From pitch to melody

Roughly, pitch is auditorily evaluated along a logarithmic scale, i.e., in terms of musical intervals. For this purpose, fundamental frequency in speech is conveniently expressed in terms of semitones above some arbitrary base-line (usually 50 Hz, which seems to be the bottom pitch for a male voice). A semitone, the pitch difference between two adjacent tones on the piano keyboard, is a 6 percent difference between two frequencies. Twelve semitones, i.e. 12 consecutive 6 percent increments (with compound interest), comprise an octave or a doubling of the frequency.[1]

There are over a hundred different computer programs (Hess 1983) that determine the rate of vocal cord vibration from a speech waveform stored in computer memory. None of these programs are perfect, but the errors they make are generally easy to detect and correct.[2] There is a dis-

[1] A convenient formula for converting the pitch interval between two tones f_1 and f_2 expressed in hertz into semitones is the following:
$c * {}^{10}\log(f_1/f_2)$, where $c = 12/{}^{10}\log(2) = 40$.

[2] In our laboratory we use a pitch determination algorithm based on the method of subharmonic summation (Hermes 1988) in combination with a pitch tracking routine. The tracking routine knows the limitations of the human voice for pitch changes over time,

concerting discrepancy between the simplicity of the melodic pattern that our ears perceive and the visual chaos that we see in pitch traces drawn by the computer pitch determination program. The hearing mechanism ignores the majority of the short term pitch fluctuations, abstracting away from irrelevant detail, and extracts only the major relevant pitch movements. A first step towards a meaningful description of the melodic properties of a language, therefore, is to reduce the raw pitch curves as determined by the pitch extraction algorithm to a series of straight lines, such that only the perceptually relevant movements are maintained. The result can be listened to after resynthesizing the utterance, keeping all the properties of the original utterance unchanged except for the pitch.[1] (further see 't Hart, Collier and Cohen 1990; Odé, this volume). This stylization can either completely be done by trial and error, or by an automatic procedure, whose output still has to be checked by hand.[2]

Stylized movements are then sorted into a limited number of categories in accordance with the intuitions of native listeners of the language (for an example of this type of research see Ebing, this volume). Finally, standardized specifications are drawn up for each category, typically by adopting the mean values measured for excursion size, abruptness and segmental alignment as the standard values. Substituting standardized movements for the original movements in the pitch trace may yield audible differences after resynthesis but should never lead to the perception of a different speech melody. The ultimate test of the descriptive explicitness of this part of the intonation grammar is a computer program that takes a raw pitch curve as its input and identifies the perceptually relevant pitch

---

so that impossible or highly improbable changes are ignored. As a result, most errors in the original pitch determination are automatically corrected; any remaining errors are so gross that they are immediately spotted by the researcher.

[1]Stylization of natural variability in speech utterances and checking the result after resynthesis is a research strategy that has pervaded experimental phonetic research since the early fifties. It was first applied to extracting the perceptually relevant changes in the resonances of the vocal tract (first and second formant), that could be made visible in wide-band spectrograms. Original spectrograms and hand-painted, stylized copies of them could be converted back into speech through a device called the Pattern Playback (Cooper, Liberman and Borst 1951).

[2]This algorithm (module STY in the LVS signal processing package) was developed at IPO Eindhoven by Hermes.

movements in terms of their distinctive features.[1]

4. Accent and stress

4.1 Linguistic structure and accent

Specific variations of duration and pitch are used in a large number of languages to make one syllable prominent within a larger domain. This syllable is called the accented syllable or simply the accent. Although multiple accents may occur within a sentence, one accent is felt to be stronger than any of the others. Accent is therefore culminative (Trubetskoy 1958). Culminativity is the property by which accent is different from (lexical) tone: successions of two or more high tones within a prosodic domain are perfectly legal, but no two accents of equal strength can coexist within one sentence. The function of accent is to mark a prosodic domain for focus, i.e., as important for the listener. There may be various reasons why a speaker chooses to put a constituent in focus. For instance, referents that are newly introduced into the discourse are typically put in focus, whereas referents that have already been identified in the preceding context are left out of focus, as is exemplified in (1). By convention, accented syllables are capitalised and focused constituents are presented in square brackets marked with+F; material out of focus is marked with -F (only crucial words are marked as +/-F):

(1) [PAris]$_{+F}$ is the CApital of FRANCE.
I like [PAris]$_{+F}$ a LOT.

However, when two or more referents are already known to the listener, but still represent a choice, each of these can be put in focus on repeated mention in the next sentence, as in (2)

(2) BerLIN and PAris are BEAUtiful CIties.
I think I like [PAris]$_{+F}$ BEST.

[1]This type of research is currently underway at IPO (Ten Bosch, Hermes and Collier 1993) for Dutch. Similar work is being done for German intonation by Möbius and Pätzold 1992.

As a first approximation each word in a +Focus domain is accented. However, since this would lead to an explosion of accents, the speaker may economize: all the words in a prosodic domain may be presented as in Focus by merely accenting the prosodic head of the constituent. Which word is the prosodic head of the constituent depends on the type of structure. For instance, in the prepositional phrase *at the back of the old HOUSE* the prosodic head is the final noun.[1] This entire constituent would be in Focus if it were the answer to the question:

(3) Q. Where did you park the CAR?
A. I parked it [at the back of the old HOUSE]$_{+F}$.

This type of focus is called broad or integrative focus (Fuchs 1984). Notice, however, that exactly the same accentuation would obtain if only the final noun *house* were in focus, as in (4), which is an example of so-called narrow focus:

(4) Q. Did you park the car behind the old BARN?
A. (No,) I parked it at the back of the old [HOUSE]$_{+F}$.

Accenting the prosodic head then always leaves an ambiguity that can only be resolved through contextual information. Accents on words other than the prosodic head do not have this ambiguity, as is exemplified in (5), which could never be the answer to either question (3) or (4):

(5) Q. Behind WHICH barn did you park the car?
A. I parked it behind the [OLD]$_{+F}$ barn.

The procedure can be repeated at the level of the word (and once more at the level of the syllable, Van Heuven 1994). The answer in exchanges (6) and (7) is identical. In (6), however, the final syllable is in narrow focus, contrasted with the final syllable of an otherwise identical word, whereas

[1]For an elaborate treatment of rules that determine the position of the prosodic head in Dutch see Baart [n.d.].

in (7) it marks broader focus since the entire word is contrasted.

(6) Q. Did you diVERT or diGEST it?
A. I di[GEST]$_{+F}$ed it.

(7) Q. Did you EAT or diGEST it?
A. I [diGESTed]$_{+F}$ it.

By definition, the syllable that is accented when a single, whole word is in focus, is the stressed syllable.[1] As before, accenting this syllable leaves a focus ambiguity than can only be resolved through context. In contradistinction to this, accenting a non-stressed syllable creates no such ambiguity. Accenting the initial syllable of *digest (vb.)* is possible in contrastive situations like (8), but could never happen if it were the answer to the question in (7):

(8) Q. Did you SUGgest or DIgest it?
A. I [DI]$_{+F}$gested it.

This conception of stress as the prosodic head on the word level only works for languages that have deterministic rules for stress placement. About half of the languages in the world have word stress. Of these, the majority have stress in a fixed position, determined by a single rule, e.g., always stress the initial syllable (e.g. Hungarian), or always stress the penultimate (e.g. Polish).[2] In other, so-called quantity-sensitive, stress languages (e.g. English, Dutch) the position of the stressed syllable is determined by more complicated rules, which typically stress the syllable that contains the largest number of segments.[3] In Dutch, about 85 per cent of

[1]This definition of stress is basically that of Bolinger 1958, where he defines stress as the docking site of the accent.

[2]Such stress rules always refer to the word edge, whether left or right. Notice that no language has a rule that stresses the middle syllable of a word. One would like to know the psychological reason behind this.

[3]In these quantity-sensitive stress rules segments in the syllable onset are ignored. Recent experimental data show that the human hearing mechanism is much less sensitive to duration variation in onset consonants than to variations in the vocalic nucleus and coda consonants (Goedemans and Van Heuven 1993).

the non-compound words receive their stress through quantity-sensitive rules (Langeweg [n.d.]). Finally, languages may have lexical stress. Here the stress position varies unpredictably from word to word, so that for each word the stress position would have to be stored in the lexicon.

When a language has free stress, there will be no integrative accent position on the word level. By our definition, such languages have no stress, they have accent only. Whichever syllable is accented, the result will always be ambiguous: the accent may signal narrow focus at the syllable level, or integrative focus on the word level.

Indonesian presents a confusing picture in this respect. On the one hand, stress is traditionally described as basically fixed on the penultimate (cf. Laksman, this volume; Odé, this volume). Yet, Ebing (1991) showed that native Indonesian listeners are unable to determine whether an accent in a particular word is in the integrative position or not: speakers were unable to produce the predicted differences in Indonesian counterparts to examples (7) and (8) above, nor were listeners able to decide which answer matched which question.

4.2 Phonetic correlates of prosodic prominence

Prosodic prominence, or culminative accent, has a dual linguistic representation. On the one hand, it has a tonal representation, a sequence of high and low tones, where abrupt changes between levels generate tonal prominence on a syllable. Tonal prominence is used to mark [+Focus] domains. The vocal cords typically vibrate slightly faster during the production of a vowel than during the production of a (voiced) consonant.[1] As a consequence of this, any syllable tends to show a shallow rise-fall pitch movement. Accent-lending pitch movements therefore have to exceed a threshold excursion size (for an average speaker something on the order

[1]This is an automatic consequence of differences in vocal tract configuration between vowels and consonants (Ohala 1978). Vowels present no obstruction to the outgoing airstream, so that the pressure drop across the glottis is relatively large, generating faster vocal cord vibration. Consonant articulation by definition involves an obstruction to the expulsion of air from the vocal tract. Therefore the intra-oral pressure is high relative to the subglottal pressure; the transglottal pressure difference diminishes during the articulation of a voiced consonant, so that the rate of vocal cord vibration drops accordingly.

of 3 semitones). The threshold level is variable, and will be perceptually adjusted (normalised) by the listener so as to optimally suit the behaviour of a given speaker. Generally, the size of the movement correlates with the perceived strength of the accent: the larger the excursion size, the stronger the accent.[1]

Moreover, for a tonal change to cause the perception of an accent, it has to be abrupt, i.e. characterized by a steep pitch movement, and it has to occur in a specific position within the syllable.[2] In Dutch, for instance, an accent-lending rise has to start at the beginning of the syllable, whereas an accent-lending fall has to be late in the syllable. If a steep rise occurs late in the syllable, or a fall early, the movement signals a break in the linguistic structure (boundary tone) but does not generate accent.[3]

The second representation of prominence is temporal. This is a hierarchical structure of metrically strong and weak syllables, whose principal correlates are temporal. Strong syllables are longer than their weak counterparts. When polysyllabic words are pronounced in a [-Focus] domain, they will no longer bear a pitch movement (Van Heuven 1987) but a stressed syllable will still be longer (by 50 to 100 per cent) than its unstressed counterpart. It will also have greater intensity and less spectral reduction (i.e. reduction towards schwa), but these differences are perceptual accent cues of lesser importance.[4] Moreover, unaccented words,

[1]This scalar effect should not be confounded with claims in the older literature on the all-or-nothing cue value of pitch movements in the perception of stress. In the experiments concerned (e.g. Fry 1958; Morton and Jassem 1965) listeners indicated stress in isolated di-syllabic words. Under such circumstances any pitch change larger than a semitone is interpreted as accent-lending.

[2]It is unclear what happens when a language has no deterministic stress rules, such as Indonesian. Possibly, any steep rise or fall may cause an accent to be heard, except for a fall at the beginning of a word and a rise at the end of a word; these latter two would then function as boundary tones. We would predict that native Indonesian listeners are less susceptible to the exact position of pitch movements within a word than, e.g., Dutch or English listeners.

[3]It is unknown whether there is a general psychophysical reason for this differential effect of rises and falls in different parts of syllables and words. One might consider the possibility that a pitch rise of fall is prominence-lending only when its course runs parallel to the intensity. This hypothesis can be checked when standardized specifications of accent-lending and boundary-marking pitch movements in other languages are concerned. I know of no research that has looked into this possibility.

[4]Intensity differences may constitute a much stronger cue to stress than has hitherto been thought, if greater intensity is implemented in a realistic way. In the traditional experi-

whether in Focus or out of Focus, are spoken some 15% faster than their accented counterparts (Nooteboom [n.y.]; Eefting and Nooteboom 1993), with a tendency for unstressed syllables to be shortened more than the stressed syllable. Similar effects of accent on overall word duration were found for English (Fowler and Housum 1987) and Indonesian (Van Zanten, this volume).

It follows from the above account that we do not consider accent to be a dichotomy. Rather we take the view that accents can be ordered along a continuous scale. The highest degrees of accent are marked by a pitch movement as well as by temporal means, whereas the lower degrees of accent are only marked by longer duration. Pitch movements are the stronger cues, but duration cues are the more robust correlates of accent.

There are indications that this account is only valid for languages with a so-called dynamic accent, such as English, German and Dutch. Beckman (1986) shows that shifts in accent position within words in Japanese (e.g. /KAta/ 'shoulder' vs. /kaTA/ 'form') are cued by tonal means only, i.e. to the exclusion of temporal and intensity cues. Obviously, much more research is needed for non-European languages before any conclusive position can be taken in this matter.

---

mental literature intensity was manipulated by changing the overall volume of one syllable relative to an other. In human speech production an increase of volume is parallelled by a change in energy distribution over the spectrum: typically energy is increased in the frequency range above 500 Hz, and decreased below 500 Hz. Both increasing intensity and shifting energy from low to high frequency bands creates a stronger stress cue, comparable in strength to duration manipulation (Sluijter and Van Heuven 1993). Differences in vowel quality are the weakest cue to accent (Fry 1965; Rietveld and Koopmans-van Beinum 1986). To complicate matters further, the order of importance among the accent cues may differ from one language to the next, possibly depending on what other phonological contrasts have to be coded in the same acoustic parameters (Berinstein 1979).

5. Introducing the next chapters

The above tutorial was intended to provide a wider perspective on the studies presented in the four chapters that form the body of this book. The chapters all deal with the prosody of Indonesian using phonetic research methods. Let me briefly characterize each research project.

5.1 Acoustic correlates of accents and boundaries in Indonesian (Odé)

This research constitutes a first approximation to the problem of identifying the acoustic factors causing the perception of accents and breaks in the linguistic structure. The methodology is correlational. Listeners are asked to identify the positions of prominent (accented) words and breaks within and between utterances. The more the listener judgements agree, the stronger the assumed accent or prosodic boundary. The perceptual strength is then correlated with selected acoustic properties of the utterances involved (typically the size of pitch movements and the duration of syllables). The claim is, of course, that those acoustic parameters that correlate best with the perceptual prominence and boundary strength are the relevant auditory cues. It should be pointed out, however, that correlational studies may well identify candidates for perceptual cues, but do not establish causal relationships. If we want to conclude that, for instance, a large pitch movement causes the perception of accent, we have to generate two utterances that are exactly the same in all respects except for the presence versus absence of the crucial pitch movement. Such a pair of utterances will never be obtained from any human speaker, since the human speaker will not be able to omit a pitch movement without also changing the temporal and spectral properties of the utterance. Therefore causality can only be established by using synthesized or resynthesized speech (see above).

5.2 Acoustic correlates of stress in Indonesian (Laksman)

This chapter presents a summary of part of Laksman's (1991) dissertation, which was completed at the Université Stendhal in Grenoble, France. Assuming that all the target words investigated have stress on the penultimate syllable, Laksman measured vowel duration (in milliseconds), intensity (in decibels) and maximum pitch value (in hertz, or number of vocal cord vibrations per second) in the final and pre-final syllables. Basically the research answers the question how well the acoustic measurements allow us to determine post hoc whether they were collected for a final (unstressed) syllable or for a pre-final (stressed) syllable. Syllable position (and thereby stress) can be estimated from the acoustic measurements quite accurately when the target words were pronounced in citation form. The separation is more complicated for targets collected as integral parts of a noun phrase. An unexpected result is that Indonesian schwa (*pepet*) does not differ in its prosodic characteristics from other vowels, even though it is claimed to be extrametrical, i.e., invisible to stress rules, in most studies on Indonesian prosody. As in the chapter by Odé, the results are preliminary and heuristic in the sense that they need to be followed up by perceptual experiments. These are currently underway, and will be reported on in the future.

5.3 Temporal correlates of focus and accent in Indonesian (Van Zanten)

In a production study, Van Zanten examines the effects of placing words in and out of focus, thereby generating and removing accent-lending pitch movements. Rather than measuring pitch phenomena she concentrates on the effects of focus on temporal organisation. Systematically varying the length of the target words from one to seven syllables, she tests the claim, derived from earlier research done on Dutch (see above), that accented words are spoken more slowly than unaccented words. Moreover, Van Zanten examines the promising possibility to look at differences in lengthening between stressed and unstressed syllables. If the penultimate syllable is the stressed position, then this syllable should be elongated

more than any of the other syllables in the word. In this sense Van Zanten's study represents yet a third approach to the problem of testing the stress position in Indonesian words.

5.4 Towards an inventory of perceptually relevant pitch movements for Indonesian (Ebing)

This is a straightforward application of the intonation research paradigm developed over the last twenty-five years at the Institute for Perception Research in Eindhoven ('t Hart, Collier and Cohen 1990) to the description of Indonesian intonation. Using spontaneous speech collected from one speaker, the perceptually relevant pitch movements are isolated in the manner outlined briefly above (section 3.3). The research has evolved to the point where a large number of pitch movements were stylized, sorted into a small number of perceptually relevant categories, and given standardized descriptions. Perceptual evaluation of the standard specifications is underway, but will not be reported in the present chapter.

REFERENCES

ABERCROMBIE, D., 1967, *Elements of general phonetics*. Edinburgh: Edinburgh University Press.

BAART, J.L.G., [n.d.], 'Focus, syntax and accent placement.' [N.p.: n.n. Unpublished doctoral dissertation, Leiden University 1987].

BECKMAN, M.E., 1986, *Stress and non-stress accent*. Dordrecht: Foris.

BERINSTEIN, A.E., 1979, *A cross-linguistic study on the perception and production of stress*. Los Angeles: University of California. UCLA Working Papers in Phonetics 47.

BERKOVITZ, R., 1993, 'Lengthening in verb-gapped constructions.' *Phonetica* [submitted].

BEZOOIJEN, R.A.M.G. VAN, 1984, *Characteristics and recognizability of vocal expressions of emotion*. Dordrecht: Foris.

BOLINGER, D.L., 1958, 'A theory of pitch accent in English.' *Word* 14:109-149.

BOSCH, L.F.M. TEN, [n.d.], 'On the structure of vowel systems. Aspects of an extended vowel model using effort and contrast.' [N.p.: n.n. Unpublished doctoral dissertation University of Amsterdam 1991].

BOSCH, L. TEN, D. HERMES, AND R. COLLIER, 1993, 'Automatic classification of intonation movements.' *Annual Research Overview Hearing and Speech Group 1992* (Eindhoven: Institute for Perception Research), pp. 14-15.

CARLSON, R., B. GRANSTRÖM, AND L. NORD, 1992, 'Experiments with emotive speech-acted utterances and synthesized replicas.' In: B.L. Derwing and J.J. Ohala (eds.) *Proceedings of the International Congress of Spoken Language Processing 1992* vol. I:671-674.

COOPER, F.S., A.M. LIBERMAN, AND J.M. BORST, 1951, 'The interconversion of audible and visible patterns as a basis for research in the perception of speech.' *Proceedings of the National Academy of Sciences* 37:318-325.

DAUER, R., 1983, 'Stress-timing and syllable-timing reanalysed.' *Journal of Phonetics* 11:51-62.

DELATTRE, P., 1965, *Comparing the phonetic features of English, German, Spanish, and French.* Berlin: Julius Gross.

EBING, E.F., 1991, 'Pilot experiment contrastieve klemtoon in Bahasa Indonesia.' [N.p.: n.n. Unpublished report, Indonesian Language Development Project ILDEP, Leiden University].

EEFTING, W.Z.F., AND S.G. NOOTEBOOM, 1993, 'Accentuation, information value and word duration. Effects on speech production, naturalness and sentence processing.' In V.J. van Heuven and L.C.W. Pols (eds.), *Analysis and synthesis of speech. Strategic research towards high-quality text-to-speech generation* (Berlin: Mouton de Gruyter), pp. 225-240.

FOWLER, C.A., AND J. HOUSUM, 1987, 'Talkers' signalling of "new" and "old" words in speech and listeners' perception and use of the distinction.' *Journal of Memory and Language* 26:489-504.

FRY, D.B., 1958, 'Experiments in the perception of stress.' *Language and Speech* 1: 126-152.

FRY, D.B., 1965, 'The dependence of stress judgments on vowel formant structure.' In: E. Zwirner, and W. Bethge (eds.), *Proceedings of the 6th International Congress of Phonetic Sciences* (Basel: Karger), pp. 306-311.

FUCHS, A., 1984, '"Deaccenting" and "default accent".' In: H. Richter and D. Gibbon (eds.), *Intonation, accent and rhythm* (Berlin: Walter de Gruyter), pp. 134-164.

GIMSON, A.C., 1969, *An introduction to the pronunciation of English.* London: Edward Arnold.

GOEDEMANS, R., AND V.J. VAN HEUVEN, 1993, 'A perceptual explanation of the weightlessness of the syllable onset.' In: *Proceedings of EUROSPEECH '93* (Berlin), vol. II:1515-1518.

GUSSENHOVEN, C., 1988, 'Adequacy in intonation analysis: the case of Dutch.' In: H. van der Hulst and N. Smith (eds.), *Autosegmental studies on pitch accent* (Dordrecht: Foris), pp. 95-121.

GUSSENHOVEN, C., AND A.C.M. RIETVELD, 1991, 'An experimental evaluation of two nuclear-tone taxonomies.' *Linguistics* 29:423-449.

HART, J. 'T, R. COLLIER, AND A. COHEN, 1990, *A perceptual study of intonation.* Cambridge: Cambridge University Press.

HERMES, D.J., 1988, 'Measurement of pitch by subharmonic summation.' *Journal of the Acoustical Society of America* 83:257-264.

HESS, W., 1983, *Pitch determination of speech signals.* Berlin: Springer.

HEUVEN, V.J. VAN, 1987, 'Stress patterns in Dutch (compound) adjectives. Acoustic measurements and perception data.' *Phonetica* 44:1-12.

HEUVEN, V.J. VAN, 1994, 'What is the smallest prosodic domain?' In: P. Keating (ed), *Papers in Laboratory Phonology III: phonological structure and phonetic form* (London: Cambridge University Press), pp. 76-98.

HOEK, J. VAN DEN, 1993, 'Pitch and duration as determinants of focal accent in Chinese. Interactions with lexical tone.' [N.p.: n.n. Lecture presented at the Second International Conference on Chinese Linguistics, Paris].

KLOSTER-JENSEN, M., 1958, 'Recognition of word tones in whispered speech.' *Word* 14:187-196.

LADEFOGED, P., 1967, *Three areas of experimental phonetics.* London: Oxford University Press.

LAKSMAN, M., [n.d.], 'L'accent en indonésien et son interaction avec l'intonation.' [N.p.: n.n. Unpublished doctoral dissertation, Université Stendhal, Grenoble, 1991].

LANGEWEG, S.J., [n.d.], 'The stress system of Dutch.' [N.p.: n.n. Unpublished doctoral dissertation, Leiden University 1988].

LILJENCRANTS, J, AND B. LINDBLOM, 1972, 'Numerical simulation of vowel quality systems. The role of perceptual contrast.' *Language* 48:839-862.

LINDBLOM, B.E.F., B. LYBERG, AND K. HOLMGREN, 1981, 'Durational patterns of Swedish phonology. Do they reflect short-term motor memory processes?' [N.p.: n.n. Unpublished paper distributed by the Linguistics Club Indiana University, Bloomington IN].

MAYER-EPPLER, W., 1957, 'Realization of prosodic features in whispered speech.' *Journal of the Acoustical Society of America* 29:104-106.

MILLER, J.D., 1961, 'Word tone recognition in Vietnamese whispered speech.' *Word* 17:11-15.

MÖBIUS, B., AND M. PÄTZOLD, 1992, 'F_0 synthesis based on a quantitative model of German intonation.' In: B.L. Derwing and J.J. Ohala (eds.), *Proceedings of the International Conference on Spoken Language Processing 1992* vol. I:361-364.

MORTON, J., AND W. JASSEM, 1965, 'Acoustic correlates of stress.' *Language and Speech* 8:148-158.

NOOTEBOOM, S.G., [n.d.], 'Production and perception of vowel duration. A study of durational properties of vowels in Dutch.' [N.p.: n.n. Unpublished doctoral dissertation, Utrecht University 1972].

OHALA, J.J., 1978, 'Production of tone.' In: V.A. Fromkin (ed.), *Tone. A linguistic survey* (New York: Academic Press), pp. 5-40.

OS, E. DEN, 1985, 'Perception of speech rate in Dutch and Italian utterances.' *Phonetica* 42:124-134.

PORT, R.F., AND M.L. O'DELL, 1985, 'Neutralization of syllable-final voicing in German.' *Journal of Phonetics* 13:433-454.

RIETVELD, A.C.M., AND F.J. KOOPMANS-VAN BEINUM, 1987, 'Vowel reduction and stress.' *Speech Communication* 6:217-230.

SLUIJTER, A.M.C., AND V.J. VAN HEUVEN, 1993, 'Perceptual cues of linguistic stress: intensity revisited.' In: D. House and P. Touati (eds.), *Proceedings of an ESCA workshop on prosody* (Lund: Department of Linguistics and Phonetics, Lund University. Department of Linguistics and Phonetics, Lund University Working Papers 41), pp. 246-249.

THORSEN, N., 1980, 'A study on the perception of sentence intonation. Evidence from Danish.' *Journal of the Acoustical Society of America* 67:1014-1030.

TRUBETSKOY, N.S., 1958. *Grundzüge der Phonologie*. Göttingen: VandenHoeck & Ruprecht.

WISE, C.M., AND L.P-H.CHONG, 1957, 'Intelligibility of whispering in a tone language.' *Journal of Speech and Hearing Disorders* 22:335-338.

ZANTEN, E. VAN, AND V.J. VAN HEUVEN, 1983, 'A phonetic analysis of the Indonesian vowel system. A preliminary acoustic study.' *NUSA, Linguistic Studies of Indonesian and Other Languages in Indonesia* 15:70-80.

ZANTEN, E. VAN, 1989, *Vokal-vokal Bahasa Indonesia. Penelitian akustik dan perseptual*. Jakarta: Balai Pustaka.