AN INVESTIGATION OF ACOUSTIC EVENTS RELATED TO SENTENTIAL STRESS AND PITCH ACCENTS, IN ENGLISH

Paul C. Bagshaw
ATR Interpreting Telephony Research Laboratories
(Visiting from the Centre of Speech Technology Research, University of Edinburgh)

ABSTRACT - An algorithm is described to abstract acoustic parameters of a speech waveform to give a scalar measure of the relative stress and pitch movement of each group of phones which can consist of a single prominence. A method of identify such groups using acoustic information is given. The abstracted parameters are used to locate sentential stress and pitch accents in English speech. These are compared with a hand-labelled prosodic transcription.

I. INTRODUCTION

We wish to label prosodic events in English speech. Prosodic events marked by hand vary considerably from labeller to labeller and may be marked inconsistently within any labellers transcription. (Pickering, et.al, in press) show that two transcribers select the same prosodic label (level, fall, rise, fall-rise, rise-fall, stressed but unaccented or unstressed) for 72% of syllables. This paper presents a method of automaticly transcribing prosodic events with the relative stress of any syllable and the extent of pitch movements being described as a scalar rather than as a discrete level. The method involves a series of abstractions of acoustic parameters which aims to isolate the prosodic variations in duration, energy and fundamental frequency from the microprosodic variations.

The grouping of phones into syllables which can constitute at most a single prominence in the utterance is used as a domain for transcribing prosodic events. The method used to produce the phone groups is described in section II. The prosodic content of each phone group is described by giving it a measure of its relative stress in the utterance and, if the group is accented, the type of pitch movement (level, fall, rise, fall-rise or rise-fall) and the relative extent of the movement. In deducing the relative stress, the prosodic variations of duration, energy and fundamental frequency (F0) are abstracted from the speech waveform acoustics. The formation of a piece-wise F0 contour to remove microprosodic variations is described in section III. The piece-wise units crossing each syllable are abstracted into one of five types of pitch movement. Each syllable is then marked as either prominent (sententially stressed) or not prominent (sententially unstressed), and if it is found to be prominent and pitch salient, it is marked as accented (section IV). These markings are compared with those transcribed by hand.

II. SYLLABIFICATION FROM ACOUSTIC PARAMETERS

An algorithm is described to group phones given by an automatic phonemic segmentation system into syllable sized items based on sonorant energy. The syllabification of speech from acoustic parameters groups phones according to the manner in which the speaker formed the utterance rather than that dictated by a set of phonological rules. The syllabification makes full use of the phone boundary and label information given by auto-segmentation and uses the sonorant energy contour of the utterance to determine their grouping.

The energy contour for a speech waveform (sampled at 20kHz) is calculated from 20ms frames at 5ms intervals so that values are synchronised with the cepstral coefficients and lower three formant frequencies used in the auto-segmentation process. Each frame is passed through a Blackman-Harris window and the frequency bins of an amplitude spectrum (512-point FFT) corresponding to the range 50Hz–2kHz are accumulated. These energy values are expressed in decibels with respect to the maximum frame energy in the utterance to form an utterance-normalised sonorant energy contour. The contour is processed by a three-frame median filter and five-frame hanning window smoother (Rabiner, et.al, 1975) in order to remove small perturbations which arise during frames of speech with low fundamental frequency (typically less than two pitch periods per analysis frame).

All minima in the energy contour are located and form candidates for syllable boundaries. The areas of silence identified by the auto-segmentation are respected and so the minima within the tenure of such areas are believed to be due to variations in background noise. Each boundary between a silence and a phone label is taken as either the beginning or the end of a syllable. The nearest candidate to such a boundary is therefore moved to align with it and all those residing within the silence section are disregarded. The regions between all the remaining energy minima are taken to be potential syllables with a start time given by the nearest left-hand-side minimum's location, and the nearest right-hand-side minimum's location giving the stop time. It is determined if the location of each of these potential syllables overlaps more than 50% of any auto-segmented vowel. If it overlaps more than one vowel segment in this way, then the vowel segment with the maximum sonorant energy is taken to be the nucleus of the syllable. If no such overlap occurs, then it is determined if the location of the potential syllable overlaps more than 50% of one of the possible syllabic consonant segments /l, m, n, r/. Again, if it overlaps more than one of these, the one with the maximum sonorant energy is selected as the syllabic nucleus. If there is insufficient overlap, the region between the minima does not correspond to a syllable unit and either the l.h.s. or r.h.s. minimum is disregarded as a syllable boundary candidate — whichever has the highest energy and does not correspond to a phone/silence boundary. The newly formed region is then taken to be a potential syllable and the process is repeated. The resultant syllabification has boundaries located at positions of minimum sonorant energy in the utterance. These boundaries may be aligned with the auto-segmentation by moving each syllable boundary to the nearest phone boundary.

A database of 453 utterances from the English language ATR conference-registration dialogues has be syllabified using the above algorithm and by using a phonologically based syllabification (Bagshaw & Williams, 1992). There is a large correlation between the two resultant syllabic domains (see table 1). The missing syllable boundaries are due to the occurrences of vowel/vowel boundaries for which there is no valley in the sonorant energy between them. When this case arises, often one of the vowels is a schwa; for example, the phonological syllabification of “my address” as /m aɪ – ə – d r ɛ s/ can be grouped on an acoustic basis as /m aɪ ə – d r ɛ s/. Conversely, extra syllable boundaries occur when the sonorant energy dips within the tenure of the phonologically based syllable at a vowel/vowel boundary or vowel/syllabic consonant boundary; for example the phones in “tour” /t ʊ ɚ/ can be grouped as /t ʊ – ɚ/ on an acoustical basis, and for the word “forms” /f ɔ r m s/ phones are grouped as /f ɔ – r m̩ s/ as its pronunciation tends towards that of “forums”.

Syllabification using acoustic parameters in this manner clusters phones with a vowel or syllabic consonant as its nucleus and containing a single burst of sonorant energy. The duration of the nucleus and the maximum sonorant energy within it are used in determining its relative prominence. The duration and energy variations are mainly attributed to phone type. These parameters are therefore Z-score normalised with respect to the phone type of the nucleus in order to compensate for segmental variations. (Campbell, 1990) uses a similar normalisation but on a phone by phone basis rather than on the basis of syllable nuclei. The mean duration and maximum energy and their population standard deviations are determined for each phone type from a training database of 200 phonemically balanced utterances. The Z-score normalisation of a nuclear phone's duration or intensity simply involves subtracting the mean value and dividing by the population standard deviation for that phone type.

An example of these processes is shown in figure 1. Part (a) shows the speech waveform and its corresponding automatic segmentation using MRPA labels (Edinburgh University's machine-readable phonemic alphabet). Part (b) gives the utterance normalised sonorant energy contour and transcription-aligned syllable boundaries with a MRPA label indicating the phone forming the nucleus. The Z-score normalised duration and energy for each syllable nucleus is given in part (c). These will be discussed further in section IV.

III. THE FORMATION OF A PIECE-WISE F0 CONTOUR

A fundamental frequency (F0) contour produced by a pitch determination algorithm (PDA) can be expected to contain values which are inaccurate, such as instances of pitch octave errors. Any PDA will also make erroneous classifications of sections of speech as voiced or unvoiced. A personal evaluation of a slightly enhanced version of the PDA described in (Medan, et.al, 1991) (which is used in this study) has been found to estimate F0 with consistently less than 1% gross pitch errors and less than 16% of speech classified as voiced or unvoiced incorrectly, when compared with F0 determined from

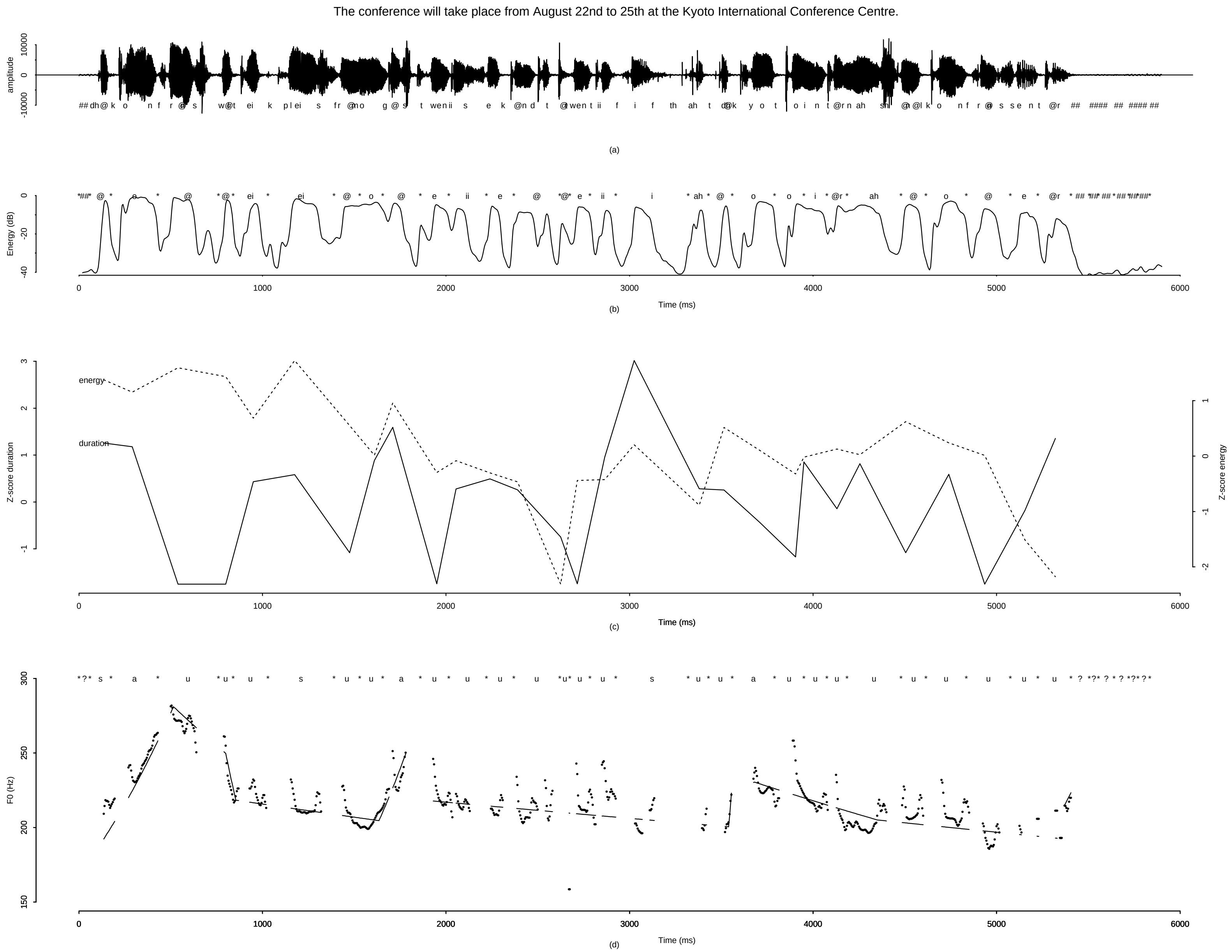

Figure 1: Example of the abstraction of acoustic features related to prosodic events

Table 1: Comparison of Phonologically Based and Acoustically Based Syllabifications

Number of syllables				
from a phonological basis	from the acoustics	Match	Missing	Extra
7299 (100.0%)	7011 (96.1%)	6980 (95.6%)	-319 (4.4%)	+31 (0.4%)

laryngograph data. In order to eliminate the majority of octave errors and reduce microprosodic perturbations, the contour is initially processed by a three-frame median filter and three-frame hanning window smoother (Rabiner, et.al, 1975). The frames of speech analysed by the PDA are in synchronisation with those used in calculating the sonorant energy contour and in the auto-segmentation. The resultant contour is an excellent estimate of the fundamental frequency of the speech waveform, but it does not form a descriptor of utterance intonation alone as microprosodic variations are also present. A process of piece-wise linear stylisation of the contour aims to eliminate such variations.

The algorithm used to perform the stylisation is based on the technique described by (Scheffers, 1988) and incorporates the robust least median of squared residuals regression (LMedR) (Rousseeuw & Leroy, 1987). The F0 values describing the contour (excluding values which equal zero to represent unvoiced speech) are converted to the semitone scale using the relationship $F0_{semitone} = 12\log_2(F0_{hertz}/55)$. Significant turning-points in the F0 contour are located, these points are modified to prevent contour discontinuities other than at the boundaries between unvoiced and voiced speech, and a new contour is generated by interpolating between them.

The following process is used to identify the turning-points. Starting with the first voiced frame, LMedR analysis is applied to a window of w frames corresponding to voiced speech, where w is initially set to 5. The final frame in this window is taken to be a turning-point candidate. The F0 value of the subsequent frame is predicted using the coefficients of the LMedR analysis. If the absolute difference between the actual and predicted F0 values is less than or equal to some level of permitted variation in F0 (1 semitone), then the candidate is not a turning-point, the window length w is incremented to include the next voiced frame, and the above process is repeated. The repetition of this process terminates when the turning-point candidate is the final voiced frame in the F0 contour. Otherwise, when the absolute difference is greater than the permitted F0 variation, either this subsequent F0 value constitutes some type of irregularity in the F0 contour or the candidate could be a true turning-point. To determine which is the case, the F0 value of the next voiced frame is also predicted. If the absolute difference between the predicted and actual values is once again greater than the permitted F0 variation, and this situation arises for all following frames up to either the final voiced frame in the contour or such that the duration of this discontinuity is greater than some minimum permitted level (100ms), which ever occurs first, then the candidate is said to be a true turning-point. Otherwise, the length of the window w is increased to include the first frame for which the absolute difference in actual and predicted F0 values was less then or equal to the permitted variation, but not those for which it was greater, and the LMedR analysis process is repeated. If the candidate was found to be a turning-point and if it corresponds to a voiced frame immediately preceding a frame of unvoiced speech, then the first frame of the next voiced region is also designated as a turning point. This process is then repeated with the length of the window w reset to 5 and the first frame of the window is set to the frame of the most recent turning-point found. The first and final voiced frames of the non-stylised contour are also assigned as turning-points.

In order to ensure that discontinuities in the stylised F0 contour only occur at unvoiced sections of speech, the fundamental frequency at each turning-point of the new contour is determined in a way which depends upon the voicing state of the frames adjacent to it. For any given turning-point (tp) at frame f_{tp} with original fundamental frequency $F0_{tp}$, the LMedR coefficients s_{tp} (slope) and i_{tp} (intercept) of the windowed points preceding the turning-point are known. The modified fundamental frequency $F0'_{tp}$ is given as,

$$F0'_{tp} = \begin{cases} 0.5(s_{tp}.f_{tp} + i_{tp} + s_{tp+1}.f_{tp} + i_{tp+1}) & \text{if frames } f_{tp}-1 \text{ \& } f_{tp}+1 \text{ voiced} \\ s_{tp+1}.f_{tp} + i_{tp+1} & \text{if frame } f_{tp}-1 \text{ unvoiced \& frame } f_{tp}+1 \text{ voiced} \\ s_{tp}.f_{tp} + i_{tp} & \text{if frame } f_{tp}-1 \text{ voiced \& frame } f_{tp}+1 \text{ unvoiced} \\ F0_{tp} & \text{if frames } f_{tp}-1 \text{ \& } f_{tp}+1 \text{ unvoiced} \end{cases} \qquad (1)$$

The new stylised contour is then created by linear interpolation of F0 between each turning-point (f_{tp}, $F0'_{tp}$) and by reseting each frame that is unvoiced in the non-stylised contour to an unvoiced state in the new one. The resultant data is then coverted back to a Hertz scale. An example of this piece-wise stylisation is shown in figure 1(d).

The F0 contours produced for the database of 453 utterances have been stylised using this method. Cepstral resynthesis of the speech from both the original F0 and the piece-wise F0 have been compared by ear. Of these, I felt that 405 (89.4%) contain no perceptual difference in prosodic content.

IV. PROSODIC ABSTRACTION

The piece-wise F0 contour will, for some utterances, contain small units which are erroneous, ie. do not correspond to part of pitch movements. Only those piece-wise units which, at some time, run through any part of a syllable nuclear phone (where F0 estimation is expected to be reliable) are treated as being part of a pitch movement. Moreover, the absolute F0 range of a piece-wise unit is not of interest as it will vary from speaker to speaker, but its extent relative to other units in the utterance is. The relative extent of each piece-wise unit is calculated by first locating a regression line which best fits the contour turning points using LMedR analysis. A by-product of the LMedR is the standard deviation, σ_{LMedR} of the points from the resultant linear model. The absolute F0 at each turning point is then converted by subtracting its modelled value and dividing by the standard deviation, σ_{LMedR}. This effectively compensates for any long term declinative tendency that may be exhibited in the fundamental frequency contour, and expresses the F0 values relative to an utterance dependent datum.

Once the relative extent of each piece-wise unit has been established, they are combined to form pitch movement descriptors. The pitch movements facilitated are level, fall, rise, fall-rise and rise-fall $\{-, \backslash, /, \vee, \wedge\}$, as given by the "British School" of intonational phonology (Crystal, 1969). Each piece-wise unit crossing any part of a syllable nuclear phone is classified as either level, fall, or rise. Let $\mathrm{F0}_{start}$ be the relative F0 height at the start of the piece-wise unit and that at the end of the unit be $\mathrm{F0}_{end}$. The piece-wise unit is classified on the following basis,

$$\text{pitch movement} = \begin{cases} \backslash & \text{if } \mathrm{F0}_{start} - \mathrm{F0}_{end} > 0.75\sigma_{\mathrm{LMedR}} \\ / & \text{if } \mathrm{F0}_{start} - \mathrm{F0}_{end} < -0.75\sigma_{\mathrm{LMedR}} \\ - & \text{otherwise} \end{cases} \qquad (2)$$

When more than one piece-wise unit crosses any particular nucleus, they are combined by initially taking all adjacent units with the same pitch movement classification and joining them into one. A join is made by setting $\mathrm{F0}_{start}$ to that of the first unit, $\mathrm{F0}_{end}$ to that of the second unit, and reclassifying using equation 2. In the database of 453 utterances, consisting of 7299 syllables, there were only 4 syllables for which more than two units remained after this process. These all contained some error which originated in the F0 estimation and are ignored. If there are two remaining units (their classifications must differ), and if either is classified as level $\{-\}$, then they too are joined in the same way. Otherwise, one is a fall $\{\backslash\}$ and the other is a rise $\{/\}$. These are combined to give a single movement classified as either a fall-rise $\{\vee\}$ or rise-fall $\{\wedge\}$ depending on their order, and the relative level at their mid point is kept. Thus, for the fall-rise and rise-fall classifications, the extent of both the onset and coda of the movement are known.

Having established the shape of the pitch movement of each syllable in this way, and with knowledge of the Z-score normalised syllable nucleus duration and intensity measures, we determine if any given syllable is prominent in the utterance (sententially stressed $\{s\}$) and if it is pitch salient (accented $\{a\}$). A syllable is marked as prominent if both its normalised duration and its normalised energy are greater than that of both its nearest neighbours, and if both are greater than 0.75 standard deviations from the mean value. (The value 0.75 is arbitrary.) It is similarly marked if either the normalised duration or normalised energy is the maximum for the utterances. A syllable is marked as accented using a decision filter three pitch movements wide (Hieronymus, 1989).

An example of such prominence detection is illustrated at the top of figure 1(d). The database of 453 utterances has been automatically prosodically marked in this way and compared with those transcribed by hand (see table 2). The transcriptions are equal for 61.6% of the syllables. Of the unstressed $\{u\}$ hand labels marked as either accented or stressed automatically, 216 were syllables with a schwa nucleus. This indicates that the hand transcriber may be marking syllables as sententially stressed only if they can be lexically stressed. There is also a noticeably large number of syllables labelled by hand as accented or stressed that are marked as automatically as unstressed, indicating that the hand labeller

Table 2: Confusion Matrix of Prosodic Transcription by Hand and by Automation

		Automatic Label			
		a	s	u	total
Hand Label	a	566 (7.8%)	177 (2.4%)	1075 (14.7%)	1818 (24.9%)
	s	128 (1.8%)	71 (1.0%)	792 (10.9%)	991 (13.6%)
	u	404 (5.5%)	226 (3.1%)	3860 (52.9%)	4490 (61.5%)
	total	1098 (15.0%)	474 (6.5%)	5727 (78.5%)	7299 (100.0%)

Correct classification rate = 4497/7299 (61.6%)

may be using acoustic parameters other than those described previously in this paper. For example, syllables whose nucleus is "fully articulated" are often marked as stressed by hand. Such measures are currently unavailable to the automatic prosodic transcription algorithm.

V. CONCLUSION

An algorithm to group phones into syllables which can consist of only one prominence has be described. This forms a domain in which to transcribe prosodic events. The domain correlates closely with a phonologically based syllabification. The piece-wise stylisation of a fundamental frequency contour to eliminate micro-prosodic variations in F0 has been further abstracted to form pitch movements for each syllable. The extend of these movements are known relative to other pitch movements in the utterance. The prominence of each syllable has been determined from these parameters and compared with a hand-labelled prosodic transcription. Although the correlation between labels (61.6%) is lower than one would hope, the reason for this may not be because the algorithm performs "poorly" but because it appears that the hand labels transcribe aspects of speech that are not apparent in the waveform acoustics used.

ACKNOWLEDGEMENTS

Thanks to Nick Campbell, Jacqueline Vaissière, Jim Hieronymus and Peter Meer for their valuable assistance and suggestions.

REFERENCES

Bagshaw, P.C. & Williams, B.J. (1992) *Criteria for labelling prosodic aspects of English speech*, In Proc. International Conference on Spoken Language Processing, Banff, Canada (forthcoming).

Campbell, W.N. (1990) *Evidence for a syllable-based model of speech timing*, In Proc. International Conference on Spoken Language Processing, Kobe, Japan, vol.1, 9–12.

Crystal, D. (1969) *Prosodic Systems and Intonation in English*, (Cambridge University Press: Cambridge).

Hieronymus, J.L. (1989) *Automatic sentential vowel stress labelling*, In Proc. European Conference on Speech Communication and Technology (EUROSPEECH-89), Paris, vol.1, 226–229.

Medan, Y., Yair, E. & Chazan. D. (1991) *Super resolution pitch determination of speech signals*, IEEE Trans. Signal Processing, ASSP–39(1), 40–48.

Pickering, B., Williams, B. & Knowles, G. (in press) *Analysis of transcriber differences in the SEC*, In Alderson, P. & Knowles, G. (eds.) Working with Speech, chapter 4, (Longman: London).

Rabiner, L.R., Sambur, M.R. & Schmidt, C.E. (1975) *Applications of non-linear smoothing algorithms to speech processing*, IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP–23(6), 552–557.

Rousseeuw, P.J. & Leroy A.M. (1987) *Robust Regression and Outlier Detection*, (Wiley: New York).

Scheffers, M.T.M. (1988) *Automatic stylization of F0-contours*, In Ainsworth, W.A. & Holmes, J.N. (eds.) Proc. of 7th. FASE Symposium, Edinburgh, vol.3, 981–987.