Appendix C

Publications

The following published papers by the author are directly connected with the research described in this thesis and are appended here.

1. Bagshaw, P.C., & Williams, B.J. (1992). Criteria for labelling prosodic aspects of English speech. *Pages 859–862 of: Proc. International Conference on Spoken Language Processing*, vol. 2. Banff, Canada.

2. Bagshaw, P.C. (1992). An investigation of acoustic events related to sentential stress and pitch accents, in English. *Pages 808–813 of: Proc. 4th. Australian International Conference on Speech Science and Technology.* Brisbane, Australia.

3. Bagshaw, P.C., Hiller, S.M., & Jack, M.A. (1993). Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching. *Pages 1003–1006 of: Proc. 3rd. European Conference on Speech Communication and Technology*, vol. 2. Berlin.

4. Bagshaw, P.C. (1993). An investigation of acoustic events related to sentential stress and pitch accents, in English. *Speech Communication*, **13**(3–4), 333–342.[1]

[1] Reproduced with permission from Elsevier Science Publishers B.V., Amsterdam.

CRITERIA FOR LABELLING PROSODIC ASPECTS OF ENGLISH SPEECH

Paul C. Bagshaw Briony J. Williams

Centre for Speech Technology Research, University of Edinburgh
80 South Bridge, Edinburgh EH1 1HN, Scotland, UK

ABSTRACT

We report a set of labelling criteria which have been developed to label prosodic events in clear, continuous speech, and propose a scheme whereby this information can be transcribed in a machine readable format. We have chosen to annotate prosody in a syllabic domain which is synchronised with a phonemic segmentation. A procedural definition of syllables based on the grouping of phones is presented. The criteria for hand labelling the prominence of each syllable, tone-unit boundaries and the pitch movement associated with each accented syllable, are described. Work to automate this process is presented and experimental results evaluating its performance are included.

I. INTRODUCTION

The need for a large corpus of prosodically labelled English speech is motivated by the use of prosodic events in training speech synthesisers, in automated foreign language pronunciation teaching, and to aid parsers used in speech recognition to disambiguate phonetically similar, but syntactically different utterances.

Speech synthesis requires a mapping from prosodic events to a set of acoustic parameters for their realisation. Parsers and the analysis of language pronunciation, on the other hand, require the reverse mapping to provide descriptors for the acoustic correlates of prosody, and semantic and pragmatic knowledge to be extracted from these correlates. The prosodic labelling of a language corpus must therefore annotate both the linguistically significant features in speech prosody and the inflections of the acoustic parameters.

We aim to transcribe sentential stress (the prominence of syllables in continuous speech) and the pitch movement associated with any accented syllables for such systems. By initially hand labelling these prosodic aspects, a set of acoustic features are sought which will form a mapping for speech synthesis, and at the same time, enable these prosodic events to be labelled automatically given the acoustic features, for parsers and language pronunciation description. The transcription system we propose is intended to be an annotation scheme for linguistically significant prosodic events in English. It is not designed to give a detailed description of every possible inflection in an F0 contour. The set of symbols (see table 1) is designed for use by both a hand transcriber of the prosodic events and for some automated procedure.

The labelling scheme described has been used to transcribe, by hand, prosodic events in a database of 453 utterances from the English language ATR conference-registration dialogues with focus[1]. An acoustic analysis of these labels attempts to establish a correlation between a set of features chosen to characterise the acoustic parameters believed to manifest prosody, and the perceived prosodic events that are transcribed.

Continuous speech is initially segmented into phone units and labelled using a HMM-based automatic segmenter (evaluated in [18]). The phones identified are grouped into syllables. Syllable boundaries are thus synchronised with the phone boundaries. The procedure employed to group phones into syllables is described in section II. Each syllable is labelled by hand as unstressed, stressed (but not accented), stressed and accented (but not nuclear), or as the nuclear accented syllable of a tone-unit. Each syllable immediately preceding a tone-unit boundary is also marked, in order to specify the boundary location. The nuclear accented syllable of a tone-unit is (according to the "British School" of intonational phonology) the final accented syllable in that tone-unit [5]. This definition of nuclear syllables and the criteria used to determine syllable prominence are addressed in section III. Each accented syllable is associated with an additional label that describes the pitch contour movement which marked it. Thus, pitch contour labelling is also synchronised with syllable boundaries. The time location of this movement may occur before, during and subsequent to the domain of the accented syllable. Pitch contour labelling criteria are described in section IV. In section V a set of acoustic features are proposed which we intuitively feel will describe the acoustic correlates of sentential stress. These acoustic features are used to form a tree-based statistical model for a small corpus of hand labelled prosodic events. This methodology is described in section VI. Its application reveals a low correlation between the acoustic features and the events labelled, which poses questions regarding the relationship between the theory and the acoustics of sentential stress. These are discussed in section VII.

II. SYLLABIFICATION

The following procedural definition is used for syllabification.

i) Phones are grouped into syllables on a phonological rather than phonetic basis. Consonantal phones (such as [m, n, l, r, s]) which may result in schwa deletion [10, pp.297–299] [6] and take on the syllabic nucleus, are therefore syllabified as if the vowels were present. Hence, *shortest* in rapid speech is syllabified as /ˈʃ ɔ – t s̩ t/, and *additional* as /ə – ˈd ɪ – ʃ n̩ – l̩/. A glottal stop that may occur before or instead of a word-final stop is treated as an instance of the underlying stop phone, and any glottalised onset to vowels is considered to be part of the vowel.

ii) Syllable boundaries are formed from the boundaries of words considered in isolation. Although in continuous speech, consonants at the end of one word can syllabify with the initial vowel of the following word [13], such resyllabification is not necessary in forming a domain in which to describe prosodic events. Thus, for example, the syllabification of *at all* differs to that of *a tall* even if /t/ is aspirated in both cases and they are phonetically identical. This approach has been adopted because the exact boundaries between syllable nuclei are not of critical im-

[1] The ATR dialogues where spoken by a female bilingual speaker of Japanese and American English.

portance, although identifying the nuclear phone is. Similarly, resyllabification is unnecessary across words that appear to blend together due to vowel deletion, as may be the case in *under a*, which is syllabified as /ˈʌ n – d r̩ – ə/.

iii) The boundaries between syllables are also determined by the presence of a morphological boundary. The boundary between a free morpheme and an inflectional suffix (except -s) or a class-II derivational affix is taken to be a syllable boundary. Thus, *hopeless* is syllabified as /ˈh əʊ p – l ə s/ rather than /ˈh əʊ – p l ə s/; and *uninteresting* is syllabified as /ʌ n – ˈɪ n – t ə – r ɛ s t – ɪ ŋ/ rather than /ʌ – ˈn ɪ n – t ə – r ɛ – s t ɪ ŋ/.

iv) On the basis of English phonotactics, any cluster of phones forming the onset or the coda of a syllable must also be a permissible word-initial or word-final cluster. According to this rule, *extra* may be syllabified as /ˈe k – s t r ə/, /ˈe k s – t r ə/, or /ˈe k s t – r ə/.

v) The 'maximal onset (and minimal coda) principle' [16] [4, pp.10–18] arbitrates between competing analyses. According to the principle, as many consonantal phones as possible form a syllable onset. Using this principle, *extra* would be syllabified as /ˈe k – s t r ə/. However, in cases when alternative boundaries are possible, stressed syllables tend to attract consonants more than unstressed ones, particularly in the case of ambisyllabic consonants such as [s, f] [8, pp.19–23]. When this final criterion is applied, the syllabification adopted for the example becomes /ˈe k s – t r ə/.

III. SENTENTIAL STRESS LABELLING

The salience of each syllable within an utterance is labelled as one of $\{u, s, a, n\}$ (see table 1) on the following basis.

Sententially stressed syllables are those that are perceived as salient due to a prominence of energy and/or duration and/or pitch [7] [12, chap 4] within an utterance. The default (and therefore intonationally unmarked) pitch movement in English is a slight downwards trend in pitch [5, 11]. This movement does not give any intonational prominence to the syllable within the declination, even if that syllable is stressed on the grounds of prominent duration and/or intensity. The same situation occurs if a stressed but unaccented syllable is one in a series of gently rising pitch movements. Where there is no pitch discontinuity, there is no accent [5]. An accented syllable must also be a stressed syllable and an accompanying pitch movement must occur during the accented syllable or on a syllable before or subsequent to the perceived accented syllable [9].

Each tone-unit of an utterance will have one peak of prominence in the form of a nuclear pitch movement. The nuclear accented syllable is the syllable on which the one obligatory pitch movement occurs in a tone-unit. This is traditionally believed to be the final accented syllable in a tone-unit [15]. At present, we make use of this traditional definition.

Tone-unit boundaries are marked by placing a diacritic $\{:\}$ on the label $\{u, s, a, n\}$ of the syllable immediately preceding the boundary. The tone-unit boundaries are identified by two phonetic features [5, pp.204–207]. Firstly, the presence of junctural features, such as slight pauses, final lengthening and rhythmic discontinuities, can signal the end of a tone-unit. However, a pause does not necessarily correspond with a tone-unit boundary in spontaneous speech, particularly in cases of disfluency. Secondly, given that the first prominent syllable, for the majority of tone-units in an utterance, is of approximately the same pitch level [5], the boundary may be signaled by some perceivable pitch change. This change can be either a step up from a falling pitch movement, or a step down from a rising pitch movement. It may be difficult to identify such pitch resets when the tone-unit onset

Table 1: Symbols for Sentential Stress and Pitch Movement Labelling

ASCII†	Symbol	Description
u	$\{u\}$	— Completely unstressed
s	$\{s\}$	— Stressed but unaccented
a	$\{a\}$	— Stressed and accented
n	$\{n\}$	— Nuclear accented
| *pipe*	$\{:\}$	— syllable immediately preceding a tone-unit boundary
\	$\{\backslash\}$	— pitch accent is a fall
/	$\{/\}$	— pitch accent is a rise
v	$\{\vee\}$	— accent is a fall-rise
^ *hat*	$\{\wedge\}$	— accent is a rise-fall
l	$\{-\}$	— level tone
<	$\{\leftarrow\}$	— pitch movement is part of the realisation of an accented syllable to the left of this syllable
>	$\{\rightarrow\}$	— pitch movement is part of the realisation of an accented syllable to the right of this syllable
- *minus*	$\{^-\}$	— the range of the pitch movement is unusually wide (increased)
_ *underscore*	$\{_\}$	— the range of the pitch movement is unusually narrow (decreased)
' *apostrophe*	$\{\triangle\}$	— pitch "peak" or level tone pitch is unusually high
, *comma*	$\{\nabla\}$	— pitch "peak" or level tone pitch is unusually low
[	$\{\vert\ \}$	— initial part of $\{\vee\}$ or $\{\wedge\}$ pitch movement is shallow
]	$\{\ \vert\}$	— final part of $\{\vee\}$ or $\{\wedge\}$ pitch movement is shallow

†The ASCII characters listed are the prosodic labels used in machine readable data.

is low and the final accent of the previous tone-unit ends with a pitch fall, or the onset is high and follows a tone-unit whose final accent ends with a rise in pitch.

IV. PITCH MOVEMENT LABELLING

The pitch contour of an utterance is labelled as a series of pitch movements at (or near) each accented syllable. A pitch movement is either a continuous pitch glide, for example over a long vocalic section of speech, or a discrete pitch jump from one level to another over a series of syllables. Each pitch movement in an utterance is labelled as one of the five categories $\{\backslash, /, \vee, \wedge, -\}$ (see table 1).

A description is associated with each and every syllable labelled as accented (or nuclear accented) to mark the direction of pitch movement on this and any following unaccented syllables. These labels should only be time aligned with an unstressed or an accented (nuclear or otherwise) syllable $\{u, a, n\}$, but not with a stressed (but unaccented) syllable $\{s\}$. (Any stressed syllable corresponding with a time aligned pitch movement label should be marked as an accented syllable.) If the pitch movement is aligned with an unstressed syllable $\{u\}$, a diacritic is applied to the pitch movement label in order to indicate whether the pitch movement is part of the realisation of the nearest accented syllable $\{a, n\}$ to the left $\{\leftarrow\}$ or the nearest one to the right $\{\rightarrow\}$. There may be more than one pitch movement associated with an accented syllable; for example, if there is a rise-fall pitch movement in the realisation of an accented syllable but the rise occurs on a preceding unstressed syllable and the fall occurs on a succeeding unstressed syllable. The uses of these diacritics enable the inflections of the F0 contour to be described while maintaining a transcription of the perceived pitch movement.

Pitch range markings are used to describe the extent of the movement in a pitch glide and the distance between levels of a pitch jump, but not for level tone. If the pitch range is distinctively wider or narrower than expected for a particular contrastive effect, it is marked with a diacritic {‾, _} on the pitch direction labels. Diacritics are also applied to these labels if the "peak" part of a pitch movement (the initial part of a fall {\\}, the final part of a rise {/}, and the mid-section of a fall-rise or rise-fall {V, Λ}) or the pitch of a level tone {−} is unusually high {△} or low {▽} for the particular speaker. In order to describe occurrences of pitch fall-rise and rise-fall with a particularly shallow rise or shallow fall, two further diacritics are included. These are used to represent, for example, fall-shallow rise as {V}.

V. ACOUSTIC FEATURES

A set of acoustic features must be extracted from the raw speech waveform in order to automatically identify syllable prominence and pitch movements. In our preliminary stages of producing an automatic prosodic labelling algorithm, eighteen features are used to describe what are we believe to be the acoustic correlates of stress (duration, intensity and fundamental frequency).

The energy and fundamental frequency of the speech waveform (sampled at 20kHz) are measured for 20ms frames of speech at 5ms intervals so that values are synchronised with the cepstral coefficients and lower three formant frequencies used in the auto-segmentation process. The fundamental frequency (F0) is determined using a slightly enhanced version of the pitch tracker described in [14]. In order to measure the signal energy, each frame is passed through a Blackman-Harris window and an amplitude spectrum is calculated using a 512-point FFT. The total energy for the frequency range of 50Hz–2kHz is determined by summation of the corresponding frequency bins. Each frame energy value is then expressed in decibels with respect to the maximum frame energy for the utterance. This process forms an utterance-normalised sonorant energy contour. Both the raw F0 contour and the energy contour are smoothed using a 3-point non-linear median filter and a 5-point hanning window [17].

The phone given by auto-segmentation which forms the nucleus of a given syllable is identified by the following procedure. The phones in the syllable are split into two groups on the basis of whether or not they are a member of the set of vocalic phones and potentially syllabic consonantal phones (currently, all vowels plus [l, m, n, r]). Each phone is associated with the maximum sonorant energy within its tenure. If there are phones in the syllable which are members of this set, then the one whose associated energy is greatest, is selected as the syllable nucleus. Otherwise, none of the syllable phones are [vowel, l, m, n, r] and the phone with the greatest maximum sonorant energy is selected. The duration associated with any syllable in determining its prominence is the duration of its nuclear phone – this will be referred to as the "syllable duration". Using the duration of the entire syllable or the duration of all consecutive sonorants in the syllable as this measure has not yet been investigated.

Each syllable in an utterance is characterised by the maximum soronant energy within its tenure (syllable energy), its "syllable duration", the maximum F0 value within its tenure, the F0 values at the beginning and at the end of the syllable, and an F0 slope in Hz per second which describes the rate of change in F0 through any voiced regions of the syllable. The syllable energy and "syllable duration" are Z-score normalised to eliminate phone-specific effects [2]. For each phone type, the mean and population standard deviation of the syllable energy/duration is determined. Then, for each token of that phone type, the syllable

Table 2: Confusion Matrix of Sentential Stress Labelling by Hand and by Automation – cyclic exclusion of each utterance during training

		Automatic Label			
		a,n	*s*	*u*	total
Hand Label	*a,n*	889 (12.3%)	72 (1.0%)	849 (11.7%)	1810 (25.0%)
	s	237 (3.3%)	72 (1.0%)	673 (9.3%)	982 (13.6%)
	u	567 (7.8%)	142 (2.0%)	3731 (51.6%)	4440 (61.4%)
	total	1693 (23.4%)	286 (4.0%)	5253 (72.6%)	7232 (100.0%)

Misclassification error rate = 2540/7232 (35.1%)

Table 3: Confusion Matrix of Sentential Stress Labelling by Hand and by Automation – all utterances used during training

		Automatic Label			
		a,n	*s*	*u*	total
Hand Label	*a,n*	1143 (15.8%)	44 (0.6%)	623 (8.6%)	1810 (25.0%)
	s	240 (3.3%)	113 (1.6%)	629 (8.7%)	982 (13.6%)
	u	334 (4.6%)	51 (0.7%)	4055 (56.1%)	4440 (61.4%)
	total	1717 (23.7%)	208 (2.9%)	5307 (73.4%)	7232 (100.0%)

Misclassification error rate = 1921/7232 (26.6%)

energy/duration is normalised by subtracting the mean and dividing by the population standard deviation. Hence, for each syllable, there are six acoustic features extracted – phone-normalised duration, phone-normalised energy, maximum F0, start-time F0, stop-time F0, and F0 slope. In automatically establishing the prominence of any syllable in an utterance, these six features for the current, previous and next syllable are used, giving a total of eighteen features per syllable.

The F0 features are also normalised so that each movement is independent of its absolute F0 values. Our intuition suggests that F0 change is the significant factor, not the absolute F0 values. Normalisation of the nine F0 parameters (the maximum F0, start-time F0, and stop-time F0 for the current, previous and subsequent syllables), is performed by determining the minimum value of these parameters and subtracting it from each. The change in F0 through the syllables is therefore described independently of the absolute height of the F0 movement.

VI. APPLICATION OF A TREE-BASED STATISTICAL MODEL

The sentential stress and pitch movements associated with accented syllables have been hand labelled in the ATR database of 453 utterances using the symbols given in table 1. The prosodic transcription was done by only one labeller.

The automatic prosodic labelling algorithm is still in its infancy and so the acoustic features described in section V are being used only to identify any given syllable in an utterances as either unstressed, stressed or accented (nuclear or otherwise). Distinguishing pitch movement types has not yet been incorporated.

The acoustic features are used as parameters to a tree-based statistical model (using "S" [3]). The model is trained on all but one of the utterances in the database. The tree classifies each hand-transcribed sentential stress label on the basis of the features given. This tree is then used to predict the labels for the utterance that was not included in the training set. These automatically generated labels are compared with those given by hand. This process is repeated in a cyclic fashion for all the utterances and the comparisons are summed. The confusion matrix (table 2) indicates the number of occurrences that each hand-transcribed label is predicted as accented $\{a, n\}$, stressed $\{s\}$ or unstressed $\{u\}$ using this process.

In order to give an indication of the dependency of the automatic labels on the method used, table 3 shows a similar confusion matrix generated when the test utterance is included in the training data.

VII. DISCUSSION

The misclassification error rate of 26.6% is quite promising given that the selection of the acoustic features that have been used is based on intuition. This, however, may not be the only contributing factor to erroneous classifications. It could be that the acoustic features are in fact closely related to the prosodic events labelled, but that the tree-based statistical model is not the most appropriate method to classify these events given the acoustic features (this is supported by the considerable difference between tables 2 & 3). Alternatively, the acoustic features presented could be insufficient to characterise the prosodic events. For example, it is likely that representing F0 movements across a three-syllable window is restrictive, given that such movements can clearly span many or part of syllables. It may be that the labelling scheme is an inadequate system for describing sentential stress and the pitch movements as perceived by the transcriber. This can be illustrated by the fact that sentential stress is not a simple binary distinction between stressed and unstressed. In ambiguous cases, the transcriber uses linguistic knowledge not evident in the acoustics. For example, the syllable in question will be marked as sententially stressed only if it can be lexically stressed. This may lead to every occurrence of schwa being marked as unstressed regardless of the acoustic evidence. With such linguistic knowledge unavailable to the tree-based model, confusions will inevitably arise between the hand labels and automatic labels.

It is most likely that the classification errors are due to some combination of all these factors, although the extent to which any one factor effects the error rate is difficult to determine. The correct-classification rate of 73.4% is, however, close to the percentage of correlating labels between two hand labellers – in the prosodic labelling of the Lancaster/IBM spoken English corpus, transcribers achieved 72% agreement for seven categories of sentential stress labels $\{\backslash, /, \vee, \wedge, -, s, u\}$ and 83% agreement for the categories "accented"/ "stressed"/ "unstressed" [1].

ACKNOWLEDGEMENTS

Thanks to Keith Edwards, Sally Bates, Alex Monaghan, Nick Campbell, Jim Hieronymus, and Bob Ladd for their valuable assistance. This work has been supported by ATR Interpreting Telephony Research Laboratories, Kyoto, Japan.

References

[1] P. Alderson and G. Knowles. *Working with speech.* Longman, London, in press.

[2] W.N. Campbell. Evidence for a syllable-based model of speech timing. In *Proc. International Conference on Spoken Language Processing*, volume 1, pages 9–12, Kobe, Japan, 1990.

[3] L.A. Clark and D. Pregibon. Tree-based models. In J.M. Chambers and T.J. Hastie, editors, *Statistical Models in S*, chapter 9, pages 377–419. Wadsworth & Brooks, Pacific Grove, California, 1992.

[4] E. Couper-Kuhlen. *An Introduction to English Prosody.* Edward Arnold (Publishers) Ltd., London, 1986.

[5] D. Crystal. *Prosodic Systems and Intonation in English.* Cambridge University Press, Cambridge, U.K., 1969.

[6] J.M. Dalby. *Phonetic Structure of Fast Speech in American English.* PhD dissertation, Indiana University Linguistics Club, Bloomington, Indiana, 1986.

[7] D.B. Fry. Duration and intensity as physics correlates of linguistic stress. *Journal of the Acoustical Society of America*, 27(4):765–768, 1955.

[8] E.C. Fudge. *English Word-Stress.* George Allen & Unwin, London, 1984.

[9] E. Gårding and Gerstman. The effect of changes in the location of an intonation peak on sentence stress. *Studia Linguistica*, 14:57–59, 1960.

[10] A.C. Gimson. *An Introduction to the Pronunciation of English.* Edward Arnold, London, second edition, 1970.

[11] D.R. Ladd. Peak features and overall slope. In A. Cutler and D.R. Ladd, editors, *Prosody: Models and Measurements*, chapter 4, pages 39–52. Springer-Verlag, Heidelberg, Germany, 1983.

[12] I. Lehiste. *Suprasegmentals.* The Massachusetts Institute of Technology Press, Cambridge, Massachusetts, 1970.

[13] I. Maddieson. Phonetic cues to syllabification. In V.A. Fromkin, editor, *Phonetic Linguistics (essays in honor of P.Ladefoged).* Academic Press Inc., London, 1985.

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

[15] J.D. O'Connor and G.R. Arnold. *Intonation of Colloquial English.* Longman, London, second edition, 1973.

[16] E. Pulgram. *Syllable, Word, Nexus, Cursus.* Mouton, The Hague, 1970.

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

[18] M.S. Schmidt and G.S. Watson. The evaluation and optimization of automatic speech segmentation. In *Proc. 2nd. European Conference on Speech Communication and Technology*, volume 2, pages 701–704, Genova, Italy, 1991.

Proc. International Conference on Spoken Language Processing, **2**, 859–862.

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).

Proc. 4th. Australian International Conf. on Speech Science & Technology, 808–813.

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

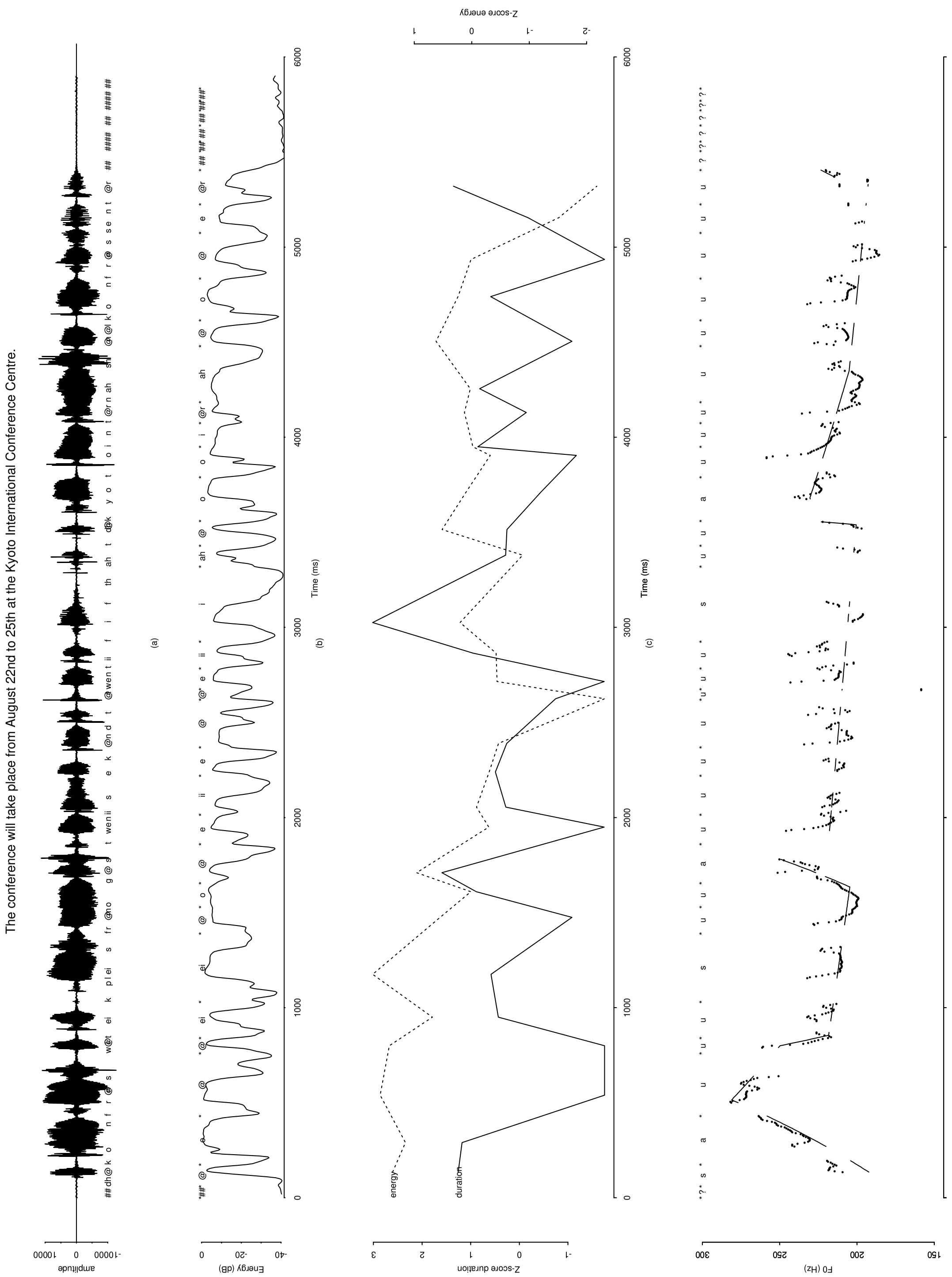

Figure 1: Example of the abstraction of acoustic features related to prosodic events

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,

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

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 $\mathrm{F0}_{\mathrm{start}}$ be the relative F0 height at the start of the piece-wise unit and that at the end of the unit be $\mathrm{F0}_{\mathrm{end}}$. The piece-wise unit is classified on the following basis,

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

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 $\mathrm{F0}_{\mathrm{start}}$ to that of the first unit, $\mathrm{F0}_{\mathrm{end}}$ 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.

ENHANCED PITCH TRACKING AND THE PROCESSING OF F0 CONTOURS FOR COMPUTER AIDED INTONATION TEACHING

P.C. Bagshaw, S.M. Hiller, and M.A. Jack

Centre for Speech Technology Research, University of Edinburgh, 80 South Bridge, Edinburgh, EH1 1HN, Scotland, UK

ABSTRACT

A comparative evaluation of several pitch determination algorithms (PDAs) is presented. Fundamental frequency estimates, $F0$, are compared with laryngeal frequency estimates, Lx. An algorithm is presented which enables Lx contours to be generated from laryngograph data. We seek the most accurate method of $F0$ extraction in order to minimise errors propagating into subsequent prosodic analysis. The super resolution pitch determinator [3] performs well relative to the other PDAs studied. Modifications made to this algorithm are described, which radically reduce the number of gross $F0$ errors and improve the classification of voiced and unvoiced sections of speech. The raw $F0$ contours produced by this enhanced algorithm are processed to form schematised contours used in computer aided intonation teaching. The series of processes used in the schematisation is described.

Keywords: *Pitch tracking, Intonation, Language teaching*

1 INTRODUCTION

The fundamental frequency of speech plays an important role in the prosodic features of stress, rhythm, and intonation. The understanding and appropriate use of prosody is an important component of foreign language learning, for both comprehension and intelligibility. Computer aided teaching of intonation therefore requires the determination of $F0$ as an initial process in the automated assessment of the speech of a non-native student. Determining $F0$ is not a simple task, and many approaches have been reported [2]. The selection of PDAs investigated here covers a range of techniques which use both time domain and frequency domain representations of speech. An evaluation of the algorithms, based on the use of laryngograph data, is described in Section 2. A method of forming a 'reference' contour from laryngograph data is presented. $F0$ contours generated by each PDA are compared with the 'reference' Lx contours. The evaluation shows that the super resolution pitch determinator has the potential to form accurate $F0$ contours from low-pass filtered speech. Errors occurring during $F0$ extraction must be minimised to prevent them from propagating into the prosodic analysis. Enhancements described in Section 3 are made to this algorithm in order to minimise $F0$ errors. The $F0$ contour produced by a PDA is, however, not manipulated solely by linguistic and paralinguistic effects. An $F0$ contour is also affected by segmental content, micro-perturbations, the speaker's anatomy and physiology, and errors involved in its determination from the speech waveform. Such $F0$ variations need to be removed in order to facilitate the comparison of a student's intonation with that of a native speaker. This is performed by a series of post-processes which schematises a raw $F0$ contour and is described in Section 4.

2 EVALUATION OF PITCH DETERMINATION ALGORITHMS

Seven PDAs are investigated. Their selection was influenced by availability, by ease of implementation, and by the desire to examine methods of $F0$ extraction which use radically different techniques.

- Cepstrum pitch determination (CPD) [4].
- Feature-based pitch tracker (FBPT) [6].
- Harmonic product spectrum (HPS) [11] [5].
- Integrated pitch tracking algorithm (IPTA) [12].
- Parallel processing method (PP) [1].
- Super resolution pitch determinator (SRPD) [3].
- Enhanced version of SRPD (*e*SRPD).

The functionality of all of the algorithms is dependent upon certain thresholds and pre-determined parameters, some of which are common across algorithms. In order to set a degree of similarity between the PDAs, all are required to present a computed $F0$ value at 6.4ms intervals. The values are limited to the ranges of 50Hz–250Hz for male speakers and 120Hz–400Hz for female speakers. In cases where a fixed-length analysis frame is required by an algorithm, the frame duration is set to 38.4ms. This duration enables at least two signal periods to reside within the frame for all $F0$ values greater than 52Hz, and allows sufficient data for cepstral and spectral analysis techniques. The speech data is sampled at 20kHz using a 16-bit analogue-to-digital converter. Some of the PDAs require a low-pass filtered version of this data, which is produced by an FIR filter with a -3dB cut-off at 600Hz and rejection greater than -85dB above 700Hz.

2.1 A Laryngeal Frequency Tracker

The 'reference' contours are created from laryngograph data recorded simultaneously with speech by using a simple 'pulse' (Fig. 1) location algorithm and deriving the duration between successive pulses.

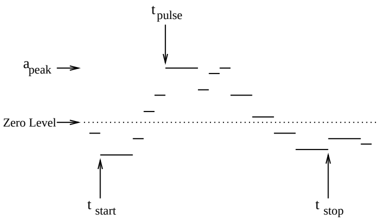

Fig. 1: Laryngograph 'Pulse'

The pulse start time t_{start} is the first sample for which the amplitude is less than zero and less than or equal to the amplitude of following samples. The pulse stop time t_{stop} is the last sample for which the amplitude is less than zero and less than or equal to the amplitude of preceding samples. The pulse width t_{width} is defined as the difference between t_{start} and t_{stop}. The pulse peak amplitude a_{peak} is the maximum amplitude of samples between t_{start} and t_{stop} (always greater than zero.) The pulse instant t_{pulse} is defined as the time of the first of these samples with an amplitude a_{peak}. For laryngograph data sampled at 20kHz, a pulse at t_{pulse} is classed as a marker of the glottal closure instant if the pulse width t_{width} is greater than four samples and the pulse peak amplitude a_{peak} is greater than some arbitrary threshold value. The duration between one pulse instant t^n_{pulse} and the next t^{n+1}_{pulse} is calculated and converted to Hertz. If the value lies within a limited range, it is taken to represent the laryngeal frequency at the time $(t^n_{pulse} + t^{n+1}_{pulse})/2$; otherwise, the duration between the pulses is considered to correspond to an unvoiced region of speech. The Lx limits are $\geq$50Hz for male speakers, and $\geq$120Hz for female speakers. There must be at least three laryngograph pulses in each voiced section. This final restriction is imposed to remove the few errors when a 'pulse' in the laryngograph data is formed by events other than glottal activity.

The accuracy with which Lx can be determined by this method is limited by the time quantisation in sampling the laryngograph signal. Each value of Lx has an error of $F_s/(F_s^2/Lx^2 - 1)$ Hz, where F_s is the sampling frequency.

2.2 Comparative Evaluation

A database containing approximately 5 minutes of speech was used for the evaluation. It was formed from sentences read by one male and one female, and was biased towards utterances containing voiced fricatives, nasals, liquids and glides, since PDAs generally find these difficult to analyse. The quantisation error in determining Lx has a mean of 0.80Hz and population standard deviation of 0.34Hz for the male speaker, and a mean of 3.33Hz and standard deviation of 0.86Hz for the female speaker. This error cannot be compensated for and affects the evaluation results for the various PDAs shown in Table 1. Durations of unvoiced or silent regions incorrectly classified as voiced by a PDA, and durations of voiced sections erroneously classified as unvoiced, are accumulated over all the utterances in the database for each speaker, and expressed as a percentage of the total duration of unvoiced (or silent) speech and voiced speech respectively. The total number of comparisons for which the difference between $F0$ and the reference Lx is higher or lower than 20% of Lx (gross errors) are expressed as a percentage of the total number of comparisons for which $F0$ and Lx represent voiced speech. The population standard deviation (p.s.d.) and the mean, absolute deviation of the Lx and $F0$ contours are given for when both represent voiced speech, and the PDA has not made a gross error.

PDA	*Unvoiced in error (%)*	*Voiced in error (%)*	*Gross errors High (%)*	*Low (%)*	*Absolute deviation (Hz) mean*	*p.s.d.*
CPD	18.11	19.89	4.09	0.64	2.94	3.60
FBPT	3.73	13.90	1.27	0.64	1.86	2.89
HPS	14.11	7.07	5.34	28.15	3.25	3.21
IPTA	9.78	17.45	1.40	0.83	2.67	3.37
PP	7.69	15.82	0.22	1.74	2.64	3.01
SRPD	4.05	15.78	0.62	2.01	1.78	2.46
*e*SRPD	4.63	12.07	0.90	0.56	1.40	1.74
CPD	31.53	22.22	0.61	3.97	6.39	7.61
FBPT	3.61	12.16	0.60	3.55	5.40	7.03
HPS	19.10	21.06	0.46	1.61	4.59	5.31
IPTA	5.70	15.93	0.53	3.12	4.38	5.35
PP	6.15	13.01	0.26	3.20	6.11	6.45
SRPD	2.35	12.16	0.39	5.56	4.14	5.51
*e*SRPD	2.73	9.13	0.43	0.23	4.17	5.13

Table 1: PDA evaluation for male speech (top) and female speech (bottom)

Any PDA producing an $F0$ contour suitable for intonation analysis must perform consistently between male and female speech. The resultant contour must accurately determine voicing so that pitch accents are not left undetected and gross $F0$ errors must be minimal for subsequent processing. CPD and HPS are therefore unsuitable algorithms for the task in hand. Of those remaining, FBPT and SRPD form the best voiced/unvoiced classifications. SRPD would be the most suitable algorithm if the voiced/unvoiced classification performance could be improved and the number of gross ($F0$ too low) errors reduced.

3 ENHANCED SUPER RESOLUTION PITCH DETERMINATOR (*e*SRPD)

The speech is initially low-pass filtered. Frames of data for which an $F0$ estimate is required at the time of sample $s(1)$ are divided into three consecutive segments of n samples,

$$\begin{aligned} x_n &= \{x(i) = s(i-n) \mid i = 1 \; to \; n\} \\ y_n &= \{y(i) = s(i) \mid i = 1 \; to \; n\} \\ z_n &= \{z(i) = s(i+n) \mid i = 1 \; to \; n\} \end{aligned} \quad (1)$$

Frames at the beginning of an utterance, for which x_n is not fully defined, are classified as 'silent'; likewise frames at the end of an utterance, for which y_n and z_n are not fully defined.

The value of n is optimised so that each segment occupies a fundamental period. The optimisation selects a value of n within a limited range, N_{min} to N_{max} samples, which is directly related to the expected range of $F0$ values for a given speaker. The minimum and maximum values of the sample sets $x_{N_{min}}$ and $y_{N_{min}}$ are determined. If the sum of these (absolute) values is less than some preset threshold for either set, then the frame is classified as 'silent'. Otherwise, the coefficient $p_{x,y}(n)$ is determined for the values of n within the limited range in steps of a decimation factor L ($L = 4$ in this investigation).

$$p_{x,y}(n) = \frac{\sum_{j=1}^{\lfloor n/L \rfloor} x(jL).y(jL)}{\sqrt{\sum_{j=1}^{\lfloor n/L \rfloor} x(jL)^2 . \sum_{j=1}^{\lfloor n/L \rfloor} y(jL)^2}} \qquad (2)$$

where $\{n = N_{min} + i.L \mid i = 0, 1, \ldots; N_{min} \leq n \leq N_{max}\}$

$p_{x,y}(n)$ is invalid if the number of zero-crossings in $x_n + y_n$ is less than 4. The locations of local maxima in $p_{x,y}(n)$ with values above an adaptive threshold (as described by Medan *et al.* [3]) form candidates for the optimum value of n. If no candidates for the fundamental period are found, the frame is classified as 'unvoiced'. Otherwise, the frame consists of 'voiced' speech, and a second coefficient, $p_{y,z}(n)$ is determined for all the fundamental period candidates. Those candidates for which $p_{y,z}(n)$ also exceeds the threshold value are given a score of 2, while the others are given a score of only 1. Candidates with a higher score are more likely to represent the true fundamental period. If there are one or more candidates with a score of 2, then all those with a score of only 1 are removed from the list of candidates and ignored. Following this, if there is only one candidate (with a score of either 1 or 2,) the candidate is assumed to be the best estimate of the fundamental period for that frame. Otherwise, the candidates are listed in order of increasing fundamental period. The candidate at the end of this list is selected to represent a fundamental period of n_M, and the m'th candidate a period n_m. Another coefficient, $q(n_m)$, is calculated for each candidate, where $q(n_m)$ is the correlation coefficient between two segments of length n_M spaced n_m apart.

$$q(n_m) = \frac{\sum_{j=1}^{n_M} s(j - n_M).s(j + n_m)}{\sqrt{\sum_{j=1}^{n_M} s(j - n_M)^2 . \sum_{j=1}^{n_M} s(j + n_m)^2}} \qquad (3)$$

The first coefficient $q(n_1)$ is then assumed to be the ideal. If a subsequent $q(n_m)$ exceeds this ideal when multiplied by 0.77 (arbitrary) then it is in turn assumed to be the new ideal. The candidate for which the value of $q(n_m)$ is believed to be ideal is taken as the best estimate for the fundamental period of the frame being analysed.

In the case when there is only one fundamental period candidate with a score of 1 and no candidates with a score of 2, there is only a small probability that the candidate correctly represents the true fundamental period of the frame. If, in such cases, the previous frame was classified as either 'silent' or 'unvoiced', then the $F0$ value describing the current, 'voiced' frame is held until the state of the subsequent frame is known. If this next frame is also not classified as 'voiced', then the frame whose $F0$ value is on hold is an isolated frame which is highly unlikely to be voiced. It is therefore re-classified as 'unvoiced'. Otherwise, the held $F0$ value is assumed to be a sufficiently good $F0$ estimate for that frame.

Biasing is applied to the coefficients $p_{x,y}(n)$ and $p_{y,z}(n)$ for values of n where the fundamental period of a new frame is expected to lie, if the two previously analysed frames were classified as 'voiced,' if the $F0$ value of the previous frame is not being temporarily held, and if the fundamental frequency of the previous frame f_0^{r-1} is less than $\frac{7}{4}$ times the fundamental frequency of its preceding voiced frame f_0^{r-2}, and greater than $\frac{5}{8} f_0^{r-2}$, ie. if it is highly probable that the fundamental period estimate of the previous frame is not erroneous. The fundamental period of the new frame n_0^r is expected to lie within the range of n closest to n_0^{r-1} for which the set of $p_{x,y}(n)$ from the previous frame are greater than zero (Fig. 2). The coefficients $p_{x,y}(n)$ and $p_{y,z}(n)$ are doubled for values of n in this range. This effectively applies a bias on the location of a maxima in the region of the fundamental period for the previous frame to form a candidate for the fundamental period of the current frame. Note, however, that the voiced/unvoiced decision is based on the presence or absence of local maxima in $p_{x,y}(n)$ which exceed the adaptive threshold. The biasing will therefore tend to increase the percentage of unvoiced regions being incorrectly classified as 'voiced'. In order to minimise this undesirable side effect, if the unbiased coefficient $p_{x,y}(n)$ does not exceed the threshold for the candidate believed to the best estimate of the frame fundamental period, then the $F0$ value for that frame is held until the state of the subsequent frame is known. If this next frame is classified as 'silent' or 'unvoiced', the former frame is re-classified as 'unvoiced'.

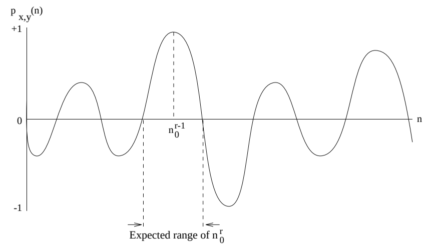

Fig. 2: Example Set of $p_{x,y}(n)$ from Previous Frame

The remainder of the processing to refine the accuracy of the $F0$ estimate is as described by Medan *et al.* [3]. The results shown in Table 1 for the enhanced SRPD algorithm (eSRPD) demonstrate the effects of these modifications on the performance of the algorithm.

4 POST-PROCESSING OF F0

The number of short sections of gross $F0$ errors that occur during $F0$ extraction are reduced by applying a non-linear smoother [7] spanning 17 frames (with a 6.4ms interval between frames). Linear smoothing, using a hanning window, is also applied to remove small perturbations in the contour, and reduce the effect of any remaining quantisation errors. A window of length l takes l consecutive values and weights the n'th value by a factor $h(n)$. The output of the linear smoother is the sum of the weighted values.

$$h(n) = \frac{1}{l+1}\left(1 - \cos\frac{2\pi n}{l+1}\right) \text{ for } 1 \leq n \leq l \qquad (4)$$

A smoothed $F0$ contour may exhibit an overall downward trend in $F0$ during the course of an utterance. Such declination may result in two accents which have the same perceived pitch having different $F0$ values. Compensation for this effect is attempted by initially applying least median of squares regression [10] to all the local maxima and to all the local minima in the $F0$ contour. An average line is taken between the two resultant linear models and used as an estimate of the declination. Declination-compensation is only applied if this average line has a negative slope. The mean and the population standard deviation of the pre-declination-compensated $F0$ contour are retained by using frequency shifting and scaling.

Z-score normalisation [9] is applied to the declination-compensated contour to enable different $F0$ values from different speakers to correspond to the same phonological pitch. An observed $F0$ value is expressed as a multiple of a measure of dispersion relative to the mean $F0$. The normalised $F0$ value is given by,

$$F0_{norm} = \frac{F0_{input} - \overline{F0}}{\sigma} \qquad (5)$$

where $\overline{F0}$ is the long-term mean $F0$ for a given speaker, and σ is the long-term population standard deviation.

Short breaks in continuity of the normalised contour are filled by linear interpolation. Breaks are only filled if they have a duration of less than 80ms and if the jump in $F0_{norm}$ across the break is less than $\sigma/2$. The contour is then smoothed again through a hanning window spanning 17 frames.

This series of processes forms a schematised $F0$ contour which is used in intonation analysis for foreign language teaching [8].

5 CONCLUSIONS

Laryngograph signals, which are recorded simultaneously with speech, have been used to derive 'reference' Lx contours. These have been used to evaluate a selection of pitch determination algorithms. The evaluation shows the enhanced super resolution pitch determinator (eSRPD) to offer improved performance relative to the other PDAs reported in this study, with less than 1.5% gross $F0$ errors and less than 16.7% of speech classified as voiced or unvoiced incorrectly. The new eSRPD algorithm has been shown to perform well independently of a speaker's sex and is unlikely to leave pitch accents undetected. This PDA is therefore the most suitable of those investigated for applications in computer aided intonation teaching where a raw $F0$ contour is smoothed, compensated for the declination of $F0$ over an utterance, normalised for speaker differences in long-term $F0$ level and range, interpolated over small gaps, and smoothed again, to form a schematised $F0$ contour. The schematised contour is used as one of the main inputs to an automatic system for intonation teaching.

REFERENCES

[1] B. Gold and L. Rabiner. Parallel processing techniques for estimating pitch periods of speech in the time domain. *Journal of the Acoustical Society of America*, 46(2, part 2):442–448, 1969.

[2] W.H. Hess. *Pitch Determination of Speech Signals: Algorithms and Devices.* Springer-Verlag, Heidelberg, Germany, 1983.

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

[4] A.M. Noll. Cepstrum pitch determination. *Journal of the Acoustical Society of America*, 41(2):293–309, 1967.

[5] A.M. Noll. *Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate*, volume 19 of *Symposium on Computer Processing in Communication*, pages 779–797. Polytechnic Institute of Brooklyn Microwave Research Institute, New York, 1970.

[6] M.S. Phillips. A feature-based time domain pitch tracker. *Journal of the Acoustical Society of America*, 77:S9–S10(A), 1985.

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

[8] E.J. Rooney, S.M. Hiller, J. Laver, and M.A. Jack. Prosodic features for automated pronunciation improvement in the SPELL system. In *Proc. International Conference on Spoken Language Processing*, volume 1, pages 413–416, Banff, Canada, 1992.

[9] P. Rose. Considerations in the normalisation of the fundamental frequency of linguistic tone. *Speech Communication*, 6(4):343–352, 1987.

[10] P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection.* Wiley, New York, 1987.

[11] M.R. Schroeder. Period histogram and product spectrum: New methods for fundamental frequency measurement. *Journal of the Acoustical Society of America*, 43(4):829–834, 1968.

[12] B.G. Secrest and G.R. Doddington. An integrated pitch tracking algorithm for speech systems. In *Proc. IEEE ICASSP–83*, pages 1352–1355, Boston, 1983.

Speech Communication 13 (1993) 333–342
North-Holland

An investigation of acoustic events related to sentential stress and pitch accents, in English

Paul C. Bagshaw

ATR Interpreting Telephony Research Laboratories Centre of Speech Technology Research, University of Edinburgh, 80 South Bridge, Edinburgh EH1 1HN, Scotland, UK

Received 7 January 1993
Revised 2 June 1993

Abstract. An algorithm is described which abstracts acoustic parameters of a speech waveform to automatically transcribe sentential stress and pitch movements. The waveform acoustics used are duration, energy and fundamental frequency. The abstractions described aim to isolate the prosodically imposed variations in these parameters. A method of syllabification from acoustic parameters is presented. The prominence of each syllable is determined using the automatic process described and the resultant transcription is compared with a hand-labelled prosodic transcription. The agreement level of 61.6% suggests that acoustic parameters other than those already used by the algorithm may be available to the human labeller.

Zusammenfassung. Es wird ein Algorithmus beschrieben, der die akustischen Parameter aus einem Sprachsignal herauskristallisiert, um automatisch die Bewegungen der Satz- und Wortbetonung zu umschreiben. Die verwendeten, akustischen Parameter sind die Zeit, die Energie und die Grundfrequenz. Die beschriebenen Ableitungen dienen zur Isolierung der durch die Betonung erzwungenen Variationen dieser Parameter. Es wird eine Methode der Silbenbildung anhand der akustischen Parameter beschrieben. Die Prominenz jeder Silbe wird anhand des beschriebenen, automatischen Verfahrens bestimmt und die sich daraus ergebende Umschreibung wird mit einer handgeschriebenen Umschreibung der Betonung verglichen. Die Übereinstimmung von 61,6% lässt vermuten, dass zur menschliche Umschreibung andere akustische Parameter verwendet werden als die im Algorithmus verwendeten.

Résumé. On décrit un algorithme qui transforme des paramètres acoustiques du signal de parole en éléments abstraits permettant de transcrire automatiquement l'accent de phrase et les mouvements intonatifs. Les paramètres acoustiques utilisés sont la durée, l'énergie et la fréquence fondamentale. Les traits abstraits qui en sont dérivés visent à isoler, au sein de ces paramètres, les variations prosodiquement imposées. On présente une méthode de syllabification à partir des paramètres acoustiques. La prominence de chaque syllabe est déterminée de façon automatique et la transcription résultante est comparée à une transcription manuelle. Le taux de concordance de 61.6% suggère que l'étiqueteur humain utilise probablement d'autres paramètres que ceux pris en compte par l'algorithme.

Keywords. Prosodic labelling; syllabification; stress; pitch movements.

1. Introduction

We wish to automatically label prosodic events in English speech. Prosodic events annotated by phoneticians vary considerably from labeller to labeller and may be annotated inconsistently within any labeller's 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 automatically transcribing prosodic events with the relative stress of any syllable and the relative height of pitch movements being described in scalar terms rather than as discrete levels. The method involves a series of abstractions of acoustic parameters, which aim to isolate the prosodic variations in duration, energy and fundamental frequency from the microprosodic variations.

Phones are grouped into syllable sized units

0167-6393/93/$06.00
SSDI 0167-6393(93)E0069-8

which are assumed to only ever be perceived as a *single* prominent unit in continuous speech. Prosodic events are transcribed in terms of these units. The method used to produce the phone groups is described in Section 2. 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 height of the movement. In deducing the relative stress, the acoustic parameters, duration, energy and fundamental frequency (F_0), are abstracted to isolate prosodic variations. The formation of a piece-wise F_0 contour to remove microprosodic variations is described in Section 3. The piece-wise sections crossing each syllable are abstracted into one of five types of pitch movement. Each syllable is then labelled as either prominent (sententially stressed) or not prominent (sententially unstressed), and if it is found to be prominent and pitch salient, it is labelled as accented (Section 4). These labels are compared with those transcribed by hand.

2. Syllabification from acoustic parameters

The algorithm described below is used to group phones into syllable sized units using the phone boundary and label information given by an automatic segmentation system, and a sonorant energy contour. The segmentation data gives the location of phones and classifies those phones which are potential syllable nuclei (vowels and syllabic consonants). The sonorant energy contour of the utterance is used to determine their grouping. 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 abstract phonological rules.

2.1. Syllabification algorithm

The energy contour for a speech waveform (sampled at 20 kHz) is calculated from 20 ms frames at 5 ms 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 (Harris, 1978) and the frequency bins of an amplitude spectrum (512-point FFT) corresponding to the range 50 Hz–2 kHz 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 the minima within 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 (energy minimum) to such a boundary is therefore moved to align with the boundary, and all those residing within the area of silence are disregarded. The regions between all the remaining energy minima are taken to be potential syllables, with a start time given by the location of the nearest minimum on the left-hand side, and the location of the nearest minimum on the right-hand side giving the stop time. We then determine whether or not the location of each of these potential syllables overlaps more than 50% of any auto-segmented vowel. If a potential syllable 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 we determine whether or not the location of the potential syllable overlaps more than 50% of one of the possible syllabic consonant segments /l, m, n/ (/r/ is included for American English). Again, if the potential syllable 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 left-hand side or the right-hand side 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.

An example of the syllabification produced using this method is shown in Figure 1 for the word "International", in the phrase "... at the Kyoto International Conference Centre". The upper part shows the speech waveform and its corresponding phonemic transcription (obtained by auto-segmentation) using MRPA labels (a machine-readable phonemic alphabet). Phone boundaries are shown by dotted lines and the continuous lines show the syllable boundaries derived on a phonological basis. The lower part gives the utterance-normalised sonorant energy contour and transcription-aligned syllable boundaries derived using the algorithm described above. The phonologically based syllabification gives five syllables, /ɪn–tɚ–næ–ʃə–nəl/, but the acoustically based syllabification gives only four syllables, /ɪn–tɚ–næʃ–ənəl/. The final syllable, /–ənəl/, may initially appear to illustrate an error in the syllabification algorithm (that of a *missing* syllable boundary). However, the sonorant energy contour shows only one peak of intensity in this section of speech, and it is transcribed by a phonetician as /–nəl/. There is an error in the automatic segmentation of the speech, but not in its subsequent syllabification.

2.2. Evaluation procedure

A database of 453 utterances from the English language ATR conference-registration dialogues

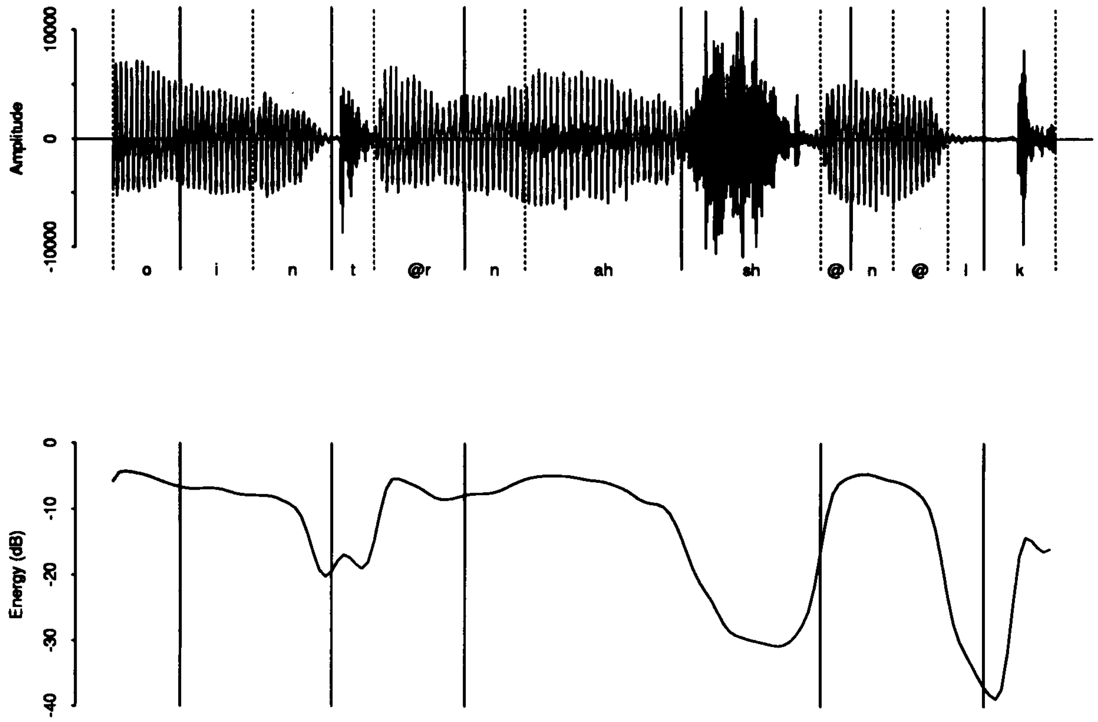

Fig. 1. Syllabification of "International".

(see (Campbell, 1992) for a brief description of this material) has been syllabified using the above algorithm and by hand using a phonologically based syllabification (Bagshaw and Williams, 1992).

In order to compare the human and machine syllabifications, it is necessary to define when they are regarded as "sufficiently similar" and when they are not. At the most stringent level of comparison, a syllable located automatically is regarded as correct only if its boundaries (both at the beginning and the end of the syllable) match exactly with those defined on a phonological basis. This method of comparison is unrepresentative of the algorithm's performance as just one miss-matched boundary would correspond to two miss-matched syllables. Furthermore, the boundaries between syllables vary even between phonological definitions – only, the syllable nuclei are well defined. Therefore, the following procedure is used to evaluate the automatic syllabification algorithm.

Initially assume that all of the syllables located automatically are *extra* syllables – i.e., that each auto-syllable is not the single match of a phonologically defined syllable. Also assume that all of the phonologically defined syllables are *missing* in the automatic syllabification – i.e., that each hand-syllable is not matched by any syllables located automatically. The phonetic transcription of every syllable is known. By definition, there will be at least one vowel and/or syllabic consonant in each syllable (defined either automatically or on a phonological basis). However, the phonetic units which form each syllable nucleus are not stipulated. The comparison proceeds by considering each hand-syllable in turn. Locate the first vowel or syllabic consonant in the hand-syllable. Then consider each auto-syllable from the beginning of the utterance to determine whether or not it overlaps with the potential nucleus. If the auto-syllable does overlap and if it has previously been assumed to be an *extra* syllable, then mark it as not being an *extra* syllable, mark the hand-syllable as *not missing*, and consider the next hand-syllable in the utterance. Otherwise, repeat the search for a matching auto-syllable using another potential nucleus in the hand-syllable until all possible nuclei have been considered.

There is a large level of agreement between the two resultant syllabic domains (see Table 1). The *missing* syllable boundaries are due to the occurrences of vowel/vowel boundaries between which there is no valley in the sonorant energy. 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 phonologically based syllable at a vowel/vowel boundary or vowel/syllabic consonant boundary; for example the phones in "tour" /tʊɚ/ can be grouped as /tʊ–ɚ/ on an acoustic basis, and for the word "forms" /fɔrmz/ phones are grouped as /fɔ–rmz/ when its pronunciation tends towards that of "forums" with schwa deletion, but a fall in energy remaining between /ɔ/ and /m/.

The above algorithm forms groups of phones with a vowel or syllabic consonant as the nucleus of each. There are no vowel/vowel pairs or vowel/syllabic consonant pairs with dips in sonorant energy between them, in any of the groups. The boundaries between groups are positioned at the point of minimum sonorant energy between nuclei (aligned to the nearest phone boundary). It is assumed that each such grouping of phones can only ever be perceived as a single prominent unit in continuous speech. These units have been

Table 1
Comparison of phonologically based and acoustically based syllabifications

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

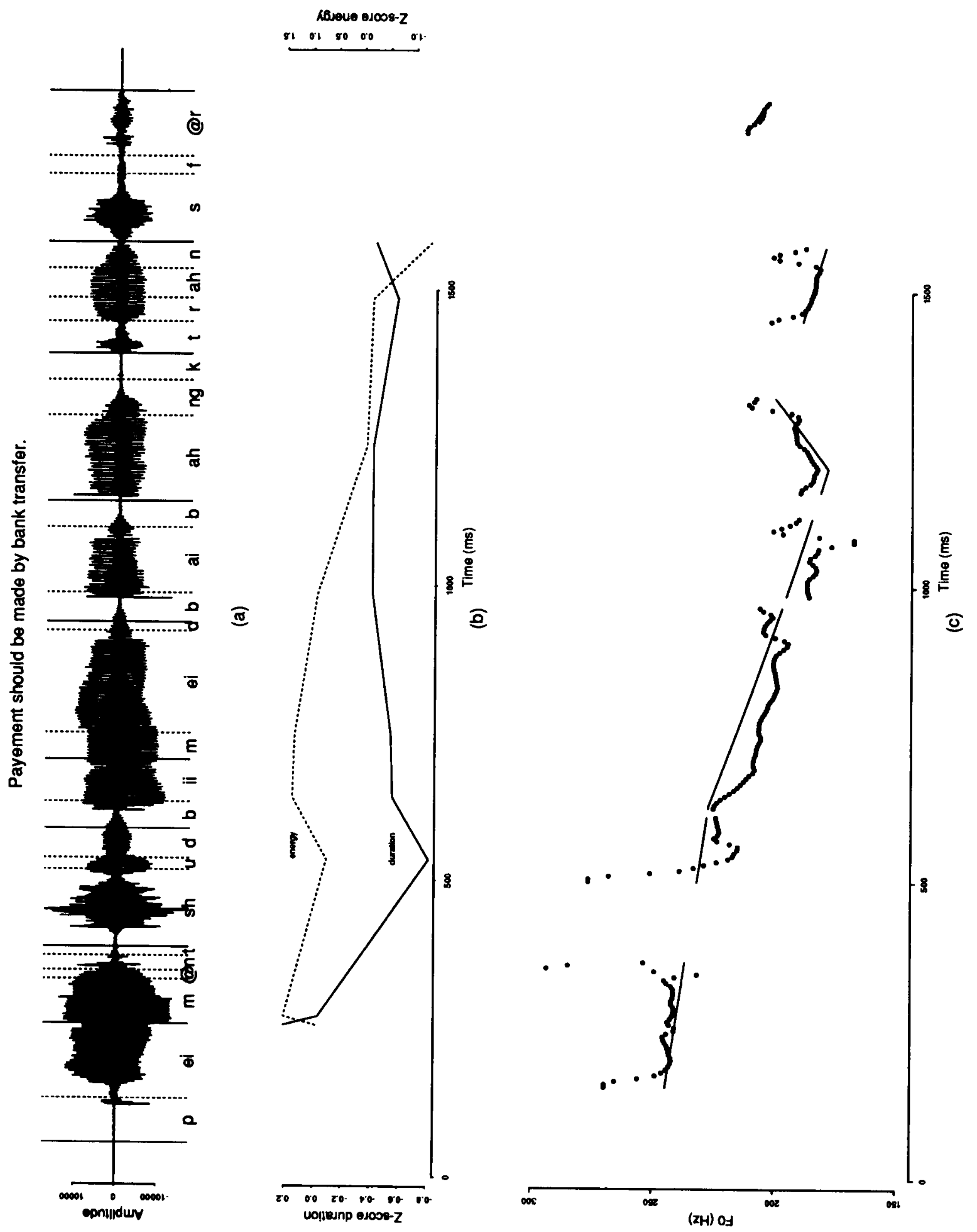

Fig. 2. Example of the abstraction of acoustic features related to prosodic events.

shown to correlate closely with syllables defined using phonologically based rules. They are, therefore, referred to as "syllables."

2.3. Duration and energy measures

Duration and sonorant energy measures are used in determining the prominence of each syllable. The duration and energy variations are mainly attributed to phone type. These parameters are therefore Z-score normalised with respect to the phone type in order to compensate for segmental variations (Campbell, 1990). The mean duration and energy and their population standard deviations are determined for each phone type from a training database of 200 utterances which provide almost total coverage of all permissible demi-syllables in English (Laver et al., 1988). The training database contains speech spoken by the same speaker in the ATR database of conference-registration dialogues and is phonetically transcribed by hand. This data is also used to train the Hidden Markov models for the automatic segmentation system. The Z-score normalisation of a phone's duration and sonorant energy simply involves subtracting the mean value and dividing by the population standard deviation for that phone type. The maximum phone-normalised duration and the maximum phone-normalised sonorant energy within a syllable are used in determining its relative prominence. (Campbell (1990) uses a similar normalisation on a phone by phone basis.)

An example of these processes is shown in Figure 2. Part (a) shows the speech waveform and its corresponding phonemic transcription (obtained by auto-segmentation). Phone boundaries are shown by dotted lines and the continuous lines show the syllable boundaries derived using the syllabification algorithm in Section 2.1. The maximum phone-normalised duration and energy for each syllable are given in part (b). These will be discussed further in Section 4.

3. The formation of a piece-wise F_0 contour

A fundamental frequency (F_0) 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. An enhanced version of the PDA described by Medan et al. (1991) (which is used in this study) has been found to estimate F_0 with consistently less than 1% gross pitch errors and less than 16% of speech classified as voiced or unvoiced incorrectly, when compared with laryngeal frequency estimates, F_x (Bagshaw et al., 1993). In order to eliminate the majority of octave errors and reduce micro-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 (LMedS) (Rousseeuw and Leroy, 1987). The F_0 values describing the contour (excluding values which equal zero to represent unvoiced speech) are converted to the semitone scale using the relationship $F_{0\,\text{semitone}} = 12\log_2(F_{0\,\text{hertz}}/55)$. An account is given below which describes how "significant" turning-points in the F_0 contour are located, then modified to prevent contour discontinuities other than at the boundaries between unvoiced and voiced speech, and how a new, stylised contour is generated by interpolating between the turning-points.

The following process is used to identify the turning-points. Starting with the first voiced frame, LMedS 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 F_0 value of the subsequent frame is predicted using the coefficients of the LMedS analysis. If the absolute difference between the actual and predicted F_0 values is less than or equal to some level of permitted variation in F_0 (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 F_0 contour. Otherwise, when the absolute difference is greater than the permitted F_0 variation, either this subsequent F_0 value constitutes some type of irregularity in the F_0 contour or the candidate could be a true turning-point. To determine which is the case, the F_0 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 F_0 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 (including any intermediate unvoiced frames) is greater than some minimum permitted level (100 ms), 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 F_0 values was less than or equal to the permitted variation, but not those for which it was greater, and the LMedS 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 F_0 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. However, a discontinuity in the stylised contour is allowed within a voiced section of speech if the turning-point is an *outlier* for either of the piece-wise sections it joins. (The detection of outliers is a part of the LMedS analysis.) In such situations, the turning point is treated as being adjacent to an unvoiced frame. For any given turning-point (tp) at frame f_{tp} with original fundamental frequency $F_{0\,tp}$, the LMedS coefficients s_{tp} (slope) and i_{tp} (intercept) of the windowed points preceding the turning-point are known. The modified fundamental frequency $F'_{0\,tp}$ is given as

$$F'_{0\,tp} = \begin{cases} 0.5(s_{tp}f_{tp} + i_{tp} + s_{tp+1}f_{tp} + i_{tp+1}) & \text{if frames } f_{tp}-1 \text{ and } f_{tp}+1 \text{ are voiced,} \\ s_{tp+1}f_{tp} + i_{tp+1} & \text{if frame } f_{tp}-1 \text{ is unvoiced and frame } f_{tp}+1 \text{ is voiced,} \\ s_{tp}f_{tp} + i_{tp} & \text{if frame } f_{tp}-1 \text{ is voiced and frame } f_{tp}+1 \text{ is unvoiced,} \\ F_{0tp} & \text{if frames } f_{tp}-1 \text{ and } f_{tp}+1 \text{ are unvoiced.} \end{cases} \tag{1}$$

The new stylised contour is then created by linear interpolation of F_0 between each turning-point (f_{tp}, $F'_{0\,tp}$) and by resetting 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 2(c). The microprosodic variations in F_0 at boundaries of voiced and unvoiced segments are eliminated whilst retaining the overall melody of the utterance.

The F_0 contours produced for the database of 453 utterances have been stylised using this method. A visual inspection has been made to compare the original F_0 contours with the piece-wise F_0 contours. The database contains 1818 syllables transcribed by hand as accented and 7316 examples of microprosodic perturbations in F_0 at voiced/unvoiced segment boundaries. The piece-wise stylisation erroneously eliminates 356 of the variations in pitch associated with an accent syllable as occurrences of short microprosodic perturbations. 610 of the microprosodic perturbations observed remain in the F_0 contour after stylisation. However, only 72 are later confused as being associated with an accented syllable by the subsequent prosodic abstraction described in Section 4.

4. Prosodic abstraction

Each piece-wise section in the stylised F_0 contour forms a possible pitch movement or part of a movement. Some piece-wise sections do not correspond to part of any pitch movement, such as those which are a direct consequence of erroneous F_0 extraction or stylisation. Thus, only those piece-wise sections which, at some time, run through any part of a syllable nuclear phone (where F_0 estimation is expected to be reliable) are treated as being part of a pitch movement. The piece-wise sections may therefore extend beyond the syllable nucleus but only those crossing the nucleus, in part or in whole, are selected. This approach compromises between using information about the movement of F_0 through vowels alone (which may be limiting for short nuclei), and using the F_0 contour of an entire syllable (where F_0 discontinuity errors may occur).

The absolute F_0 range in an utterance will vary from speaker to speaker and from utterance to utterance. F_0 piece-wise sections are, therefore, normalised for each utterance to give relative F_0 heights. The relative height of each piece-wise section is calculated by first locating a regression line which best fits the contour turning-points using LMedS analysis. A by-product of the LMedS is the standard deviation, σ_{LMedS} of the points from the resultant linear model. The absolute F_0 at each turning-point is then converted by subtracting its modelled value and dividing by the standard deviation, σ_{LMedS}. This effectively compensates for any long term declinative tendency that may be exhibited in the fundamental frequency contour, and expresses the F_0 values relative to an utterance dependent datum.

Once the relative height of each piece-wise section has been established, they are combined to form pitch movement descriptors. The pitch movement descriptors used are level, fall, rise, fall–rise and rise–fall $\{-, \backslash, /, \vee, \wedge\}$ (Crystal, 1969). Each piece-wise section crossing any part of a syllable nucleus is classified as either level, fall or rise. Let $F_{0\,\text{start}}$ be the relative F_0 height at the start of the piece-wise section and that at the end of the section be $F_{0\,\text{end}}$. The piece-wise section is classified on the following basis:

$$\text{pitch movement} = \begin{cases} \backslash & \text{if } F_{0\,\text{start}} - F_{0\,\text{end}} > 0.75\sigma_{\text{LMedS}}, \\ / & \text{if } F_{0\,\text{start}} - F_{0\,\text{end}} < -0.75\sigma_{\text{LMedS}}, \\ - & \text{otherwise}. \end{cases} \quad (2)$$

When more than one piece-wise section crosses any particular nucleus, they are combined by initially taking all adjacent sections with the same pitch movement classification and joining them into one. A join is made by setting $F_{0\,\text{start}}$ to that of the first section, $F_{0\,\text{end}}$ to that of the second section, 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 sections remained after this process. These all contained some error which originated in the F_0 estimation. If there are two remaining sections (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 {\} and the other is a rise {/}. These are combined to give a single movement classified as either a fall–rise {∨} or rise–fall {∧} depending on their order, and the relative level at their mid-point is kept. Thus, for the fall–rise and rise–fall classifications, the relative height of both the onset and coda of the movement are known.

Having established the shape of the pitch movement over each syllable in this way, and with knowledge of the Z-score normalised syllable nucleus duration and energy 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 labelled as prominent if both its normalised duration and its normalised energy are greater than that of both its nearest neighbours (end-points being inherently lower), and if both are greater than 0.75 standard deviations from the mean value. (The value 0.75 is arbitrary.) It is similarly labelled if either the normalised duration or normalised energy measure is the maximum for the utterance. A syllable is labelled as accented using a decision filter three pitch movements wide developed by Hieronymus (1989). This categorisation of the scalar measures of prominence based on duration and energy, to a single discrete level (stressed or unstressed) is necessary only in order to compare the automated annotation with that made by hand.

The database of 453 utterances has been automatically prosodically labelled in this way and compared with those transcribed by hand (see Table 2). The transcriptions are equal for 61.6% of the syllables. Note, however, that the balance between syllables being transcribed automatically as prominent (accented or otherwise) or as not prominent, is dependent upon an arbitrary threshold. It is therefore only possible to observe general trends from the confusion matrix. Of the unstressed {u} hand labels transcribed as either accented or stressed automatically, 216 were syllables with a schwa nucleus. This suggests that the hand transcriber may be annotating syllables as sententially stressed only if they can be lexically stressed, and that the auto-segmentation process may have more confusions in classifying a schwa than for some other vowel. There is also a noticeably large number of syllables labelled by hand as accented or stressed that are labelled automatically as unstressed, suggesting that the hand labeller may be using acoustic parameters other than those described previously in this paper. For example, syllables whose nucleus is "fully articulated" are often labelled as stressed by hand. Such measures are currently unavailable to the automatic prosodic transcription algorithm. These errors may also be accounted for by the need to refine the Hieronymus decision filter used in locating the accented syllables and the need to enhance the F_0 piece-wise stylisation procedure. 428 of the confusions between accented and non-accented syllables arise because of erroneous F_0 stylisation.

5. Conclusions

An algorithm based on the abstraction of acoustic parameters to annotate sentential stress and pitch accents has been presented. A method of grouping phones into syllables which can consist of only one prominence has been described. This forms a domain in which prosodic events are transcribed. There is a large level of agreement (95.6%) between this domain and a phonologically based syllabification. The piece-wise stylisation of a fundamental frequency contour aimed at eliminating microprosodic variations has been illustrated. Piece-wise sections of an F_0 contour are abstracted to describe pitch movements over each syllable. The extend of these movements are known relative to other pitch movements in the utterance and they are used in locating accented syllables. The prominence of each syllable has been determined using the automatic process described and have been compared with a hand-labelled prosodic transcription. An agreement level of 61.6% has been found. This level of agreement is lower than one would expect given the large body of research which has indicated the acoustic correlates of prominence in English to be duration, energy and fundamental frequency. It appears that the hand labels transcribe aspects of speech that are not apparent in the waveform acoustics used. Possible reasons for the differences in the hand and automated transcriptions have been discussed. It is an area of on-going research to establish what contributes to the er-

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%)

(a = accented, s = stressed but unaccented, u = unstressed).
Correct classification rate = 4497/7299 (61.6%).

rors and to determine ways of overcoming them by refining the algorithm.

Acknowledgments

Thanks to Nick Campbell, Jacqueline Vaissière, Jim Hieronymus and Peter Meer for their valuable assistance and suggestions. I am also grateful to Steve Isard and Mervyn Jack for their assistance in the preparation of this document.

References

P.C. Bagshaw and B.J. Williams (1992), "Criteria for labelling prosodic aspects of English speech", *Proc. Internat. Conf. on Spoken Language Processing, Banff, Canada*, Vol. 2, pp. 859–862.

P.C. Bagshaw, S.M. Hiller and M.A. Jack (1993), "Enhanced pitch tracking and the processing of F_0 contours for computer aided intonation teaching", *Proc. European Conf. on Speech Communication and Technology, Berlin*, Vol. 2, pp. 1003–1006.

W.N. Campbell (1990), "Evidence for a syllable-based model of speech timing", *Proc. Internat. Conf. on Spoken Language Processing, Kobe, Japan*, Vol. 1, pp. 9–12.

W.N. Campbell (1992), "Prosodic encoding of English speech", *Proc. Internat. Conf. Spoken Language Processing, Banff, Canada*, Vol. 1, pp. 663–666.

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

F.J. Harris (1978), "On the use of windows for harmonic analysis with the discrete Fourier transform", *Proc. IEEE*, Vol. 66, No. 1, pp. 51–83.

J.L. Hieronymus (1989), "Automatic sentential vowel stress labelling", *Proc. European Conf. on Speech Communication and Technology, Paris*, Vol. 1, pp. 226–229.

J. Laver, C. Bennett, I. Cohan, J. Dalby, D. Davies and M. McAllister (1988), ATR/CSTR speech database project, Status report, No. 1 (Centre for Speech Technology Research; Univ. of Edinburgh, UK).

Y. Medan, E. Yair and D. Chazan (1991), "Super resolution pitch determination of speech signals", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-39, No. 1, pp. 40–48.

B. Pickering, B. Williams and G. Knowles (in press), "Analysis of transcriber differences in the SEC", *Working with Speech*, ed. by P. Alderson and G. Knowles (Longman, London), Chapter 4.

L.R. Rabiner, M.R. Sambur and C.E. Schmidt (1975), "Applications of non-linear smoothing algorithms to speech processing", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-23, No. 6, pp. 552–557.

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

M.T.M. Scheffers (1988), "Automatic stylization of F_0-contours", in: *Proc. 7th FASE Symposium, Edinburgh*, ed. by W.A. Ainsworth and J.N. Holmes, Vol. 3, pp. 981–987.