PROSICE: A spoken English database for prosody research

Mark Huckvale and Alex Chengyu Fang

1. Introduction

Prosody - the study of the intonation, stress and rhythm of speech - is now assuming a greater importance in phonetics, phonology and speech technology than ever before. Once regarded as subservient to studies of segmental structure, it is now being seen as providing the 'framework' which holds different levels of phonetic description together. The recent past has seen novel views of the phonology of intonation (e.g. Pierrehumbert, 1980), a new interest in prosodic phrase structure and prominence (e.g. Liberman and Prince, 1977) and the rise of autosegmental or non-linear accounts of phonetic description which integrate metrical structure with phonetic substance (e.g. Clements and Keyser, 1983). The role of prosody is also changing in speech synthesis and recognition. In speech synthesis, the success of concatenative systems - whereby recorded segments of speech are glued together to make novel utterances - has meant that the key issues have changed from segmental to supra-segmental quality (Klatt, 1987). In speech recognition the increasing emphasis on dialogue systems has meant more research is taking place into the automatic determination of prosodic structure for the purposes of utterance disambiguation (e.g. Wightman and Ostendorf, 1995)

Contemporaneously with the development of prosody research has been the increasing influence of corpus-based research throughout speech technology and experimental phonetics. This has been driven by the huge appetite of current speech recognition research for large quantities of controlled recordings. As an example of this trend in prosody, the prediction of segment durations in speech synthesis is now commonly generated from a multiple regression analysis performed upon a database of transcribed spoken speech (van Santen, 1993).

The combination of these two trends has created a demand for publically-available corpora of spoken recordings for the scientific research and technological application of prosody. In this chapter we look at the requirements and existing corpora and describe a new spoken English database with novel characteristics. Our database is derived from ICE-GB, the British component of the International Corpus of English (ICE).

2. Requirements for Prosody Research

The prosodic elements of main interest (above the level of the syllable) are of three kinds: *prosodic phrase structure*, which describes how words are grouped into prosodic units, and which is strongly related to the syntactic structure of utterances; *phrasal prominence*, which describes how components gain particular emphasis or focus in an utterance, and which is related to semantic and discourse structure; and *intonation markers*, which describe the particular tune types which are related to discourse functions. Hence prosody research requires speech recordings with a

particularly wide range of linguistic annotations - from discourse functions, through semantic and syntactic annotations, to phonetic properties of pitch, duration and loudness. Scientific studies are currently involved in relating higher-level linguistic annotations to phonological models, and phonological models to phonetic realisations, while technological studies are perhaps more concerned with a simple means of determining 'neutral' readings of texts, possibly moving from a shallow syntactic analysis directly to phonetic values of duration and pitch.

While each type of study will have its own specialist requirements, a corpus for prosodic analysis is often associated with high-quality recordings of naturally produced speech, with fundamental frequency tracks, and with phonetic and linguistic annotations[1]. Surprisingly, very few databases of British English currently meet these requirements.

The most relevant corpora that are currently available are:

- SCRIBE contains passages of spontaneous speech that were recorded in an anechoic chamber with a simultaneous laryngograph recording which provides a robust fundamental frequency trace.

- EUROM0 (Fourcin and Gibbon, 1994) contains 4 readings of a 2 minute passage, which have been phonetically annotated to fine detail, and also have laryngographic signals.

- SEC (Knowles, 1993 and 1994) contains transcriptions of broadcast material of about 20,000 words with prosodic annotations. MARSEC (Knowles, 1995) adds to this material phonological intonation marking, wordclass labels and phrase-level syntactic analyses.

- London-Lund Corpus (LLC) of Spoken English (Svartvik and Quirk, 1979), which contains transcribed spoken speech with prosodic markings.

However, these speech databases form a rather strange picture. The EUROM0 corpus has fine phonetic detail and fundamental frequency, but no prosodic or syntactic analysis. The SCRIBE corpus contains spontaneous speech, but it is not annotated at all. The LLC corpus has manually transcribed prosodic annotations, but no quality annotated signals. The SEC corpus does not have time aligned annotations, while the MARSEC corpus, which is otherwise the most sophisticated, has prosodic and grammatical annotations, but is based on a rather limited syntactic analysis of a mixed group of speakers. Furthermore the broadcast medium and the mixture of speakers makes fundamental frequency analysis in MARSEC difficult. Of these, only the LLC corpus has been used extensively for prosodic research so far (Altenberg, 1987; Svartvik 1990).

3. Design Aims

Our objective with the PROSICE corpus has been to combine the best features of existing databases to generate recordings with a complete set of characteristics appropriate for prosody research, within certain resource limitations. This chapter describes the first component, PROSICE-1, which is based on read speech from a single speaker.

The design objectives of PROSICE-1 were:

- High-quality recordings
- Accurate fundamental frequency information
- Genuine spoken texts
- Annotations at both wordclass and syntactic levels linked to signal
- Relevant phonetic annotations

The resource constraints were:

- The use of existing texts and grammatical analysis
- No manual annotation
- The size of a single CD-ROM for distribution

The impracticality of performing manual annotation has meant that we could not provide a phonological level of prosodic annotation. This limitation is potentially very serious in that it precludes the typical division of levels found in scientific research in prosody (e.g. Altenberg, 1987), where syntactic information is projected onto phonological levels of prosody. However, we feel that such phonological level marking is in itself rather limiting. Firstly prosodic annotations, such as those found in the LLC corpus for example, bring with them particular theoretical models of prosody - they mark 'tone units', implying a certain definition of intonational phrasing; and secondly they are subjective, resulting in a level of disagreement between annotators. As far as technological applications are concerned, an intermediate representation between words and phonetic realisation of prosody introduces the requirement for two mapping processes rather than one. Furthermore the trend in speech recognition is for the phonological model to be in the application rather than in the data. It is reasonable to incorporate prosodic phrasing and phrasal prominence as features of a realisation process, but the free parameters in such a system should be calculated from the phonetic evidence directly. Speech recognition systems use phonological segments to model the functions of the sounds in words, but they do not require the training data to be phonologically annotated.

We have based PROSICE-1 on recordings in a form which guarantees the highest acoustic quality, and then indexed them for access via time-aligned annotation at both wordclass and syntactic levels[2]. In section 6 below, we describe the levels of phonetic annotation that we have been able to provide for PROSICE-1. We hope that other workers will be able provide some phonological annotations, particularly to allow comparisons between applications of prosody corpora.

4. Choice of texts

The choice of the texts to record was primarily decided by the immediate applications of PROSICE-1 in the design of text-to-speech systems. Thus a few guidelines in text selection can be easily formulated. First of all, the texts should have their origins in the written genre. Secondly, they should be written mainly for spoken delivery. It is also desirable that texts for this study should be diversified in both style and subject matter. These considerations naturally led us to the scripted texts in the ICE-GB corpus, which are further divided into *broadcast news*, *broadcast talks*, and *non-broadcast speeches*. These three categories range from *S2B-001* to *S2B-050*, covering 50 texts totalling about 100,000 words. For detailed descriptions of the ICE text categories, see Chapter ** in this volume.

Our next question was which of those three categories best serve our purpose. We left out *non-broadcast speeches*, which largely consists of university public lectures, where speakers hesitate, stutter, or correct themselves in a conscious effort to try to 'talk' to the audience, instead of simply 'reading' from their scripts. As a result, such texts, or rather transcriptions, are usually characterised by self-corrections and voiced pauses, and thus are not suitable for this study. We then opted for *broadcast talks*, a sub-category where, unlike *broadcast news*, speakers are the actual composers of their deliveries, where the prosodic phrasing and phrasal prominences have their intended meanings. New recordings of the texts from this group might also enable an interesting comparison with the original recordings.

There are 20 texts in this category, coded sequentially from *S2B-021* to *S2B-040*. Table 1 lists their titles, sources, and dates in text code order. The column labelled *Text Nos*. indicates the number and the sequence of extracts in a composite text[3] . For instance, *S2B-022* has two components; the first one was transcribed from *The River Times* on ITV recorded on 21 June 1991 and the second from *Nature* on BBC 2 TV produced on 5 March 1991. Those without such number indications are non-composite texts.

Table 1: Text titles and sources in broadcast talks

5. Recording

The texts were re-recorded in the anechoic chamber at the Department of Phonetics and Linguistics at UCL. Recording of the sound pressure signal was made using a B&K 2231 microphone positioned 50cm in front of and 10cm to the side of the speaker's mouth. A simultaneous Laryngograph signal was also recorded, using surface electrodes placed on the neck at the level of the thyroid cartilage. The Laryngograph maintains a small current through the neck, and is able to provide an indication of vocal fold contact area (Fourcin and Abberton, 1971). In Figure 1 you can see that the Laryngograph signal is greatest when the vocal fold contact is greatest, immediately after a vocal fold closure.

Figure 1: Sp, Lx, Tx and word annotations

The two signals were amplified, filtered and digitally recorded directly onto computer disk with 16-bit linear quantisation at 20,000 samples/second. The recorded signals were also captured on Digital Audio Tape for archival purposes.

The selection of the 20,000 samples/second sampling rate was chosen as a suitable compromise between audio bandwidth and capacity of the publication CD-ROM. At this sample rate, we could expect to hold 8 texts on one CD-ROM. At the full CD-audio sample rate of 44,100 samples/second, we would only have been able to store half as much. The 20,000 samples/second rate is also very common for spoken language databases (it was used, for example, in the 8-language EUROM1 database).

To prompt the speaker, the printed texts were inserted into soft plastic wallets, which considerably reduced the amount of noise created by page turning. We investigated, but discarded, the possibility of teleprompting via a computer monitor - we felt that this would compromise the speed and prosody of the delivery. Either the prompter must be controlled by the speaker - which adds to his cognitive task - or by an assistant. In the latter case the speaker may be confused by the jerkiness of the scrolling text.

Misreadings and corrections are natural consequences of reading a text out loud. We did not perform any editing on the finished signal, so all deviations from the text are marked and annotated. The speaker read through each text 3 times before the recording - the aim was to make sure that there were no surprises for the speaker, but to avoid a carefully rehearsed intonation (the speaker did not listen to the original broadcast recording).

6. Phonetic Annotation

6.1 Orthographic alignment

To form the necessary connection between signal and syntactic structure, it was desirable to find the temporal location of each word unit in the parse tree. We addressed this problem in a novel way.

From each text as it was actually spoken (including corrections), we collated all the different words (typically 800 different words for a 2000 word text)[4] . The list of words was entered into a tele-prompting system and the speaker re-recorded the words individually in the anechoic chamber. Alongside the speech pressure signal and the Laryngograph signal, a third signal generated by the prompting system was recorded. This prompting signal comprised 'clicks' made by the prompting system as each prompt word was displayed. Thus for each text, we finished with a recording of about half an hour of individual words. These recordings were first made onto an 8-channel digital audio recorder at 44,1000 samples/sec and later transferred to computer disk at 20,000 samples/second.

The presence of the prompting signal combined with the list of words in the order recorded (chosen to be random) allowed the identification of the position of each individual word in this second recording. A small amount of hand-editing was required to tidy up some mistakes made during this recording - less than 2% of words. From the identified individual words, we then constructed a procedure to identify the positions of words in the original recording as follows.

We first formed a frequency-domain representation of both sets of data using a 26-channel filterbank analyser (this was based on an extension of the 19-channel vocoder described by Holmes (1980)). This analyser provides a vector of 26 log energies every 10ms from the speech pressure signals. The actual alignment was made using this representation. This type of analysis is highly suitable for our purposes because a simple Euclidean distance function can be computed between any two spectral vectors which will give a direct measure of the similarity between any two 10ms speech segments. The filterbank output can be seen in a grey-level representation of energy in Figure 2.

Figure 2: pauses, long-section annotations, Fx, spectrographic representation

The alignment of the individual words with the passage can be performed using a procedure called dynamic programming (DP). The aim here is to find a suitable compression/extension of the concatenated individual words such that the total distance between the aligned spectral vectors is at a minimum. The DP method guarantees that the alignment found will have the lowest possible cumulative distance. The actual implementation is a modified version of the 'one-pass' DP algorithm used for speech recognition by Bridle and Brown (1979).

Although the alignment could have been computed over the whole passage, there were good reasons for arbitrarily segmenting the original passage and performing the alignment over smaller regions: (i) less computer time and memory are required, (ii) alignment errors cannot extend beyond one alignment region.

The segmentation of the original passage was performed automatically on the basis of 'major pauses' in the signal. A program located regions of the signal bounded by major pauses and by prompting the operator allowed the word string for each region to be identified. Each region could then be aligned by DP separately. See Figure 2. The criteria for the identification of a major pause is given in the next section.

From the identified regions, text word string, and recordings of the individual words in the string, the DP algorithm finds the best alignment and hence the start and stop time for each word. Each alignment is actually made starting from the middle of the pause preceding the region to the middle of the pause closing the region. This allows the DP algorithm to determine the starting and ending times of the region to within 10ms (whereas the pause-finding algorithm only locates regions within 50ms)

6.2 Pauses

Major pauses are defined as regions of the signal (i) having energy less than 60dB lower than the peak energy for the whole recording, (ii) extending over at least 250ms. In practice this was done by computing the signal energy in 50ms frames over the whole signal and then smoothing this using a triangular weighting filter 250ms wide. The smoothed energy in the signal was then used to set a gap threshold at -60dB relative to peak. Pauses were then regions of 250ms or longer which consisted of frames below this threshold (the use of 250ms is compatible with the definition by Goldman-Eisler (1972)). As will be seen below, such pauses were highly correlated with major clause boundaries.

Minor pauses were taken to be regions of silence of between 50 and 250ms that occurred between words. It is impossible to locate such pauses simply by looking at the speech signal, because stop-gaps (silent regions caused by oral closures prior to a plosive burst) may commonly be of such durations. Thus pauses of 100ms or greater are very common *inside* words. We addressed the problem of locating minor pauses by integrating it into the word alignment procedure. Using the one-pass DP algorithm, it is a simple matter to allow optional pauses to be inserted between words if this helps the overall alignment score. Thus only in circumstances - between words - where a small silent gap would aid alignment is a minor gap inserted. A typical 2000-word text provided 260 major gaps and 65 minor gaps with this procedure.

6.3 Fundamental frequency contour

For the determination of the pitch of the signal, the Laryngograph output was processed in the following manner.

The Laryngograph signal was filtered with a zero-phase non-recursive digital band-pass filter operating between 40 and 2000Hz to remove low-frequency energy due to gross larynx movement and residual high-frequency noise. A peak-following algorithm - which was allowed to rise as fast as the signal but decay at a restricted rate - was used to locate and measure individual pitch periods (see Figure 1).

The pitch period markers give an accurate and instantaneous measurement of fundamental frequency (Fx). However, since the onsets and offsets of voicing can produce false period values (as the vocal folds start or cease vibration) the fundamental frequency trace calculated from the pitch period data was subjected to a five-point median filter (equivalent to 25ms). This removed glitches without compromising the measured frequency values at other times. A contour is shown in Figure 2.

6.4 Pitch accents

While the fundamental frequency contour provides a value of the pitch of the signal every 5ms, it is also useful to have some characterisation of the pitch through an entire word. This would make it easier to relate the higher level labelling to phrasal prominence.

The aim is to determine a set of parameters which are related to the size and nature of any pitch accent occurring within a word. The parameters chosen are: (i) mean Fx through word, (ii) the standard deviation of Fx, (iii) the mean rate of change of Fx, (iv) the standard deviation of (iii), and (v) the mean acceleration of Fx. Together these parameters identify words that have unnaturally high or low Fx or have large changes in Fx, they identify whether words have largely rising or falling contours and whether these are rapid or extended in time, they also differentiate between a rise-fall contour and a fall-rise contour. The measurements are only made on the voiced regions of a word as located by the orthographic alignment.

The fundamental frequency units used were in *cents*, one hundredths of an octave, with the median Fx value for the passage taken to be zero.

7. Aligning syntactic analysis to speech signals

The word alignment process generates a file which indexes the graphic word to its temporal location in the acoustic signal file. The next task was to align these references to the syntactic tree structure of the sentence[5] , a necessary integration that enables cross references between the acoustic recording, the graphic word, and the corresponding syntactic category and function. Figure 3 shows a sample of the two separate files before the alignment.

Figure 3: Sample temporal and syntactic files

In the time-word file, upper-case words bound within colons (:) are not lexical items. They are special mark-up symbols that indicate pauses, coughs, etc. So far, we have found the following indications necessary:

:CLICK: Discernible noise.
:COUGH: Cough.
:ERR: Misreading of the original script during recording. Each item in the misreading is preceded by this symbol.
:GAP: Major pauses.
:MGAP: Minor pauses.
:?: Unclear utterance.

A computer program was written in C to align and merge the two files automatically. The final output takes the format as shown in Figure 4. Each aligned syntactic tree is preceded by three numbers indicating: serial number, number of nodes, and number of leaves. For instance, we can read in Figure 4 that this tree is the fifth in the whole text (<#5>), that there are 14 nodes altogether, and that there are 6 leaf nodes. Each leaf node is annotated with a digit indicating the number of lexical items occupying this node. This indication is especially helpful in the case of compound nouns (*Judge Meyer, The Hague*), complex prepositions *(in accordance with, by means of*), certain conjunctions (*rather than, as if*), certain marginal modal auxiliaries (*need to, ought to*), and semi-auxiliaries (*appear to, be about to*), which are treated as single units and ditto-tagged according to the ICE wordclass annotation scheme (cf Chapter ** in this volume). Each item in the leaf node is followed by a time value that indicates its position in the actual digital recording and also indexes the other phonetic annotations.

Figure 4: Sample temporally and syntactically aligned file
----

8. Correlations between syntactic categories and pauses

Pauses in speech are by no means of random occurrence, since they tend to divide up the stream of speech into grammatically and lexically relevant chunks (Quirk et al, 1985:II.19). Extensive studies have been carried out to investigate their correlation with syntactic units. It has been established that pauses in read speech are mainly due to the influence of graphic arrangements in the text (Stenström, 1990:211). It has also been established that pauses specifically mark the sentence and the clause (Goldman-Eisler, 1972). There has also been an emphasis on the use of pauses to identify 'information units' larger than the clause (Brotherton, 1979; Beattie, 1983; and Chafe, 1987; cf. Stenström, 1990:213). As the only automatically measurable property of read speech that strongly correlates with syntactic units, pauses have been recommended as an anchor-point for the integration of orthographic transcriptions and the digitised acoustic file to create a time-aligned speech database (Knowles, 1994:97).

In this section, we describe an investigation of pauses in PROSICE-1. The primary aim of this experiment was to test if correlation with syntactic structures was shown by the automatically determined pause annotations. A secondary aim was to find out to what extent clause elements apart

from the sentence and the clause coincide with pauses. The clause elements considered in this experiment include five phrases - adjective phrase, adverb phrase, noun phrase, prepositional phrase and verb phrase - in addition to the clause and the sentence.

The experiment was based on the first re-recorded text, *S2B-025*, which is entitled *For He Is an Englishman*, broadcast on BBC Radio 4, on 5 November 1990. This text has 2,014 running words (tokens), or 843 different word forms (types), which fall into 132 parsing units (PU)[6] . We can thus calculate that there are roughly 15.25 words per parsing unit. Except for three (one interjection and two noun phrases), all parsing units correspond to the sentence in terms of structure[7] . The actual recording of this text covers a duration of 11 minutes or 660 seconds, representing a speech rate of 3.05 words per second, or 183.09 words per minute[8] . Altogether, there are 309 pauses, both major and minor, yielding an average of approximately 6.5 words per pause.

Table 2 lists the frequency distribution of pauses in relation to the syntactic categories. The first column, *Category*, lists the types of syntactic categories, ie, sentence, clause, etc. *Frequency* is divided into two columns: the first column indicates the observed frequency of a certain category and the second column the frequency of that category with initiating pauses. *Percentage* also has two columns, the first one displaying the percentage of a certain category that has initiating pauses and the second indicating the proportion of pauses falling into that category. Thus, we can read, for instance, that there are altogether 161 adjective phrases, 8 of which have initiating pauses. These eight occurrences occupy 5% of the total number of adjective phrases and 2.6% of the total number of pauses.

Table 2: Frequency distribution of pauses among major syntactic categories

We observe that of all the pauses in the text, only 18 (5.8%) coincided with the start of elements other than the seven categories, or 94.2% of all the pauses correspond to syntactic structures. The results confirm that pauses are in general a reliable indication of the start of a canonical syntactic structure. The predominant correlation is between the sentence start and the pause. All the sentence starts in the recording correlate with pauses, accounting for nearly half (48.5%) of all the pauses. This result strongly confirms the finding of previous studies that sentences are marked by their temporal cohesion in reading (Goldman-Eisler, 1972:103), but not clauses; of the 106 clauses in our text, only 38 (35.8%) are found co-occurring with pauses[9] . To quote Goldman-Eisler (ibid., 110),

> the speaker who thinks on his feet organizes his message in highly cohesive sentence units with a clear hierarchical structure whereby constituent clauses are temporally integrated into the sentence frame ... to a far greater extent than sentences are into the whole discourse.

Clause elements, accordingly, show an even greater integration into the sentence. Of the five phrases, the verb phrase has the highest correlation with pauses (13.9%), with the noun phrase

having only 4.8%. Nonetheless, these five phrases also reveal a varying degree of temporal integration; the verb phrase and the prepositional phrase both demonstrate an integration stronger than the sentence and the clause but looser than the other three phrases. Considering that both of these two phrases generally require nominal complementation and are thus more complex than the other three, we could hypothesize that while sentence demarcation remains the more prominent function of pauses in read speech, some rhetoric pauses are needed with complex phrase structures that demand complements. Manual inspection of the text revealed that verb phrases with initiating pauses typically begin with the word *be*[10] , which occurred 16 times out of 38. Here are a few examples.

[1] *:GAP: A two-thirds majority but no appeal :GAP: <u>was</u> needed for a lamp post in a street* [S2B-025-47]

[2] *:GAP: One of the difficulties :GAP: <u>was</u> that there was no reliable book on the subject* [S2B-025-51]

[3] *:GAP: One :MGAP: by-product of this essentially rural work :GAP: <u>was</u> that with the late Robert Ayckman :GAP: I became and remain :GAP: a waterways revivalist* [S2B-025-57]

[4] *:GAP: The pre-nineteen :MGAP: fourteen Canal Commissions' admirable report :GAP: <u>was</u> wholly ignored* [S2B-025-62]

These four examples demonstrate two characteristics that probably explain the occurrences of pauses. The first is a heavy subject, as in Examples [1] and [4]. The second characteristic is a heavy subject complement as in [2] and [3], where *be* introduces *that*-clauses as subject complements. [3] actually has both a heavy subject (*one by-product of this essentially rural work*) and a heavy subject complement (*that with the late Robert Ayckman I became and remain a waterways revivalist*). Whether these two observations can be generalised needs to be tested on more data. Preposition phrases with initiating pauses, on the other hand, are typically those that can be termed as complex prepositions ([5] and [6]) and complex conjunctions as in [7]:

[5] *:GAP: The latter feared the boroughs :GAP: <u>because of</u> their masses of urban voters :GAP: and were looking for allies* [S2B-025-28]

[6] *:GAP: :COUGH: :GAP: Commons are common :MGAP: only to certain individuals who happen to share rights in the land :GAP: <u>in common with</u> some other individuals* [S2B-025-76]

[7] *:GAP: These ways were left out of the act :GAP: <u>with the result that</u> astute gravel merchants and estate developers :GAP: exploit loopholes in the law :GAP: to discommon and ravage* [S2B-025-124]

The relatively high proportion of adverb phrases with initiating pauses could be conveniently explained by the use of adverbs as sentential disjuncts. Disjuncts 'have a superior role as compared with the sentence elements; they are syntactically more detached and in some respects "superordinate", in that they seem to have a scope that extends over the sentence as a whole' (Quirk et al., 1985:8.121).

In conclusion, we can say that this experiment confirmed pauses to be a reliable demarcation of the conventional sentence in fluent read speech. Though often similar to the sentence in terms of syntactic structure, clauses are much less characterised by initiating pauses and thus generally demonstrate subordination to the sentence. We would therefore expect that phrases, which are clause elements, would co-occur with pauses much less often than do clauses. However, nearly half of the pauses (46%) fell on the five phrases considered in the experiment. While we may account for some of their occurrences in the one text, a detailed separate study with substantial data is necessary for any sound generalisation.

9. Further work

PROSICE-1 is a novel database for prosody research. Its level and quality of syntactic analysis go beyond other spoken corpora. It has very high-quality audio speech signals, very accurate fundamental frequency information derived from the simultaneous Laryngograph recording, and accurate word annotations derived from whole-word alignment and a number of other phonetic annotations. It is also a corpus of a good size, comprising over 16,000 words and nearly two hours of speech. We intend to make the whole database available on CD-ROM at nominal cost to researchers.

We hope to develop PROSICE in a number of ways. Firstly, we expect to add more annotation to the database by, for example, relating to phrasal prominence the semantic features encoded in the ICE grammatical annotation scheme. Secondly we hope to make new recordings of spontaneous speech and analyse and present them in a similar way. Thirdly we hope that other workers will join with us in providing further annotations: phonological annotations such as the ToBI system (Silverman *et al*, 1992), or phonetic segment annotations which will allow the identification of segment duration changes adjacent to prosodic boundaries.

10. Notes

[1] The database should also be accompanied by a clear statement about Intellectual Property Rights. It is still very common to see private databases unavailable to the scientific world at large.

[2] For detailed descriptions of ICE wordclass and syntactic annotation schemes, see relevant chapters in this volume.

[3] Composite texts are those comprising separate extracts from a single or different sources.

[4] A word here was defined as a string having a word-class label in the syntactic analysis.

[5] For a description of the ICE syntactic parsing scheme, see Chapter ** in this volume.

[6] An ICE parsing unit roughly correponds to the conventional notion of *sentence*, though there are cases where a parsing unit is only a syntactic phrase.

[7] We distinguish the sentence from the clause. A sentence consists of either one sentence or a number of coordinated sentences. In this treatment, clauses - whether finite or nonfinite - are always subordinate.

[8] In comparison, Lumley (1933) noted an average rate of 1.8 words per second for politicians, 2.7 words per second for educators, 2.9 words per second for preachers, and 3.2 words per second for reporters (cf. Altenberg 1987:16). Altenberg (ibid., 22) found an average of 2.4 words per second for the LLC Corpus.

[9] In Stenström 1990, nearly half (48%) of the clauses in spontaneous speech have initiating pauses, revealing the effect of the cognitive process of planning and selecting.

[10] The word *be* is grouped into various uses by the ICE tagging scheme. The four examples include *be* as the passive auxiliary ([1] and [4]) and the copula ([2] and [3]).

References

Altenberg, B. (1987). *Prosodic Patterns in Spoken English - Studies in the Correlation between Prosody and Grammar for Text-to-Speech Conversion*. Lund: Lund University Press.

Beattie, G. (1983). *Talk*. Milton Keynes: Open University Press.

Bridle, J.S., and M.D. Brown (1979). Connected word recognition using whole word templates. In *Proceedings of the Institute of Acoustics Autumn Conference, Windermere November 1979*, pp 25-28.

Brotherton, P. (1979). Speaking and not speaking: Process for translating ideas into speech. In A.W. Siegman and S. Feldstein (Eds), pp 179-209.

Chafe, W. (1987). Cognitive constraints on information flow. In R. Tomlin (Ed), pp 21-51.

Clements, G.N., and S.J. Keyser (1983). *CV Phonology: A Generative Theory of the Syllable*. MIT: MIT Press.

Fourcin, A.J., and E. Abberton (1971). First applications of a new Laryngograph. In *Medical and Biological Illustration 21*, pp172-182.

Fourcin, A.J., and D. Gibbon (1994). Spoken language assessment in the European context. In *Literary and Linguistic Computing 9*, pp 79-86.

Goldman-Eisler, F. (1972). Pauses, clauses, sentences. In *Language and Speech 15*, pp 103-113.

Holmes, J.N. (1980). The JSRU Vocoder. In *IEE Proceedings 127*, Part F., No. 1.

Klatt, D.H. (1987). Review of text-to-speech conversion for English. In *JASA 82*, pp 737-793.

Knowles, G. (1993). From text to waveform: Converting the Lancaster/IBM spoken English corpus into a speech database. In C. Souter and E. Atwell (Eds), pp 47-58.

Knowles, G. (1994). Annotating large speech corpora: building on the experience of Marsec. In *Journal of Linguistics*, 13, pp 87-98.

Knowles, G. (1995). Recycling an old corpus: converting the SEC into the MARSEC database. In G. N. Leech, G. Myers and J. A. Thomas (Eds), pp 208-219.

Leech, G.N., G. Myers and J.A. Thomas (Eds) (1995). *Spoken English on Computer: Transcription and Mark-up*. London: Longman.

Liberman, M.Y., and A. Prince (1977). On stress and linguistic rhythm. In *Linguistic Inquiry 8*, pp 249-336.

Lumley, F.H. (1933). Rates of speech in radio speaking. In *Quarterley Journal of Speech, 8*, pp 393-403.

Pierrehumbert, J.B. (1980). *The Phonology and Phonetics of English Intonation*. PhD dissertation. MIT.

Quirk, R, S. Greenbaum, G. Leech, and J. Svartvik (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Siegman, A.W., and S. Feldstein (Eds) (1979). *Of Speech and Time: Temporal Speech Patterns in Interpersonal Contexts.* Hillsdale, N.J.: Laurence Erlbaum.

Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. (1992). ToBI: A Standard for Labelling English Prosody. In *Proc. 1992 Intl. Conf. Speech & Language Processing, Banff, Canada*.

Stenström, A. (1990). Pauses in monologue and dialogue. In J. Svartvik (Ed), pp 211-252.

Souter, C., and E. Atwell (Eds) (1993). *Corpus-based Computational Linguistics*. Amsterdam: Rodopi.

Svartvik, J. (Ed) (1990). *The London-Lund Corpus of Spoken English - Description and Research*. Lund: Lund University Press.

Svartvik, J., and R. Quirk (1979). *A Corpus of English Conversation*. CWK Gleerup Lund.

Tomlin, R. (Ed) (1987). *Conherence and Grounding in Discourse*. Amsterdam/ Philadelphia: John Benjamins Publishing Company.

van Santen, J.P.H. (1993). Timing in text-to-speech systems. In *Proc. EuroSPEECH-93, Berlin*, pp 1397-1406.

Wightman, C.W., and M. Ostendorf (1995), Automatic labeling of prosodic patterns. In *IEEE Trans. Speech and Audio Processing, 2*, pp 469-481.

Text Code	Text Nos.	Title	Source	Date
S2B-021	1-4	Journalists' monologues on presidential wealth	LBC Radio	07-07-91
S2B-022	1	The River Times	ITV	21-06-91
S2B-022	2	Nature	BBC 2 TV	05-03-91
S2B-023	1	Can You Steal It?	BBC Radio 1	02-03-91
S2B-023	2	Spirit Level	Radio Oxford	20-01-91
S2B-023	3	From Our Own Correspondent	BBC Radio 4	27-04-91
S2B-024	1	Viewpoint '91: Poles Apart	ITV	30-04-91
S2B-024	2	40 Minutes	BBC 2 TV	08-11-91
S2B-025		For He Is an Englishman	BBC Radio 4	05-11-90
S2B-026		The World of William	BBC Radio 4	05-02-91
S2B-027		Castles Abroad	ITV	21-06-91
S2B-028	1	Lent Observed	BBC Radio 4	21-11-90
S2B-028	2	Lent Observed	BBC Radio 4	26-02-91
S2B-029		The Reith Lecture, No. 2	BBC Radio 4	02-11-90
S2B-030	1	Address to the Nation	BBC Radio 4	17-01-91
S2B-030	2	Address to the Nation,	BBC Radio 4	18-01-91
S2B-030	3	Labour Party Political Broadcast	BBC 1 TV	13-03-91
S2B-030	4	Social Democratic Party Political Broadcast	Channel 4	13-03-91
S2B-031	1-2	The Police Debate	BBC Radio 4	10-02-91
S2B-031	3	The Week's Good Cause	BBC Radio 4	30-03-91
S2B-032	1	Opinion: King or Country	BBC Radio 4	07-11-90
S2B-032	2	The Police Debate	BBC Radio 4	10-02-91
S2B-033		Barry Norman's film '91	BBC 1 TV	12-03-91
S2B-034		Analysis	BBC Radio 4	16-05-91
S2B-035	1-2	The Class Debate	BBC Radio 4	24-02-91
S2B-036	1-2	The Class Debate	BBC Radio 4	24-02-91
S2B-037		The Scarman Report	BBC Radio 4	16-06-91
S2B-038	1	Medicine Now	BBC Radio 4	12-03-91
S2B-038	2	Medicine Now	BBC Radio 4	19-03-91
S2B-038	3	The Week's Good Cause	BBC Radio 4	17-03-91
S2B-039	1-3	From Our Own Correspondent	BBC Radio 4	02-04-91
S2B-040	1-3	From Our Own Correspondent	BBC Radio 4	27-04-91

Table 1: A list of texts in broadcast talks

Temporal File	Syntactic File
54.10000 It 54.90000 looked 55.16000 interesting 55.79000 GAP: 56.16000 even 56.35000 if 56.47000 underpaid 57.17000 :GAP:	PU S(act,decl,indic,cop,past,unm) SU NP() NPHD PRON(pers,sing) {It} V VP(act,indic,cop,past) MVB V(cop,past) {looked} CS AJP(ingp) AJHD ADJ(ingp) {interesting} A AVP(add) AVHD ADV(add) {even} A CL(-su,indic,montr,pass,pastp,sub,unm) SUB SUBP() SBHD CONJUNC(subord) {if} V VP(indic,montr,pass,edp) MVB V(montr,edp) {underpaid}

Figure 3: A sample of time-word and syntactic tree files

```
<#5> 14 6
PU S(act,decl,indic,cop,past,unm)
 SU NP()
  NPHD PRON(pers,sing) 1
   {It 54.10000}
 V VP(act,indic,cop,past)
  MVB V(cop,past) 1
   {looked 54.90000}
 CS AJP(ingp)
  AJHD ADJ(ingp) 1
   {interesting 55.16000}
 A AVP(add)
  AVHD ADV(add) 1
   {:GAP: 55.79000}
   {even 56.16000}
 A CL(-su,indic,montr,pass,edp,sub,unm)
  SUB SUBP()
   SBHD CONJUNC(subord) 1
    {if 56.35000}
  V VP(indic,montr,pass,edp)
   MVB V(montr,edp) 1
    {underpaid 56.47000}
    {:GAP: 57.17000}
```

Figure 4: A sample output of temporal and syntactic alignment

Category	*Frequency*		*Percentage*	
Sentence	150	150	100.00	48.5
Clause	106	38	35.8	12.3
Verb phrase	273	38	13.9	12.3
Prepositional phrase	197	20	10.2	6.5
Adverb phrase	109	9	8.3	2.9
Adjective phrase	161	8	5.0	2.6
Noun phrase	581	28	4.8	9.1
Others		18		5.8
Total	1577	309		100.00

Table 2: Frequency distribution of pauses among major syntactic categories

[1] The database should also be accompanied by a clear statement about Intellectual Property Rights. It is still very common to see private databases unavailable to the scientific world at large.

[2] For detailed descriptions of ICE wordclass and syntactic annotation schemes, see relevant chapters in this volume.

[3] Composite texts are those comprising separate extracts from a single or different sources.

[4] A word here was defined as a string having a word-class label in the syntactic analysis.

[5] For a description of the ICE syntactic parsing scheme, see Chapter ** in this volume.

[6] An ICE parsing unit roughly correponds to the conventional notion of *sentence*, though there are cases where a parsing unit is only a syntactic phrase.

[7] We distinguish the sentence from the clause. A sentence consists of either one sentence or of coordinated sentences. In this treatment, clauses - whether finite or nonfinite - are always subordinate.

[8] In comparison, Lumley (1933) noted an average rate of 1.8 words per second for politicians, 2.7 words per second for educators, 2.9 words per second for preachers, and 3.2 words per second for reporters (cf. Altenberg 1987:16). Altenberg (ibid., 22) found an average of 2.4 words per second for the LLC Corpus.

[9] In Stenström 1990, nearly half (48%) of the clauses in spontaneous speech have initiating pauses, revealing the effect of the cognitive process of planning and selecting.

[10] The word *be* is grouped into various uses by the ICE tagging scheme. The four examples include *be* as the passive auxiliary ([1] and [4]) and the copula ([2] and [3]).