Proceedings of the Institute of Acoustics

STATISTICAL RELATIONSHIPS BETWEEN AUDITORY AND ACOUSTIC RECORDS OF INTONATION: DESIGN OF A DATABASE

Nawal Ghali, Simon Arnfield and Peter Roach

Speech Laboratory, Department of Psychology, University of Leeds, U.K.

1. INTRODUCTION

In our attempts to make generalisations about intonation we depend heavily on the validity of the method used to represent prosodic information. In order to evaluate the relationship between intonation transcription and the physical properties in the speech signal we need a large sample of transcribed recordings; this paper describes work on such a corpus which also provides a considerable amount of grammatical information.

Many methods have been developed for recording pitch movements as heard by the trained analyst. Some attempt to record pitch movements with maximal phonetic accuracy (and hence with some redundancy) while others rely on a prior phonological analysis to make possible a more economical coding with minimal redundancy. Ideally, any good intonation transcription should make it possible to generate an acceptable fundamental frequency contour that closely resembles that of the original speech, though explicit statements of this as a goal are comparatively recent. However, we do not know in quantitative terms how successful we can expect this process to be. Experimental evidence on the fallibility of human judgements about prosody tend to be based on rather small samples. What is needed is a large body of human-transcribed speech in computer-readable form that will enable us to explore in statistical terms the relationship between the trained expert's auditory transcription and the acoustic analysis of the same data carried out by computer.

We are working on a project (funded by ESRC and shared with Lancaster University) to convert the Spoken English Corpus (Knowles et al, forthcoming[1]) into a machine-readable database stored on CD-ROM. This corpus, the original work on which was funded by IBM UK, comprises around 6 hours of radio broadcasts and other talks, and has already been prosodically transcribed in its entirety by two experts; the text is in machine-readable form with numerical codes for tone-marks. The prosodic transcription that was chosen for the analysis is a variety of the type of transcription commonly referred to as "Standard British", but it differs in some significant respects from the most widely adopted versions such as O'Connor and Arnold [2], particularly in that it does not segment the intonation-unit into pre-head, head, nucleus and tail. Each pitch-accent is therefore marked with one of the available tones, with the convention that the last pitch-accent in the intonation unit is deemed to be the nucleus.

The corpus has been grammatically tagged word by word, and an automatically generated parse applied. The textual form of the corpus is therefore very richly annotated. Our project is currently working on the digitisation of the corpus and the acoustic analysis of it, and this will be

followed later this year by work on automatically aligning the text, syllable by syllable, with the acoustic signal, so that the recording corresponding to any portion of the corpus can be retrieved very easily from the disk.

The research work described in this paper is intended to produce three principal deliverables:

(i) A machine-readable version, stored on CD-ROM, of the six-hour corpus of digitised recordings on which the Spoken English Corpus is based. This machine-readable corpus is called MARSEC (MAchine-Readable Spoken English Corpus).

(ii) A set of linked files providing textual, grammatical, acoustic and prosodic annotations of the recordings, all with a common time reference.

(iii) A statistical methodology for examining the relationship between the expert auditory transcription of the existing SEC and the acoustic parameters that can be extracted from the recorded signal.

In addition, the project is evaluating alternative transcription systems for work of this sort. The existing SEC was transcribed using an approach that would not necessarily be regarded as ideal in the context of present-day prosodic research. This part of the research is primarily the field being worked on by our partners in the University of Lancaster. The following outline is based on the three headings given above, and follows them in the order given.

2. DELIVERABLE (i): *Machine-Readable Speech Data*

The corpus (i.e. the full set of recording) lasts for around 6 hours. We decided that it should be converted into digital form in pieces lasting no longer than 1 minute (to enable microcomputer-based speech workstations to handle chunks of the corpus without the need for further editing). One minute of speech takes up a little less than 2 mbyte.

2.1. Filenames and subdirectory structure:

There are 11 subdirectories, corresponding to the 11 categories of the corpus (A through M, excluding I and L). The files have been named according to the section number, the 1-minute chunks and, at the end, we have added either b (if the prosodic transcriber was BJW) or g (if the transcription was done by GOK). The extension .sig represents signal files. For example, the file A0101b.sig is part A, section 1, first minute, transcribed by BJW.

2.2 Digitising the recordings:

We used a PC configured as a SAM workstation, since we intend to use the SAM conventions and protocols as far as possible: these have become the de facto standard for most collaborative European speech research. The files were created with AU21DSK, a program produced by the

manufacturers of the OROS AU21 board that is used in the SAM workstation. We used a sampling rate of 16 KHz. The OROS board automatically applies appropriate anti-aliasing filtering.

2.3 Editing and storing:

We have used the PTS signal editing package to edit these files. Each file has been limited to within one minute, breaking at a pause (a clear silence). The recording includes also the first word following the pause, and the following file starts from the word before the pause; in this way we preserve the pause itself. The breaks have been marked on a master copy of the text.
The recording were initially archived on Exabyte magnetic tape, and was then transferred on to a single CD-ROM disk, which will be available through the ESRC Data Archive[1].

3. DELIVERABLE (ii): *Cross-referencing mechanism for the corpus.*

The MARSEC corpus will consist of several versions of the data as follows:

* acoustic waveform
* fundamental frequency waveform
* intensity waveform
* phonetic transcription
* syllabic division transcription
* prosodically annotated transcription
* punctuated transcription
* word tag transcription
* parse treebank

This section describes a mechanism for cross-referencing from any file to the equivalent position in any other file. To be able to do this several complicated indexes need to be created. The phonetic transcription is aligned automatically with the acoustic waveform by HMM. This is the key step in allowing cross-referencing between the acoustic data and the textual data. The next step is to match the phonetic transcription with the syllabic transcription. This should be fairly straightforward since the phonetic transcription will have been generated from the prosodic text (as will the syllabic transcription) although they might contain some minor differences. The syllabic transcription will have been generated from the prosodic transcription and as such will be easy to match backwards, especially if some re-alignment data (such as word boundaries) is included in the syllabic transcription (maybe only temporarily). The task of aligning the prosodic transcription with the word-tag

---

[1] ESRC Data Archive, University of Essex, Colchester CO4 3SQ.

transcription has mostly already been achieved. Matching the treebank and punctuated versions with the word-tag is relatively straightforward, and will be easier when the errors that have been introduced by post-editing of the punctuated version and the word-tag version have been resolved.

3.1 CROSS-REFERENCING

There are thirteen basic types of cross-reference possible, shown diagrammatically in Fig.1:

3.1.1.1 acoustic -> phonetic transcription
3.1.1.2 acoustic -> fundamental frequency / intensity.
3.1.1.3 acoustic -> syllabic, prosodic, word-tag, punctuated, treebank

3.1.2.1 F0/intensity -> acoustic
3.1.2.2 F0/intensity -> phonetic
3.1.2.3 F0/intensity -> syllabic, prosodic, word-tag, punctuated, treebank

3.1.3.1 phonetic -> acoustic
3.1.3.2 phonetic -> fundamental frequency / intensity
3.1.3.3 phonetic -> syllabic, prosodic, word-tag, punctuated, treebank

3.1.4.1 text -> acoustic
3.1.4.2 text -> fundamental frequency / intensity
3.1.4.3 text -> phonetic
3.1.4.4 text -> syllabic, prosodic, word-tag, punctuated, treebank

Full details of the interlinking of these levels are given in MARSEC documentation which will be distributed to users of the corpus.

3.2 FILE FORMATS AND NAMING CONVENTIONS IN THE SEC

The SEC consists of various versions of the data. This section explains what files exist and their file formats. The original version of the corpus was produced from tape recordings of radio broadcasts and some other talks. From these the unpunctuated transcription was produced which was punctuated by volunteers and prosodically transcribed by Gerry Knowles and Briony Williams. The punctuated version was then used to produce the word tag version, and later the parse treebank (not described in this paper). The new version of the corpus will also include five more versions: acoustic waveform, fundamental frequency waveform, intensity waveform, syllabic division transcription, phonetic transcription.

AUDITORY AND ACOUSTIC RECORDS OF INTONATION

3.2.1 UNPUNCTUATED VERSION

Not distributed. This version is exactly the same as the punctuated version (below) except that no punctuation symbols are included. There are 53 files in 11 categories:

A01 A02 A03 A04 A05 A06 A07 A08 A09 A10 A11 A12
B01 B02 B03 B04
C01
D01 D02 D03
E01 E02
F01 F02 F03 F04
G01 G02 G03 G04 G05
H01 H02 H03 H04 H05
J01 J02 J03 J04 J05 J06
K01 K02
M01 M02 M03 M04 M05 M06 M07 M08 M09

Each category deals with a different type of speech style. See 'A Manual of Information to Accompany the SEC Corpus' for more information.

3.2.2 PUNCTUATED VERSION

In many instances these files were produced by volunteers punctuating the given unpunctuated text and represent the original "script" that the reader is supposed to have used in producing the recording. However, in some cases the original scripts were in fact available and have been used. This has allowed some variability to creep into the corpus where the original script was not followed exactly by the speaker. Due to the nature of the task of punctuating a text derived from a recording and not being able to hear the recording there will inevitably be some spurious punctuation. File name conventions are as for the unpunctuated version and the files exist in the 'pun' subdirectory. Some editing has been done and is noted by comments such as [change of speaker], [live commentary omitted], or [interview omitted]. Each file contains some header information contained within square brackets, stating the text number, title, speaker(s) and broadcast notes. For example:

[001 SPOKEN ENGLISH CORPUS TEXT A01]
[In Perspective]
[Rosemary Hartill]
[Broadcast notes: Radio 4, 07.45 a.m., 24th November, 1984]

Good morning. More news about the Reverend Sun Myung Moon,
...

3.2.3 PROSODIC VERSION

The prosodic version was split into three directories: bjw, gok and dup. bjw and gok contain the files prosodically transcribed by Briony Williams and Gerry Knowles respectively whereas dup contains sections from the corpus that have been transcribed by both bjw and gok. The same file name conventions have been used but with some slight changes. In some cases one of either bjw or gok will have transcribed the whole section and the other will have done a small section for comparison purposes. In these cases the whole transcribed section will occur in one of bjw or gok and the repeated section will occur in dup. In other cases each will have done half a section with a small overlap. In these cases the same file name will occur in each subdirectory bjw, gok with the overlap occurring in dup. It is therefore true to say that some of the information in dup is entirely repeated (as in the last case above) whereas some information (the first case above) only exists here. Header information is included as for punctuated files but with the addition of a line to indicate the transcriber. Unfortunately no indication is given as to where (in this recording) this file comes from. So, for example, in section C where the only recording is C01 which was split between the transcribers it is impossible to discover whether bjw/C01 precedes gok/C01 or vice-versa without examining the text for clues. The main differences in file format from the punctuated version is the omission of punctuation (except apostrophes in words such as "don't" and hyphens) and the inclusion of prosodic information.

Prosodic information is marked with a set of codes. There are three marks used for tone unit boundaries: | || and #240; the latter is also used to mean low rise-fall, but fortunately this only occurs once in the corpus on line 148 of bjw/F04: "#240following". Other prosodic symbols are _ (meaning low-level) and the remainder which comprise a # followed by a 3 digit number. The full list of prosodic symbols, as they appear in the corpus, is: | || _ #161 (high fall-rise), #162 (high rise-fall), #246, #247 (low fall-rise), #171 (low rise), #172 (high rise), #173 (low fall), #174 (high fall), #163 (high level), #248 (stressed unaccented syllable), #165 (pitch raising), #166 (pitch lowering), #240 (low rise-fall); #249 was originally used as synonymous with #248. The code #240 is used as a "hesitation tone-unit boundary" and was only transcribed by GOK. In addition to this (* bracketed words erased *) and (* unfinished tone group *) also occur.

Example of prosodic transcription file:
[001 SPOKEN ENGLISH CORPUS TEXT A01]
[In Perspective]
[Rosemary Hartill]
[Broadcast notes: Radio 4, 07.45a.m., 24th November, 1984]
[Transcriber: BJW]
#166Good #174morning || #165#174more #249news about the #163Reverend _Sun #248Myung #174Moon | _founder of the Unifi#174cation #248Church | who's #161currently in #248jail | for #174tax evasion || ...

3.2.4 WORD-TAG VERSION

This actually exists in two formats, "vertical" and "horizontal". In practice the horizontal version is only of use because it is easier to read and this will not be described here. The vertical tag format is that produced by the CLAWS (Garside et al [3]) tagging suite developed for the LOB corpus and exists in the vtag subdirectory. The horizontal tag files exist in the htag subdirectory. The same file name conventions used by the punctuated version are used. The format of the vertical tag file contains six columns, each line having a single entry such as a single word or punctuation symbol. Column 1 is the file name; column 2 is the line number in the punctuated version from which this entry came; column 3 is a three digit number, the first two of which indicate the word number on the line and the third digit is used to number each punctuation symbol following the word; column 4 contains the word-tag; column 5 is the word/punctuation entry; column 6 contains residual information such as marking manual editing (@), some enclitics (< and >), some compound/non-standard words (*), ditto forms etc. End of sentences are marked by inclusion of end of sentence markers. These are tagged as 5 hyphens and the word is 43 hyphens. A number of differences (from the punctuated version) have been introduced into the vertical tag format due to editing -- these largely take the form of inserted/deleted/modified punctuation. One notable difference is that A01--A06 have had square brackets changed to parentheses, whereas elsewhere they remain square. Header information is removed except for the title and author, and lines such as [change of speaker] and [speech extract omitted]. This information is tagged just as other information. Here is an example:

```
A01  2 001 (    (                              @
A01  2 010 IN   In
A01  2 020 NP   Perspective
A01  2 021 )    )                              @
A01  3 001 (    (                              @
A01  3 010 NP   Rosemary
A01  3 020 NP   Hartill
A01  3 021 )    )                              @
A01  5 001 ----- ------------------------------------------
A01  5 010 JJ   good
A01  5 020 NN   morning
A01  5 021 .
A01  5 022 ----- ------------------------------------------
A01  5 030 AP   more
A01  5 040 NN   news
A01  5 050 IN   about
A01  5 060 ATI  the
A01  5 070 NPT  Reverend
```

AUDITORY AND ACOUSTIC RECORDS OF INTONATION

4. DELIVERABLE (iii): *Statistical Treatment*

This is the area in which least progress has been made so far, since it is intended that it will form the bulk of the work in the second year of the project. The problems to be addressed are familiar enough to anyone with experience of working with acoustic records of prosodic phenomena, and for the present we shall simply summarise the main problems as we see them. Each F0 and intensity file generated by acoustic analysis of the data files is in the form of a continuous vector corresponding roughly to pitch and loudness. The relationship between the two acoustic parameters on the one hand and intonation on the other is not fully understood, and will be a major focus of interest in the research. We feel that acoustic studies of intonation have concentrated too exclusively on fundamental frequency as the correlate of the auditory percept, and we hope to establish that some weighted function of fundamental frequency and intensity will produce a better match.

In carrying out auditory transcription, the auditory system appears to perceive pitch as continuously varying. In the SEC, major changes in the pitch are marked on the text with one of a fixed number of tone marks (e.g. / for rising movement, \ for falling). In the simplest kind of comparison, we could calculate *for a given speaker* what acoustic characteristics correspond to the tone marks \ and / on single syllables. We would then be able to predict what tone mark a human analyst would use to represent a given set of F0 values for such items. A number of factors cause complications:

1. Different speakers have different pitch ranges. In studying more than one speaker, therefore, we need to look not at *absolute* F0 values but at *relative* ones (relative to the speaker's normal pitch range.

2. Although on single syllables the relationship between (auditory) pitch and (acoustic) F0 is easy to see, the tone mark actually predicts pitch behaviour over all the syllables which follow the tone-marked syllable up to the next tone mark or the tone-unit boundary which follows. In terms of the auditory analysis of intonation used for the SEC, therefore, there is no categorical difference between the following:

/ no / nobody / nobody went there / nobody went there without a ticket

- though the F0 record will look very different. The same tone mark therefore represents a considerable variety of F0 contours.

3. The tone marks seem to imply a continuous pitch movement, but the F0 track stops and reverts to baseline when voicing ceases; this happens at silent pauses, of course, but also at voiceless consonants. So a sentence like 'Eat sweet potatoes' will have several large gaps in the F0 data that will cause problems for pitch-reading algorithms, though the human auditory system is able to "connect up" between the voiced parts.

4. It will not be sufficient to be able to state context-free rules to convert between acoustic and auditory values. The phonological categories used in intonation transcription are realised in very different ways in different contexts, particularly as regards position in the tone-unit. Consequently an analysis based on isolated fragments of speech could only be regarded as a tentative preliminary approach. One of the most valuable aspects of the material we are working with is the wealth of examples of relatively unrestricted connected speech. We hope that this will permit us to make more progress on this research area than has been possible in earlier work.

5. REFERENCES

[1] G.O.KNOWLES, L.TAYLOR AND B.WILLIAMS *The Lancaster/IBM Spoken English Corpus*, Longman (forthcoming).

[2] J.D.O'CONNOR AND G.F.ARNOLD *The Intonation of Colloquial English*, (2nd.Ed), Longman (1973).

[3] R.GARSIDE, G.LEECH AND G.SAMPSON *The Computational Analysis of English*, Longman (1987).