10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

The first spoken intonation corpus (1909): a re-assessment of Daniel Jones's *Intonation Curves*

Michael Ashby

UCL (University College London), UK
m.ashby@ucl.ac.uk

Abstract

In 1909 the UCL phonetician Daniel Jones (DJ) published a small volume entitled *Intonation Curves*, which was based on the exhaustive analysis of eight commercially available gramophone records of the day, covering English, French and German. The method was to listen repeatedly to the records, lifting the needle at numerous successive points and plotting the final pitch heard so as to trace continuous lines on a musical stave, along with the orthographic text and a detailed phonetic transcription. The result, an approximately 20-minute body of recordings with time-aligned representations on several levels, deserves recognition as an early spoken corpus.

We have located and digitized copies of most of the discs used in the original study, re-creating the corpus and making possible a replication, though the noisy 110-year-old recordings present challenges in acoustic analysis. Evaluation of various pitch-trackers shows that DJ's auditory method matches or outperforms the most successful present-day algorithms. Plotting the modern f_0 determinations on a stave for comparison with DJ's reveals that his detailed intonation curves are remarkably accurate.

We review the significance of the work, and its relationship to the later development of formal models of intonation for the three languages.

Index Terms: Daniel Jones, spoken corpus, pitch tracker, intonation, gramophone record, history of phonetics

1. Introduction

Intonation curves [1] is an early work by the UCL phonetician Daniel Jones (1881–1967) [2]. It was published in 1909 when he was still in his twenties. The work is based on the exhaustive analysis of eight commercial gramophone records of the day, chosen by DJ to represent 3 languages (English, French and German) and to sample 2 styles, literary and conversational. Aligned with the intonation curves themselves, Jones gives a 'detailed' segmental transcription of the speech on the records, and on the facing page an orthographic version and another 'standard' transcription, which is his own approximately phonemic transcription of the text.

DJ thus identifies a set of 8 recordings (about 20 minutes of speech in total) and provides time-aligned representations of them on several levels. So whereas in one way the work is a sort of 'phonetic reader' (i.e., a teaching resource for language learners), from another point of view it can be seen as a very early example of an annotated speech corpus. It precedes by many decades the earliest spoken corpora generally noted in surveys of corpus linguistics [3].

.

2. The intonation curves

2.1. Pitch axis

DJ represents intonation with continuous lines drawn on a musical stave (see Figure 1). The use of musical notation was not uncommon at the time, not only for intonation and tone, but in acoustic work more widely. After all, the musical notation of pitch is essentially a logarithmic scale of frequency. It is important to realize that Jones intends his representations to be taken literally, at concert pitch. In Figure 1, for example, the onset pitch in the first line, shown as F in the bass clef, means specifically what the notation implies: in this case, 174.6 Hz. Musical notation may appear quaint from a modern perspective, but the intonation representations are in principle entirely quantitative.

2.2. Time axis

The time axis makes no claim to quantitative representation. DJ divides the intonation curves into 'bars', each bar corresponding to one phonetic syllable. The width of each such

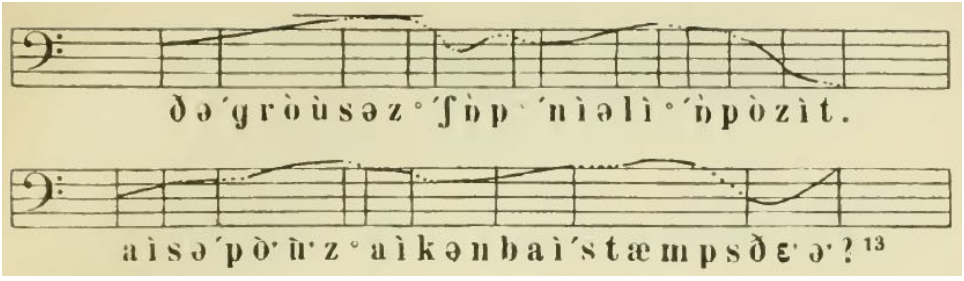

Figure 1: Two sample lines from DJ's analysis of the 'English Conversation' record ([1, p. 19]. The text is *The grocer's shop nearly opposite. / I suppose I can buy stamps there?*

10.21437/SpeechProsody.2020-25

Table 1: DJ's corpus of 1909 as used in *Intonation Curves* ('CG cat no' = Gramophone Company catalogue number). Surviving copies of all the English and French recordings have been located and digitised for this study. The Geman recordings have not yet been found.

Language	GC cat no	Title	Speaker
English	1315 III	Passage from Shakespeare's *Richard II*	Sir H. Beerbohm Tree
English	1356	Poe's *The Bells* (verses 1–3)	Canon Fleming
English	1286	Conversation from Langenscheidt's *Englisch*	Bernard MacDonald
French	31171 II	Passage from Rostand's *La Samaritaine*	Sarah Bernhardt
French	31253	Lafontaine's *Le Corbeau et le Renard*, and *Le Loup et l'Agneau*	Louis Delaunay
French	31284	Conversation from Barlet and Rippmann's *French Life and Ways*	unknown
German	11968 II	Passage from Schiller's *Wallenstein*	Max Montor
German	41319 III	Passage from Goethe's *Faust*	Otto Sommerstorff

bar on the page is simply the space taken up by the symbols required to transcribe it. So, though some are narrow and some wide, the width has no relationship with real duration.

3. The recordings

The eight gramophone records on which the analysis was based were commercially available at the time of the study (there was already a considerable market in spoken word recordings). DJ's intention was plainly that the user would acquire copies of the records and use them in in combination with his work. DJ gives full particulars of the records, which has made it possible to seek surviving copies of them and attempt a partial verification of his findings. They are listed in Table 1. For the present research, copies of six records (the three English discs, and the three French) have been located and digitized. The two German recordings have so far not been found. It was evidently part of DJ's original design to use 9 recordings, but he was unable to find a German recording in conversational style to match those found for French and English. The English and French 'conversational' recordings are, of course, of *scripted* conversational material—already at this date a long-established language-teaching genre.

4. Methods of f_0 estimation

4.1. DJ's mode of working

The method of working was to listen repeatedly to the records, lifting the needle from the record at successive points and noting, as Jones says, the 'the impression of the sound heard at the instant when the needle is lifted' [1, p. v]. The same method was also used in making the detailed segmental transcriptions: DJ says '…it is generally necessary to take several observations for the quality of a given sound and several more for its pitch'.

The interrupted listening technique which DJ applied, which may perhaps be seen as an early form of 'gating' [4], probably has the effect of favouring a certain mode of listening. He was selectively attending to very short portions of the signal, and furthermore listening analytically to just one characteristic—pitch. As a result, there was a greatly reduced load on short term memory, and little or no need to impose a linguistic categorization on what he was hearing. Any listener under those conditions is likely to be pushed away from a 'speech mode' of perception towards a psychoacoustic mode of response.

4.2. Contemporary instrumental methods

DJ was well aware that objective records of pitch could be obtained 'by means of tracings of voice vibrations, obtained by the use of a kymograph or otherwise' [1, p. iv]. In fact, he gives a detailed demonstration of how this may be done in the first edition of one of his own books [5, pp. 179–182]. But he comments that 'the work of preparing curves by this method is so laborious, that no one has ever yet analysed texts of sufficient length to be of any practical value to language students' [1, p. iv]. His own approach aims to 'combine as far as possible scientific accuracy with practical utility'.

DJ gives no indication of how much labour his own approach entailed, but a rough estimate based on the extent of the material suggests that around 75,000 individual judgements would have been required.

Reasonably accurate automatic extraction of voice pitch with analog hardware, resulting in an immediately displayed intonation curve, did not become possible until the late 1930s [6].

4.3. Pitch extraction algorithms

4.3.1. Dominance of Praat

Replication of DJ's study requires the application of a pitch-extraction algorithm to the digitized recordings. For many speech researchers, the question of selecting an acoustic analysis tool would not arise, Praat [7] being the unquestioned choice.

For example, a survey of the 774 papers accepted for ICPhS2015 [8] shows that 420 make some use of Praat—somewhat more than 54%. Similarly, Strömbergsson [9] conducted a bibliometric survey specifically of pitch tracker utilization in research papers over the period 2010–2016, and found Praat specified in 80 out of 141 publications (about 57%). It is probably no exaggeration to say that over recent years at least half of all published research in phonetics (including prosody) has depended in some degree on Praat.

It is true that in the same study Strömbergsson found Praat the best performer of a number of tools assessed empirically against a 'ground truth reference' of laryngographic f_0 determinations. But speech researchers have probably been too ready to be reassured by that and similar comparative

evaluations, while forgetting that that test signals employed were clean speech recorded in studio conditions. The real-world performance of algorithms when presented with flawed signals may be different. The early recordings which constitute DJ's corpus are challenging on account of high levels of noise, and the curtailed and uneven frequency characteristic of recordings made acoustically (the electrical recording process was not introduced until 1925).

4.3.2. The RAPT algorithm

The present study was conducted using SFS [10], which offers a menu of different pitch extraction algorithms. All were evaluated on sample recordings, and it was quickly established that the most successful results were obtained with 'fxrapt', an implementation of the RAPT algorithm [11]. Preliminary trials were also made using Praat, but with discouraging results. It is likely that this study would have been abandoned at an early stage had Praat been the only system evaluated for f_0 estimation.

5. Methods of comparison

Raw curves of f_0 extracted from the recordings are not directly commensurate with DJ's musical representations. Two methods of effecting the comparison suggest themselves. Either (i) sample points on DJ's curves can be converted to values in Hz, and these correlated with values taken from corresponding points on the f_0 curves (we call this 'DJ > f_0'), or (ii) the f_0 curves may be mapped on to a musical stave matching DJ's, permitting direct visual comparison (call this 'f_0 > DJ').

5.1. Method DJ > f_0

We first examine the most obvious method of mapping, DJ > f_0. To convert DJ's musical representations into f_0 estimates, positions of successive points on the curves drawn on the stave must be estimated graphically. A section of French conversation 25 syllables in length was taken from the beginning of the French Conversation record. A scanned image of the printed page was viewed in high magnification and measured with an on-screen cursor. The *y*-coordinate of each measured point was transformed into the logarithm of the corresponding frequency by means of a lookup table constructed to provide one row for each pixel on the vertical axis of the scan. Annotations were placed at corresponding measurement points in the audio file, resulting in 50 pairs of f_0 determinations at sampling points in 25 syllables, displayed as a scatterplot in Figure 2. The correlation is high (r^2 = 0.77), indicating an excellent degree of agreement.

It will be seen that the intercept of the fitted regression line is negative, tending to suggest that DJ's judgements are on the high side. But this merely indicates that he probably chose to play this record somewhat faster than the modern standard of 78 rpm (he had a variable-speed clockwork gramophone, and some record-to-record variation in reproduction rates had therefore been anticipated). Further, much of the scatter in Figure 2 is probably owing to (unavoidable) misalignment in time of the selected sample points rather than to errors in DJ's judgement. The impossibility of an exact mapping from DJ's time scale to real time is an irremediable shortcoming of the DJ > f_0 mode of comparison.

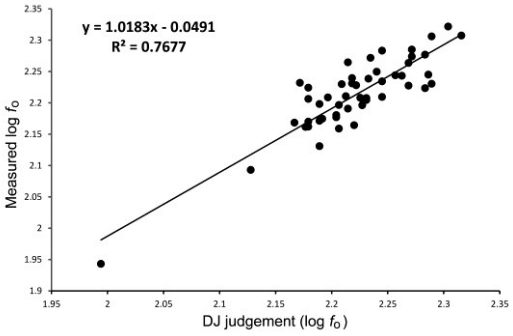

Figure 2: Correlation of automatically-made estimates of f_0 with DJ's estimates in a sample of 25 syllables.

5.2. Method f_0 > DJ

It is straightforward to produce a system which will accept output from a pitch tracker and plot it on a suitable logarithmic scale along with constants (to draw the stave lines) and other graphic elements. To copy the pseudo-time of DJ's bar lengths, we need to annotate the audio files so as to match the phonetic segments of DJ's 'detailed' transcription. The width (in pixels) occupied by each phonetic symbol can be determined, and a script can then stretch or compress the time scale in each region so as to give the required width when the f_0 curve is plotted. Finally, the original and the automatically-derived version can be juxtaposed for comparison. An example from the English Conversation record appears as Figure 3.

6. Results

Figure 3 indicates an astonishing level of agreement, better indeed than might be expected from two present-day pitch tracking algorithms evaluated in parallel. Notice the pitch trough in the first syllable of *letters*, the small peak located in the second syllable of *cashing*, and how the greater part of the final rise is correctly placed in the second (weak) syllable of *orders*. Overall, DJ's judged pitch is about 4% flat compared with the modern analysis—suggesting that he was playing this

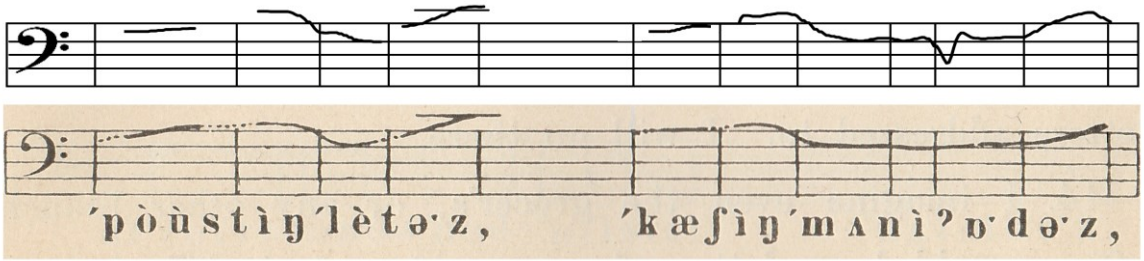

Figure 3: Top: automatically plotted intonation curve for two phrases from the English Conversation record (male speaker); below: DJ's version. The text is *posting letters, cashing money orders.*

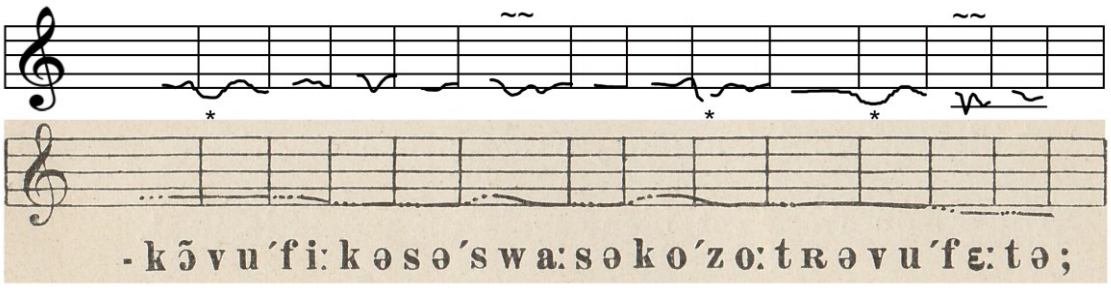

Figure 4: A line from Sarah Bernhardt's recitation of *La Samaritaine* Top: automatic f_0 extraction; below: DJ's estimates. The text is *Qu'on vous fit, que ce soit ce qu'aux autres vous faites*, 'Do unto others [as you would have] them do unto you'.

particular record at approximately 75 rpm. The degree of agreement exemplified in Figure 3 is entirely typical of that found when samples are selected throughout the work.

To illustrate the full extent of agreement across widely differing speakers and styles, Figure 4 shows a specimen from the one female speaker in the corpus, in a stylized verse performance in French. Note the switch to the treble clef. While it is true that the almost chanted rendering of this piece means that it has little interest from the point of view of French intonation, the long stretches of approximately level pitch are interesting for at least three reasons. First it is relatively straightforward to compare DJ's pitch standard with the modern one, without uncertainty over the time alignment of the points compared. Second, the evident intention to maintain the same level pitch across vowels which are separated by consonants gives an opportunity to study pitch perturbations ('micro-intonation') caused by the consonant articulations. Third, the speaker employs a very prominent vibrato, evidently for emotional effect, which takes the form of a rapid fluctuation in pitch at around 7–8 Hz, selectively applied to prominent syllables. Both the involuntary pitch perturbation and deliberate vibrato are likely to be detectable in the physical record of f_0, but it is interesting to enquire whether they are noted in DJ's auditory analysis.

The measured f_0 curve shows downward excursions at 3 locations (marked * in Figure 4) probably as a result of aerodynamic resistance during voiced fricatives. DJ does not note these perturbations.

The fluctuations in f_0 evident in [swaː] and [fɛː] (marked with ~~ in Figure 4) are to be attributed to deliberate vibrato. DJ's representation effectively smoothes these to minor pitch prominences (peaks or falls) associated with the stressed syllables. In both cases, therefore, DJ appears to be abstracting the linguistic component of the pitch variation from the stylistic and aerodynamic confounds superimposed on it.

7. Discussion

It is unlikely that the originality and extraordinary accuracy of DJ's work have ever been properly appreciated. A few contemporaries with access to the recordings may have come close; a 1910 reviewer hailed *Intonation Curves* as 'the most elaborate, accurate and careful transcription of intonation existing anywhere' [12, p. 82]. But only an observer with the skill and patience to repeat some of DJ's observations could know this.

Later commentators are typified by 't Hart et al., who put the work in a category of 'dubious empirical status' and say 'impressionistic auditory descriptions remain difficult to interpret and may not be representative of other listeners' perceptions' [13, p. 4]. But as has been demonstrated, DJ's judgements—though not presented in a superficially numerical form—are quantitative, precise, objective and verifiable. It is hard to see in what respect they fail to be 'empirical'.

Of course, DJ presents raw data and makes no attempt at formal systematization. 'Professor Jones gives us graphic representations of intonation, but no theory' [14, p. 3]. The early history of formal systems for intonation analysis has yet to be studied in any detail (but see [15]). Hermann Klinghardt (1847–1926) published pioneering works on the intonation systems of all three languages, English, French and German, while the most significant early British theorist was H. E. Palmer (1877–1949). Though both Klinghardt and Palmer were senior to DJ, all their relevant publications are later than 1909, and both acknowledge a debt to DJ, specifically citing *Intonation Curves* [16, p. v; 17, p. vii].

It has probably been generally assumed that systems of intonation description were necessarily first advanced on the basis of impressionistic or invented data [13, p. 187] and only later put into relation with empirical data from actual speakers. In fact, as this study shows, 'elaborate, accurate and careful' intonation data was available for English, French and German several years before the earliest attempts at formal intonation description in any of the three languages.

8. Acknowledgements

The author is grateful to John Coleman (University of Oxford, UK) for discussions in the planning stage of this research, and to Petr Roesel (Johannes Gutenberg-Universität Mainz, Germany) for valuable help during a protracted international search for a copy of the 'English Conversation' recording.

9. References

[1] D. Jones, *Intonation curves. A collection of phonetic texts, in which intonation is marked throughout by means of curved lines on a musical stave*. Leipzig ; Berlin: B G Teubner, 1909.

[2] B. S. Collins and I. M. Mees, *The real Professor Higgins: The life and career of Daniel Jones*. Berlin: Mouton de Gruyter, 1999.

[3] A. Lüdeling and Merja. Kytö, Eds., *Corpus Linguistics*, vol. 1. De Gruyter Mouton, 2008.

[4] F. Grosjean, ‘Spoken word recognition processes and the gating paradigm’, *Perception & psychophysics*, vol. 28, no. 4, pp. 267–83, 1980.
[5] D. Jones, *An outline of English phonetics*. Leipzig und Berlin: B.G. Teubner, 1918.
[6] J. Obata and R. Kobayashi, ‘A direct-reading pitch recorder and its applications to music and speech’, *The Journal of the Acoustical Society of America*, vol. 9, no. 2, pp. 156–161, Oct. 1937.
[7] P. Boersma and D. Weenink, Praat: doing phonetics by computer. [Computer program]. Version 6.1.08, retrieved 5 December 2019 from http://www.praat.org/ 2019
[8] The Scottish Consortium for ICPhS 2015, Ed., *Proceedings of the 18th International Congress of Phonetic Sciences. Glasgow, UK*. Glasgow: Glasgow University Press, 2015.
[9] S. Strömbergsson, ‘Today’s most frequently used F0 estimation methods, and their accuracy in estimating male and female pitch in clean speech’, in *Proceedings of INTERSPEECH 2016*, San Francisco, USA, 2016, pp. 525–529.
[10] M. Huckvale, ‘Speech Filing System’, 2013. [Online]. Available: http://www.phon.ucl.ac.uk/resource/sfs/. [Accessed: 20-Mar-2015].
[11] D. Talkin, ‘A robust algorithm for pitch tracking (RAPT)’, in *Speech coding and synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. New York: Elsevier, 1995, pp. 495–518.
[12] R. Weeks, Review of [1], *Le Maître Phonétique*, pp. 82–83, 1910.
[13] J. ’t Hart, R. Collier, and A. Cohen, *A perceptual study of intonation: An experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press, 1990.
[14] H. Klinghardt and M. de Fourmestraux, *French Intonation Exercises. Translated and adapted by M.L. Barker*. Cambridge: W. Heffer & Sons, 1923.
[15] A. Cruttenden, ‘The origins of nucleus’, *Journal of the International Phonetic Association*, vol. 20, no. 1, pp. 1–9, 1990.
[16] H. Klinghardt and G. Klemm, *Übungen im englischen Tonfall*, etc. Cöthen, 1920.
[17] H. E. Palmer, *English intonation, with systematic exercises*, Cambridge: Cambridge : Heffer, 1922.