The Punch and Judy man: one or more speakers? A study of speaker variability

Sandra P. Whiteside (1), Gavin Dempster (2) and Sheila Williams (2).

(1) Centre for Language Engineering, University of Sheffield, Department of Speech Science, 18 Claremont Crescent, Sheffield, S10 2TA, United Kingdom.

(2) Centre for Language Engineering, University of Sheffield, Department of Computer Science, Regent Court, 211 Portobello Street, Sheffield, S1 4DP, United Kingdom.

1. ABSTRACT

Speaker variability is investigated by contrasting the voices of characters in an unscripted Punch and Judy show created by a single puppeteer (Bob Arkley - BA), without the use of any mechanical or other voice transformation devices. The three characters studied are Barnacle Bill the sailor, Mr Punch and Judy. The two investigators who were not present at the show experienced no difficulty in distinguishing auditorily between the characters in contrastive contexts, apparently on the basis of accent and voice quality. We present here the preliminary results of a variety of analyses to identify the characteristics which may be contributing to the individual perceived identities of the characters.

2. BACKGROUND & INTRODUCTION

This pilot study forms part of a much wider investigation into speech and speaker characterisation (ESPRIT Working Group No. 6298, VOX) and is intended partly as an exercise in the integration of articulatory, perceptual and acoustic approaches. The work reported here is a preliminary study of a data set which is supplementary to a core speaker style database [1].

The data used in this study represents three perceptually distinct voices produced by a single speaker (BA). The issue here is whether we can characterise each of the voices by examining fundamental frequency patterns, formant frequency patterns, voice quality and accent attributions and whether each of the voices can be characterised by a unique set of parameters. However it must be acknowledged that the acoustic-phonetic method here represents only a single and simplified approach to the much wider question of speaker recognition which has become central to forensic speaker recognition. (see [2]).

3. DATA & ANALYSIS

3.1 THE DATA

The speech data examined here was recorded on DAT tape at an impromptu performance of a Punch and Judy show. The show took place in a large rehearsal room during a break in rehearsals of an amateur dramatic production. The audience were members of the drama group, their families and friends, and ranged in age from about 8 to 70 years. The show is adaptive and performed unscripted following "story-lines". Only the voices of characters created by the puppet-master are analysed here.

3.1.1 The Speaker. The puppet-master is aged 44 and uses no artificial voice transformation devices. He has been performing Punch and Judy shows and developing the various characters since about the age of 7.

3.1.2 Transcription and selection of data samples. For this investigation, we first produced a full orthographic transcription of the whole show, which lasted approximately 35 minutes. Minor and major prosodic boundaries were assigned perceptually using Reyelt's method [3], displaying a high level of agreement between all three authors.

For each of the three characters we identified the longest single continuous utterance which could be found with minimal background noise, for a Vocal Profile Analysis (VPA, [4, 5]), phonetic realization fundamental frequency and vowel formant frequency analyses. However, for the purposes of some of the analyses (e.g. VPA) relatively extensive passages from each voice are to be preferred. For formant analyses, recordings must have very low background noise levels. This therefore largely limited the choice of data samples for the current investigation.

Therefore, although the VPA for each character was not limited to these selected passages, they formed the main focus of the analyses. The orthographic transcriptions of the selected passages can be found in Appendix I. Due to the lack of constraints on data production, the selected passages varied in length. However, the average word length for each of the characters were similar in length [6].Phonetic transcriptions were prepared for each of these selected passages.

3.2 THE ANALYSIS TECHNIQUES

3.2.1 Accent attributions. Our informal accent attributions were based on our impressions from the full 35 minute performance.

3.2.2 VPA and phonetic realization analysis. Vocal Profile Analysis (VPA) is a formal protocol [4, 5], to record perceived attributes of voice quality in terms of articulatory settings. Our analysis was limited to the recorded samples and was therefore performed without visual cues. As we have relatively little experience with this technique, we limited our responses to the allocation of laryngeal features to either neutral, moderate or extreme categories rather than using more detailed estimations of scalar degrees. Phonetic realizations were deduced from phonetic transcriptions.

3.2.3 F0 analysis. Fundamental frequency analyses were performed using a Kay CSL Model 4300 speech analysis workstation on selected portions of the data samples shown in Appendix I. For this process, periods of silence, audience noise which obscured a character's voice and whispered speech by character was removed for analysis. In addition the samples had to be divided up into shorter sections for analysis on the KAY CSL, which can only cope with short stretches of speech. The division of the speech samples is also given in Appendix I. The following denote the total lengths of speech analysed for each character: Barnacle Bill: 25.087 seconds; Mister Punch: 15.461 seconds and Judy: 29.736 seconds. The settings (default values) and methods used for the F0 analysis were as follows: i) Frame length = 20 msec; ii) Frame advance = 20msec; iii) Pitch voicing cutoff (max. zero cross) = 25%; iv) Zero crossing clipping level = 15; v) Pitch peak threshold (minimum peak) = 100; vi) Continual adjustment of analysis range to accommodate the wide variability of the characters' voices; vii) Manual checking of the computed values to ensure accuracy of results.

3.2.4 Vowel formant frequency analysis. An analysis of vowel formant frequencies was performed on the selected speech samples shown in Appendix I, using LPC analysis and wide band FFT spectrograms derived by the Kay CSL Model 4300 system. This use of two analysis methods was used to monitor the accuracy of the LPC analysis whereby the results of the LPC analysis were overlaid onto FFT spectrograms. The settings (default values were used except where otherwise specified) used for the LPC analysis were as follows: i) a frame length of 20 ms; ii) a filter order of 12 except where for weaker signals, a filter order of 14 was used; iii) a pre-emphasis weighting on 0.9; iv) window weighted analysis; v) Autocorrelation method. Three readings of F1, F2 and F3 and their bandwidths (where possible) were taken for the vowels [ɜ], [i], [u] and [æ] for each character at quarterly intervals over the length of the vowel in particular word contexts (see Table 3). Average values to the nearest 10 Hz were then calculated. (See Table 3).

4. RESULTS AND DISCUSSION

4.1 ACCENT ATTRIBUTIONS

From the initial informal analysis over the complete recording, all three authors agreed that in sum Judy presented an RP accent, the accent of Mr Punch was variously described as (West) Midlands and Birmingham and Barnacle Bill was attributed to Bristol, West Country (naval). The more detailed analysis of phonetic realizations was limited to the selected utterances.

Indicants of an RP accent in the phonetic realization of Judy's speech include consistently the velar nasal [ŋ] word finally in the *-ing* morpheme, [ɪ] as the word-final segment in 'quickly', 'baby' and 'nappy' and the production of the vowel [ʌ] in words like 'Punch', 'stuffing'. However, the utterances analysed here, on the whole, reflect the informal RP of a native speaker rather than the more precise enunciation indicative of the "adoptive-RP" defined by Wells [7]. Variants which support this interpretation include place assimilation (e.g. '... what mister ...' = [p m]), the consistent dropping of unstressed word-initial *h*- ('..see him..' - ['si ɪm]) and Sibilant Yod Assimilation ('..as you ..' - [əʒu]).

Again following Wells [7], we found confirmation to support the attribution of Barnacle Bill to the Bristol region in the preponderance of rhoticity in the vowels of 'world' and 'downstairs' for example. In addition the following realizations were found which are typical of the Bristol vowel system - 'whole'- ['ho:ɬ], 'wide' - ['wo̞ɪd˺], 'morning' - ['mɔɻnɪn] and 'that' - ['ðaʔ]. Glottalizing of word-final /t/ was also found in words like 'what', 'secret', 'that' and 'right'.

The speech of Mr Punch showed some but few of the phonetic realizations normally associated with a Birmingham accent. For example, the following vowel realizations were found - 'know' - ['nʌʊ], 'hair' - ['hɛ:] and 'me' - [mɔi]. The expected velarized nasal plus plosive ending ([ŋg])for the *-ing* morpheme was not found. We deduced that most of the Birmingham flavour of the accent came from vocal quality and a tendency to raise and slightly diphthongise vowels as exemplified above.

Other features present throughout the speech of all three characters were the lack of distinction between clear and dark /l/, Yod Coalescence (J: 'would you' ['wʊʤu]; BB: 'did you' - ['dɪʤu]; MP: 'mind you' - ['maɪnʤu]), H - dropping and /t/ glottalization. However, the latter two features were not consistent in Mr Punch's data sample.

4.2 VOCAL QUALITY

The VPA indicated differences in the perceived preferred laryngeal settings of the three characters. Judy displays modal voice with intermittent falsetto and moderate to extreme larynx lowering. Barnacle Bill has clearly modal voicing with the intermittent raising of pitch and a neutral larynx quality. For Mr Punch it was difficult to ascertain whether the phonation was true falsetto or extremely high modal voicing and the larynx position appeared to be extremely raised. There was some audible whisperiness in all three voices, extreme only in the case of Judy.

4.3 ACOUSTIC ANALYSIS

4.3.1 Fundamental Frequency Analyses. The detailed results of the data samples (Appendix I) can be found in Table 1. A summary of the results can be found in Table 2. These values are represented longitudinally for all three speakers in Figure 1 on a log scale.

Filename (Refer to App. I)	Mean F0 (Hz) (s.d. Hz)	No. of periods over which F0 is calculated
BB1	163 (30)	74
BB3	171(31)	40
BB4	196 (57)	138
BB5	234 (78)	76
BB6a	194 (58)	61
BB6b	267 (65)	68
BB7	175 (33)	64
BB8	258 (96)	147
BB9	188 (42)	52
BB10	170 (34)	35
BB11	192 (40)	140
BB12	202 (64)	76
BB13	261 (60)	50
MP2	308 (42)	117
MP3	273 (34)	109
MP5	366 (54)	109
MP6	346 (51)	80
MP7	279 (38)	135
MP8	252 (51)	147
J1	219 (69)	140
J2	248 (48)	82
J3	168 (32)	105
J4	200 (33)	90
J5	192 (30)	129
J6	216 (52)	104
J7	193 (26)	160
J8	220 (47)	98
J9	217 (55)	111
J10a	242 (58)	48
J10b	304 (59)	51
J11	179 (29)	81
J12	312 (64)	38
J13	219 (51)	75

Table 1 - Fundamental Frequency Analysis Results.

Character	Total length of data sample (seconds)	Overall mean (Hz)	Overall s.d.(Hz)
Barnacle Bill (BB)	25.1	209	71
Mister Punch (MP)	15.5	299	60
Judy (J)	29.7	214	55

Table 2 - A Summary of the Fundamental Frequency Analysis.

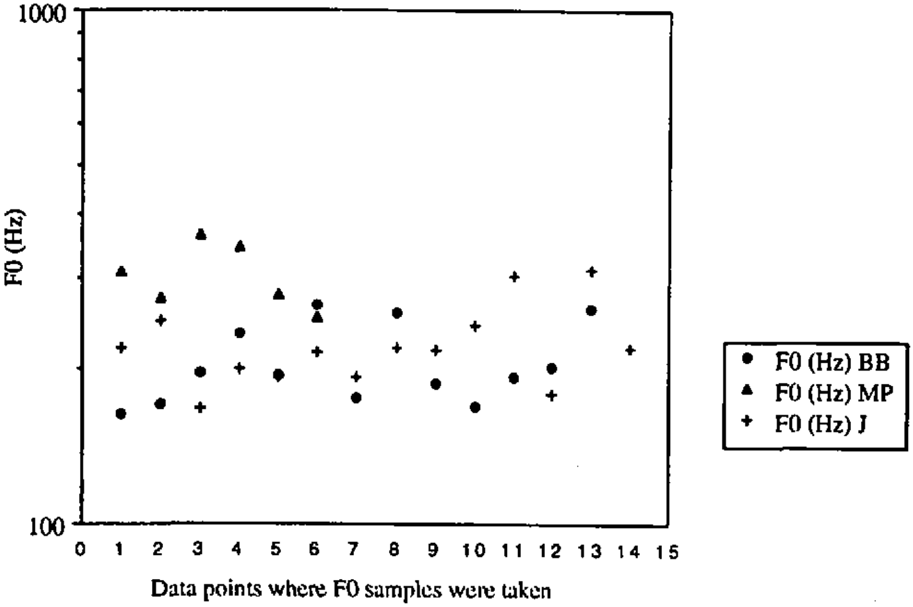

Figure 1. Longitudinal F0 data for Barnacle Bill (BB), Mr Punch (MP) & Judy (J).

The results in Tables 2 and 3 and Figure 1 indicate that the perceived higher pitch range for Mr Punch was supported by the overall F0 mean. However, F0 does not distinguish markedly between the average pitch perception attributed to Barnacle Bill and the lower pitch attributed to Judy, where the difference in the overall mean between the two characters is only 5 Hz. However, Barnacle Bill displays greater variability in F0 than Judy, which may explain why Judy was perceived as being lower pitched. What is interesting to note is that for Barnacle Bill (209 Hz) and Judy (214 Hz), the overall mean F0 values are close to the mean value of 233Hz observed for trained women actresses by Cowan in 1936, as cited by Linke [8]. Contrastingly, the mean F0 value for Mr Punch (299 Hz) clearly exceeds this. It is likely that much of the F0 variability is also dependent upon the content of the performance and the nature and extent of the characters interacting with the audience.

4.3.2 Vowel Formant Structure

The structure of the single clearest example from each character, of each of the four stressed vowels analysed, is shown in Table 3. Formant frequency values (and where possible formant bandwidths (b_n)) are given in Hz Given the high levels of background noise it was often difficult to derive formant frequencies and given the spontaneous and unscripted nature of the speech, it was not possible to control phonetic context. A two dimensional vowel chart [9] showing F2-F1 (Hz) versus F1 (Hz) for a group of selected vowels is given in Figure 2. The selected vowels are highlighted (*) in Table 3 and the Appendix.

Vowel	Character & sample	F1 (Hz)	b1 (Hz)	F2 (Hz)	b2 (Hz)	F3 (Hz)	b3 (Hz)	Context
ɜ	BB5	450	-	1210	-	2040	-	'world'*
	MP3	590	-	1290	-	2500	-	'girls'*
	J1	570	-	1450	-	2350	-	'girls'*
i	BB6	400	-	1925	-	2700	-	'see'*
	MP2	320	100	2180	130	2780	120	'see'*
	J5	380	230	2290	180	2860	250	'see'*
u	BB12	470	140	1470	220	2270	130	'you'*
	MP2	675	-	1550	-	2500	-	'you'*
	J4	410	110	1340	220	2270	280	'do'*
	J5	370	70	1640	80	2420	260	'doing'
æ	BB4	710	160	1610	170	2630	100	'that'
	BB8	780	50	1550	180	2640	90	'that'*
	BB11	760	30	1380	100	1980	219	'fact'
	MP7	660	-	1570	-	2470	-	'having'*
	MP5	840	270	1580	240	2730	200	'and'
	J9	760	200	1490	150	2410	240	'whacking'
	J9	800	200	1330	2000	2520	170	'that'*

Table 3 - Vowel Formant Values for the three characters.

The ratio of the formant values for the example of [ɜ] from Judy's speech indicate a relatively neutral vocal tract setting and the position of the vowel in Figure 2 indicates this neutral quality. From this, we can estimate a rough approximation of the vocal tract length from the formant frequency values of F1 to F3, using the Pipe model of the vocal tract [9]. This gives us an estimate of 18.7 cm for Judy's vocal tract length, which exceeds the average male vocal tract length of 17 cm [10]. It must be stressed that this is only a rough approximation and that one would need to consider factors such as: (i) the "radiation impedance" of the lips [9] and (ii) energy losses in the vocal tract through cavity walls and soft tissue, for example, for a more precise estimation. However, the evidence lends some support to our perceptions of Judy having a lowered laryngeal setting, thus effecting the perception of a 'deeper' voice.

Mr Punch's formant frequencies for F1 and F2 of the vowel [ɜ] suggests a vowel quality which is a little more open than neutral, if we combine an acoustic and auditory representation [9] as shown in Figure 2. Barnacle Bill however shows a closer quality for [ɜ] than the other two characters. From Table 3 we see that there is a marked difference in the third formant of Barnacle Bill at 2040 Hz, which is lower than either Mr Punch or Judy's F3 values. This can be explained by its rhotic nature and supports the attribution of Barnacle Bill's accent to Bristol.

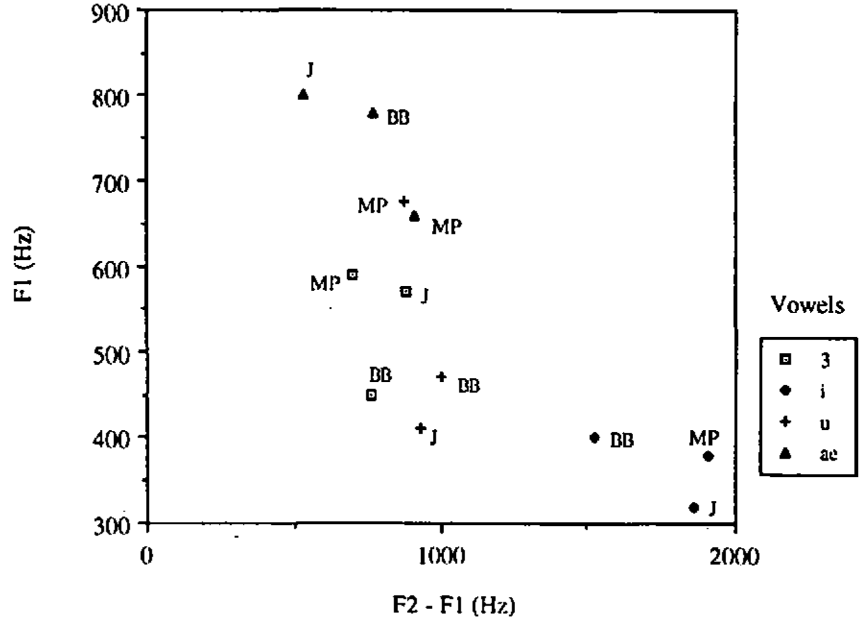

Figure 2. Vowel chart of selected vowels for the three characters

For [i], the vowel chart shows that Mr Punch has a more fronted vowel than the other two characters with Judy having the closest vowel quality, with Barnacle Bill showing a more centralised vowel quality. However, it is worth saying at this point, that these differences may be due to the different phonetic contexts they were spoken in. The disparity between the phonetic contexts is further illustrated by the range of formant frequency values for [u] which shows that Judy has the highest and most backed vowel of the three characters. Barnacle Bill shows a more fronted and more open vowel quality for [u] which maybe explained by its phonetic context where it precedes [æ]. Mr Punch's realisation of [u] suggests that it is open and backed. For [æ] Judy shows the most backing and openness for this vowel. While Barnacle Bill shows an open but more fronted vowel quality close to an [a] position on the IPA chart [11]. Mister Punch shows a much closer though still backed vowel quality (which in fact falls close to his realisation of [u]). What is interesting to note from Figure 2 is that of the three characters, only Judy's formant frequency data forms a vowel triangle with a neutral vowel falling within this triangle.

5. SUMMARY AND CONCLUSIONS

This is a preliminary study of a relatively large data set. From these initial analyses it seems that there are a multiplicity of factors which contribute to the 'different' speaker characteristics of the three characters. These include: differences in perceived laryngeal settings; different accent manifestations which are not always constant for the whole portrayal of each character; differences in fundamental frequency and formant frequency characteristics. In spite of the clear perceptual distinction between the three characters, clues remain to the speaker's underlying accent and speaking style. We plan a more detailed analysis of this data as part of our ongoing work on speech and speaker characterization.

6. REFERENCES

[1] V. Péan, S. Williams and M.Eskénazi, *The Design and Recording of ICY, a Corpus for the Study of Intraspeaker Variability and the Characterization of Speaking Styles*, Proc. EUROSPEECH '93, Berlin, Germany, 1993.

[2] H. J. Künzel. *Current Approaches to Forensic Speaker Recognition*, Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny Switzerland, pages 135-141. 1994.

[3] M. Reyelt, *Experimental Investigation on the Perceptual Consistency and the Automatic Recognition of Prosodic Units in Spoken German*, Proc. ESCA Workshop on Prosody, Lund, Sweden, Lund University Working Papers 41, 1993.

[4] J. Laver, *The Gift of Speech*, Edinburgh University Press, 1991.

[5] J. Mackenzie Beck, *Vocal Profile Analysis: User's Manual*, Queen Margaret College Edinburgh and University of Edinburgh, 1991.

[6] S. Williams, S. Whiteside, G. Dempster, *The Punch and Judy man: Speaker or speakers ?*, Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny Switzerland, pages 131-134, 1994.

[7] J. C. Wells, *Accents of English* (Volumes. 1 and 2), Cambridge University Press, 1982.

[8] C. E. Linke, A study of pitch characteristics of female voices and their relationship to vocal effectiveness, *Foilia Phoniatrica, 25*, 173 - 188, 1973.

[9] P. Ladefoged, *A Course in Phonetics* (3rd Edition), New York: Harcourt, Brace, Jovanovitch, 1993.

[10] P. Lieberman, S. Blumstein, *Speech physiology, speech perception and acoustic phonetics*, Cambridge University Press, 1988.

[11] International Phonetic Association, *International Phonetic Alphabet..* revised to 1993.

Appendix I - The data used for the analyses

Barnacle Bill

"All right I'll just make sure there's nobody listening out the back. Hang on a minute. No it's all right, there's nobody there. Now now listen, the secret is, that Mister Punch, is the laziest man in the whole wide world. All he ever does is sleep. Do you know. I went to see him this morning. I said come on Mister Punch, I said. We're off to Rotherham to do a show. But do you know what he said. Do you know what that old rapscallion said to me. I'll tell you. He said (snoring sounds). Just like that. Cos he was fast asleep. In fact if you were to have a listen, you'd hear him snoring downstairs, right now. You have a listen, and see if you can hear him. (next bit very unclear) Hang on. (pause) There did you here that ?"

The division of the data used in the analysis was as follows (each sample is represented in Fig. 1 as a data point):

BB1 "I'll just make sure there's nobody listening out the back" (Length 1.912sec.; Remarks: Crackly.)

BB3 "now the secret is" (Length 1.159sec.)

BB4 "that mister punch is the laziest man in the whole wide" (Length 3.227sec. Remarks: Laughter from an audience member is clearly audible over "man in the").

BB5 "world all he ever does is sleep" (Length 1.785sec.)

BB6 "do you know I went to see him this morning I said come on mister punch I said" (Length 3.085sec.)

BB7 "we're off to Rotherham to do a show" (Length 1.453sec.)

BB8 "but do you know what he said do you know what that old rapscallion said to me" (Length 3.529sec.)

BB9 "I'll tell you he said" (Length 1.343sec.)

BB10 "he was fast asleep" (Length 1.001sec.Remarks: A bit crackly and quiet.)

BB11 "in fact if you were to have a listen you'd hear him snoring downstairs" (Length 3.214sec. Remarks: Audience member can be heard breathing over "you were to have a listen".)

BB12 "right now you have a listen and see if you can hear him" (Length 2.087sec. Remarks: "right now" is quiet; "you have a listen" is 'normal' loudness; "and see if you can hear him" is also quiet, and slightly whispered.)

BB13 "there did you hear that" (Length 1.292sec. Remarks: Loud voice, which echoes off the hall walls.)

Total Length: 25.087 secs.

Mister Punch

"Right get off (?). I'll teach her to wallop me on the head. Mind you. You, should see my little boy, boys and girls. Do you know, he's ever so handsome, just like me. And he's ever so intelligent just like me. And he's unbelievably quiet, just like me. Mind you. He does take after his mummy for one thing. He's got terribly messy hair and he hates having his hair combed. Oh look, look, look . He's coming. Aah"

The division of the data used in the analysis was as follows (each sample is represented in Fig. 1 as a data point):

MP2 "wallop me on the head mind you you should see my little"(Length 2.558sec. Remarks: Quiet.)

MP3 "boy boys and girls do you know he's ever so handsome"(Length 2.336sec. Remarks: Quiet. Crackly. "just like me" removed because of audience member's laughter over it.)

THE PUNCH AND JUDY MAN

MP5 "and he's unbelievably quiet just like me"(Length 2.794sec. Remarks: Relatively high amplitude is due to unbelievably'.)
MP6 "does take after his mummy for one thing"(Length 1.643sec. Remarks: Quiet. "mind you" removed because of audience member's laughter over it.)
MP7 "he's got terribly messy hair and he hates having his hair combed" (Length 2.957sec. Remarks: Quiet.)
MP8 "oh look look look he's coming aah" (Length 3.173sec. Remarks: Quiet."aah" is approximately 1sec long)
Total length: 15.461 seconds

Judy
"Quickly boys and girls while mister punch is gone, would you all do me a favour ?" "Yes, I mean, well you heard what Mister Punch said, about babysitting means that you sit on the baby. That would be a very dangerous thing to do, wouldn't it ?" "Yes. So if you see him doing anything stupid, or well anything dangerous to the baby - you know, like er, well like setting fire to the baby's nappy, or or or, or stuffing him into a washing machine, or any ... or whacking him, yes. Anything like that , would you all shout me as loudly as you can." "Oh good. Then I can come back and sort him out. Right. I'll see you all later on then boys and girls. Bye bye"

The division of the data used in the analysis was as follows (each sample is represented in Fig. 1 as a data point):
J1 "quickly boys and girls while mister punch is gone would you all do me a favour" (3.382sec. Remarks: Quiet.)
J2 "I mean well you heard what mister punch" (Length 1.716sec. Remarks: Quiet. Noisy. "yes" removed because of audience noise. "said" removed because of audience member murmuring over it.
J3 "about babysitting means that you sit on the baby" (Length 2.450sec. Remarks: Quiet.)
J4 "that would be a very dangerous thing to do wouldn't it" (Length 2.074sec. Remarks: Quiet.)
J5 "so if you see him doing anything stupid or" (Length 2.924sec. Remarks: Quiet. "yes" removed because of audience responses over it.)
J6 "well anything dangerous to the baby" (Length 2.551sec. Remarks: Audience member can be heard faintly talking in the background. "to the baby" is said fairly quietly.)
J7 "you know like er well like setting fire to the baby's nappy" (Length 3.408sec. Remarks: A child's voice can be heard over "baby's nappy"; however,Judy is talking fairly loudly at this point.)
J8"or or or stuffing him into a washing machine" (Length 2.265sec. Remarks: "or or" removed because of child's voice over it. "machine" has a fairly quiet child's voice over it.)
J9 "any ... or whacking him yes anything like that" (Length 2.527sec.)
J10 "would you all shout me as loudly as you can" (Length 2.153sec.)
J11 "oh good then I can come back and sort him out"(Length 1.801sec. Remarks: First part of "oh" removed because of audience noise. "sort him out" has an audience member talking over it, but fairly quietly.)
J12 "right I'll see you all later on then boys and girls bye bye" (Length 2.485sec.)
Total length: 29.736 seconds