

PROCESSING ‘YUP!’ AND OTHER SHORT UTTERANCES IN INTERACTIVE SPEECH

Nick Campbell, John Kane and †Helena Moniz

Centre for Language and Communication Studies
Trinity College Dublin
Ireland

nick@tcd.ie, kanejo@tcd.ie, helenam@l2f.inesc-id.pt

ABSTRACT

The detection of short utterances in conversational or interactive speech is essential to the proper processing of meaning in spoken interaction. Short, simple utterances are extremely common, and because of their highly variable prosody, carry many different forms of subtle interpersonal information. This paper reports on our approach to this problem and describes some corpora we are working with as well as the results of an analysis showing overlapping segments to be significantly different in their prosodic characteristics.

Index Terms— Spoken interaction, non-vocal outbursts, multimodal processing, conversational speech corpora

1. INTRODUCTION

Conversational speech is interactive. Unlike read speech or broadcast speech, it relies on constant feedback from the listener(s) and is therefore characterised by a fragmented and repetitive form of speaking, as the meaning is adaptively and collaboratively built up by both partners throughout the interaction [1, 2]. This form of speech typically contains many short utterances and frequent changes of speaker as interactivity is high throughout.

Figure 1 (from [3]) plots the speech activity of two partners throughout a 30-minute telephone conversation. It shows speech density, measured as a ratio of speech to non-speech timings per utterance for time-aligned utterances of each speaker. The number of utterances per speaker differ, but the lines in the lower part of the plot show average density for each moment of the conversation. It is clear from the plot (a) that both partners are active throughout, and (b) that the density of one varies as a reciprocal of that of the other.

Figure 2 plots the averaged speech density measures for three conversations. Here the plot for one speaker of each pair is inverted for graphical purposes to allow a more immediate comparison of their reciprocity. At the top is the pair shown in Figure 1, in the middle that of the best-aligned pair of speakers ($r=-0.752$) and at the bottom that of the worst-aligned pair

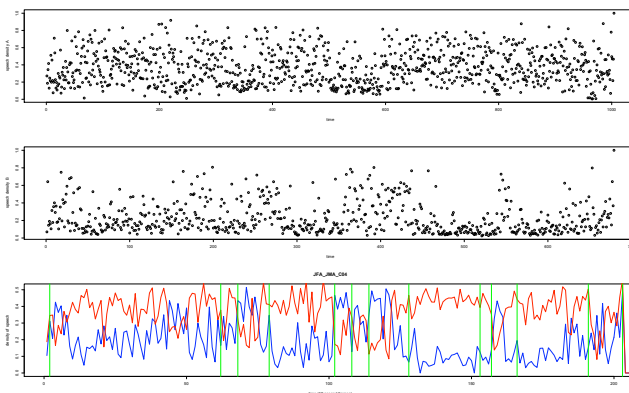


Fig. 1. Speech activity plotted as the ratio of speech to non-speech per utterance for two speakers in a telephone conversation; a high value here represents almost continuous speech activity, with long utterances and short gaps, whereas a low value indicates a brief utterance, typically giving feedback, in a long period of non-speech. The upper plot shows speaker A ($n=1004$ utterances), and the middle plot speaker B ($n=679$ utterances) on the same time axis of 30-minutes. The difference in the number of utterances is clear, as is the complementarity of the distributions. The lower plot shows the windowed and 10-second average time-aligned values for both speakers superimposed. Vertical green bars mark a change of topic. The reciprocity of their speech activity is clearly shown.

of speakers ($r=-0.120$) out of one hundred recorded thirty-minute telephone conversations. It is striking even in the worst case how closely aligned these plots appear.

The point being illustrated by these two figures is that *both* partners are each highly active throughout all parts of every conversation and that the number of short utterances is correspondingly extremely high. These short utterances can be very repetitive and vary only minimally with respect to their transcription. The information they carry is primarily in their prosody, which must be specially processed separately.

†FLUL/INESC-ID, Portugal.

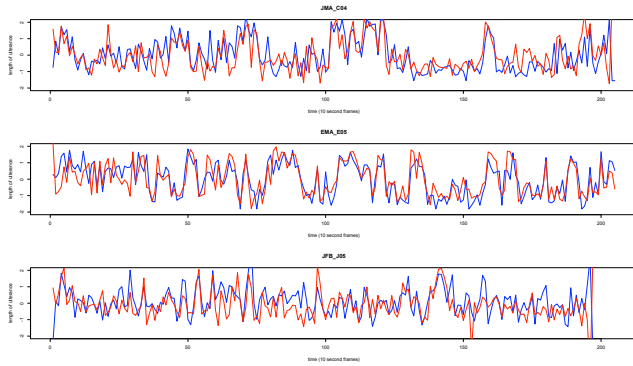


Fig. 2. Comparing three different thirty-minute conversations, showing the speech density plots for Speaker A with the inverse speech density plots for each Speaker B. This view of conversational activity allows us to simply compare and measure the difference in changes of utterance density across time. The top plot shows JFA as in Figure 1 ($r=-0.624$), and the lower two plots show the most (JFA_EMA_E05, $r=-0.752$) and least coordinated (JFA_JFB_J05, $r=-0.120$) examples from the corpus as determined by this measure.

2. THE ESP AND D64 CORPORA

The data reported above are taken from Japanese speech (part of the JST Expressive Speech Processing (ESP) Corpus [4]). The ESP corpus was recorded in Japan between 2000 and 2005 and includes 1,500 hours of interactive speech captured in everyday situations using head-mounted microphones in conjunction with body-worn mini-recorders. In the ESP_C telephone speech subset of the corpus, we found that the hundred most common words accounted for more than half the total utterance count of the corpus. These were equivalent to the English words “yup”, “yeah”, “uhuh”. “umhm”, etc., and were often simply repeated (“yeah yeah”, “yeah yeah yeah”, up to a maximum of seven repetitions) and included much laughter though the conversations were rarely humorous. All-wood has reported very similar short-utterance distributions for several other languages [5].

Certain common interjections such as “honma?!” (an Osaka dialect equivalent to the English word ‘really’) were produced under an extremely diverse range of prosodic conditions resulting in each expressing an entirely different interpersonal message. These words would be transcribed identically by a speech recogniser and could not be adequately expressed by a speech synthesiser without special annotated markup on the input, and they remain a challenge with respect to discourse classification [6, 7]. We counted 3,500 tokens of ‘honma’ which were perceptually classified into more than 20 different types of pragmatic utterance, with finer degrees of subtlety distinguished within each class [8].

Figure 3 shows speech activity from a five-person conversation from the follow-up FreeTalk Multimodal Conver-

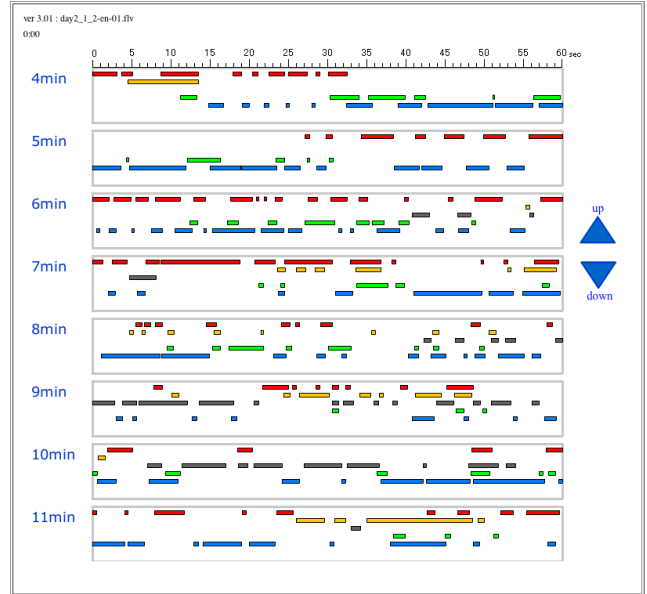


Fig. 3. Plots of time-aligned speech activity from a 5-party conversation (part of the FreeTalk Corpus) showing how fragmented and interactive such speech activity typically is. Each speaker is plotted in a different colour. Whereas clear sections of continuous speech can be seen in each one-minute time-slice, the number of short often-overlapping utterances far outweighs the number of longer utterances.

sation Corpus [9], in which the main language was English, and reveals similar disproportionality between long and short utterances. We can assume that the longer sections of continuous speech represent well-formed grammatical utterances which could be adequately processed by a speech synthesiser or recognition system. However, it is clear from the figure that our present technology, with no prosody or speaking-style-specific analysis would fail to adequately process *more than half* of the utterances here.

Our present work extends this interactive speech data collection to include multimodal sources including motion-capture, high-definition video, 360-degree industrial video, and multiband audio recordings. Being based now in Dublin, it centres around English as a common language and includes participants from a broad range of cultural backgrounds. The work extends previous audio-only work carried out on a very large corpus of conversational speech towards the combined audio and video processing of a large multimodal corpus of social interaction.

The D64 Corpus, recorded in Dublin in 2009, incorporated recordings from 12 audio lines, 5 high-definition video, 2 360-degree industrial video and 6 Optitrak motion-capture devices. The recordings took place over a period of two days during which 5 participants engaged in a series of unstructured and unscripted informal and casual social interactions.

The third author performed an analysis of two sections of this corpus. The first consisted of recordings made during the initial setting-up of the equipment, and the second of recordings made on the same day during more relaxed social chatting. She selected 1,228 short utterances (many of them just the one word "yup!") from this material by introspective listening and excised the waveforms representing these utterances from the recordings of head-mounted or body-worn microphones for each speaker. There were 480 samples from the first session, and 748 samples from the second session.

3. SHORT UTTERANCES IN TWO TYPES OF CONVERSATIONAL SPEECH

In all we distinguished 12 semantic types of utterance in the samples selected for further analysis including Discourse markers (m) *Like, so, you know*, Grunts (g) *Humhum*, Backchannels (b) *Okay, yes, right*, Interjections (i) *God!, Perfect!*, Onomatopoeia (o) *Grrrrr*, Filled pauses (f) *Um, uh, er*, Elongations (e) *The:*, Repetitions (r) *The the; I I*, Substitutions (s) *How he for what she*, Deletions (d) *But they're (abandoned linguistic material)*, Truncations (t) *Abso- (for absolutely)*, and Complex sequences of disfluencies (c) *The only uh the only*.

The most frequent events that we encountered in these recordings were backchannels, filled pauses, complex sequences of disfluencies, repetitions, grunts, and interjections as enumerated in Table 1 below. Acoustic feature analysis was performed for each waveform segment, extracting values representing minimum, mean, and maximum of power and pitch (F0), position of the F0 peak, amount of voicing, values of the first and second harmonics, third formant, spectral tilt, and normalized duration of the events. A principal component analysis showed these to be independent and showed the

Table 1. Showing types and counts of selected short utterances per speaker for four speakers. Total sample counts are shown on the right and at the bottom.

	s1	s2	s3	s4	
backch (b)	92	55	84	235	446
complex (c)	29	3	31	65	128
delet (d)	6	0	9	11	26
elong (e)	9	1	8	21	39
filler (f)	8	6	48	81	143
grunt (g)	37	37	12	26	112
interj (i)	25	15	9	61	110
markers (m)	8	2	11	30	51
onomat (o)	0	0	0	2	2
repet (r)	24	0	45	57	126
subs (s)	3	0	1	4	8
trunc (t)	6	1	4	6	17
	247	120	262	599	1228

first component to be most influenced by power, the second by fundamental frequency, and the third by the voice quality parameters and duration. The first five components together accounted for more than 70% of the variance. A support vector machine given the first 5 components was able to correctly predict utterance type at 42.7% using 10-fold cross validation. For this 12-category task, chance prediction rate would be around 9%.

Importance of components:

	Cmp.1	Cmp.2	Cmp.3	Cmp.4	Cmp.5	Cmp.6
St. Dev	1.744	1.622	1.212	1.188	1.124	0.956
Prop of Var	0.217	0.188	0.105	0.100	0.090	0.065
Cumulative	0.217	0.405	0.510	0.611	0.701	0.767

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
fmean	-0.204	-0.451		-0.301	-0.202	-0.277
fmax	-0.147	-0.483		-0.312	-0.181	-0.205
fmin				0.477	-0.457	-0.483
fpct		-0.118	-0.224		0.497	-0.400
fvcd	-0.261		0.344	-0.354	0.360	
pmean	-0.443	-0.235		0.337		0.118
pmax	-0.297	-0.326	-0.236	0.365		0.231
pmin	-0.404		0.369	0.268		
ppct		-0.166		0.153	0.466	
h1h2	-0.264			-0.320	-0.312	0.111
h1a3	-0.390	0.359	-0.318			
h1	-0.373	0.345				0.157
a3	0.200	-0.181	0.496	0.101		0.413
dn		-0.255	-0.502			0.438

4. VOICE QUALITY ANALYSIS

It was noted that mean pitch and voice-quality values were different for overlapping speech segments and significant differences were also found in acoustic parameters between the two sessions. Accordingly, we performed a more detailed analysis of voice quality using glottal gradient parameters derived from inverse filtering instead of the raw waveform [10].

We focussed on the RCG parameter (similar to H1a3) which provides a measure of spectral slope and corresponds to the rate of closure of the glottis, and on GOG which measures first-formant prominence. A lax vocal tract apparatus or losses at the glottis (which can result from a different vocal fold setting when the speaker is more relaxed) can cause a weakening of the energy at the first formant. Lower RCG and GOG values indicate a tense voice quality and high values indicate a breathier/more lax quality.

A comparison of the results confirmed that RCG was significantly lower for overlapping segments ($t=3.118$ p -value = 0.0018) and that it was overall higher in the second session than in the first ($t=3.792$). This finding was reinforced by the GOG measure which was also lower for overlaps ($t=2.512$) and higher for the second session ($t=4.772$).

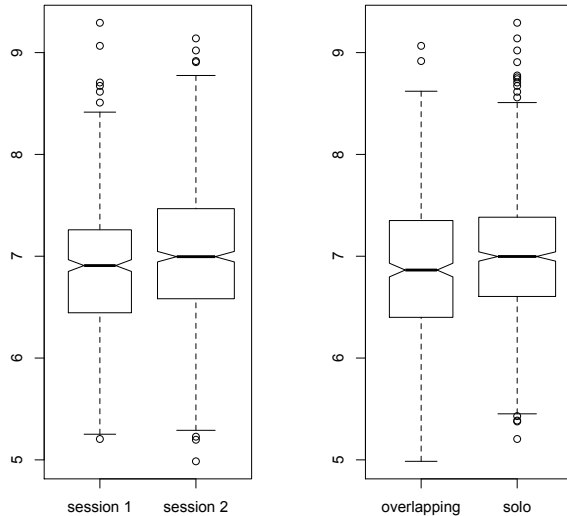


Fig. 4. Showing plots of parameter RCG (representing a measure of tenseness in the voice) for four participants, factored by both Sessions (left) and Overlapping speech (right).

When overlap is included as a factor in the svm training, prediction error decreases by 10%, yielding an accuracy of 47.2% for the 12-class categorisation.

5. DISCUSSION

We can explain these findings in terms of both vocal effort and social relationships. The first session was tense and technologically oriented whereas the second was relaxed and socially interactive. Speakers voices may be more breathy when they are relaxed or when they are engaged in friendly social discourse. Similarly, when interrupting or speaking over another, they use more vocal effort.

Interestingly, there was a difference in the direction of change for one person across sessions. We infer from the above that this person who entered midway during the first session may have enjoyed watching the setup, but perhaps became a little tense when the ‘official’ recording started.

6. CONCLUSION

This paper has presented some of our findings from analyses of conversational speech corpora of Japanese and English. It has shown that contrary to the “no gap no overlap” theory of speech interaction [2] there is considerable overlap found in our data. We explain this as due to the need for conversation participants to indicate their engagement in the discourse and to provide constant feedback to the speaker. The same patterns of speech activity are found for English as for Japanese.

Whereas many of the utterances we examined in the second part of the paper are textually very similar, we have

shown that speakers make significant use of both prosody and voice quality to demonstrate their engagement and discourse intentions. Future work will include this voice quality information in advanced prosodic feature extractors for social interaction analysis.

7. ACKNOWLEDGEMENTS

This work was done at TCD thanks to Science Foundation Ireland (SFI Stokes Professorship Award 07/SK/I1218). The second author is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) The work of Helena Moniz is supported by FCT grant SFRH/BD/44671/2008.

8. REFERENCES

- [1] Kendon, Adam, (1990) *Conducting Interaction: Patterns of Behaviour in Focused Encounters*. Cambridge: Cambridge University Press.
- [2] Sacks, Harvey, Schegloff, Emanuel A., & Jefferson, Gail (1974). A simple systematic for the organization of turn-taking in conversation. *Language*, 50, 696-735.
- [3] Campbell, Nick and Scherer, Stefan, (2010) “Comparing Measures of Synchrony and Alignment in Dialogue Speech Timing with Respect to Turn-Taking Activity”. in *Proc Interspeech 2010*, pp.2546-2549.
- [4] The Japan Science & Technology Agency’s (JST) Expressive Speech Processing (ESP) Corpus
- [5] Allwood, J. (1993), “On dialogue cohesion”. *Gothenburg papers in Theoretical Linguistics* 65, Göteborgs Universitet, Department of Linguistics.
- [6] Bunt, H. (2000), “Dialogue pragmatics and context specification”. Harry Bunt & William Black (eds.), *Abduction, Belief and Context in Dialogue*, vol. 2, Amsterdam:Benjamins, pp. 139-166.
- [7] Bunt, H. (2010), “Multifunctionality in dialogue”. *Computers, Speech and Language*, special issue on dialogue modeling, Yorick Wilks, editor.
- [8] Campbell, Nick and Erickson, Donna, (2004) “What do people hear? - a study of the perception of non-verbal affective information in conversational speech”, *Journal of the Phonetic Society of Japan*, 8, (1), pp.9 - 28
- [9] The FreeTalk Multimodal Conversation Corpus — <http://sspnet.eu/2010/02/freetalk/> (also available in interactive form from <http://www.speech-data.jp>)
- [10] Luggner, M., and Yang, B., (2006) “Classification of different speaking groups by means of voice quality parameters”, *SprachKommunikation*.