Exploring the implications for feedback of a neurocognitive theory of overlapped speech

D. Neiberg, J. Gustafson

Centre for Speech Technology (CTT), TMH, CSC, KTH, Stockholm, Sweden

[neiberg,jocke]@speech.kth.se

Abstract

Neurocognitive evidence suggests that the cognitive load caused by decoding interlocutors speech while one self is talking is dependent on two factors: the type of incoming speech, i.e. non-lexical feedback, lexical feedback or non-feedback; and the duration of the speech segment. This predicts that the fraction of overlap should be high for non-lexical feedback, medium for lexical feedback and low for non-feedback, and that short segments has a higher fraction of overlapped speech than long segments. By normalizing for duration, it is indeed shown that the fraction of overlap is 32% for non-lexical feedback, 27% for lexical feedback and 12% for non-feedback. Investigating non-feedback tokens for the durational factor gives that the fraction of overlap can be modeled by linear regression and logarithmic transform of duration giving a $R^2 = 0.57$ ($p < 0.01$ for F-test) and a slope $b(2) = -0.04$ ($p < 0.01$ for T-test). However, it is not enough to take duration into account when modeling overlap in feedback tokens.

Index Terms: neurocognitive theory, feedback, lexical feedback, non-lexical feedback

1. Introduction

The perhaps simplest observation of human conversation is that overwhelmingly one speaks at a time. This observation can be considered as the first principle in turn-taking theory [1], but other observations include the presence of short feedback in overlapped speech [2] and having an overlap up to one second in speaker shifts is quite common [3]. This indicates that there are exceptions to the one-at-time principle which might be related to the neurocognitive decoding of feedback and short phrases. There are relatively few neurocognitive theories of turn-taking. The literature include a hybrid computational model of Ymir Turn-Taking Model (YTTM) and the Augmented Competitive Queuing (ACQ) [4], an oscillator model for timing phenomena [5] and the role of mirror neurons and the motor cortex [6]. This paper aims to formulate a neurocognitive theory for the occurrence of feedback and short phrases in overlap and verify the predictions.

The most basic observation of conversation, the one at a time principle, can be hypothesized to be caused by overlapped phonological or semantic processing systems for speech recognition and production. Two candidates for this overlap are Wernicke's area (Brodmann Area 22), located in posterior superior temporal gyrus (STG), and Broca's area (Brodmann Area 44), located in posterior inferior frontal gyrus. Both areas has shown highly correlated activation in (fMRI) functional magnetic resonance imaging experiments during speech perception and production [7]. They points towards evidence which indicate that the overlapped regions are also involved in tasks which puts load onto the phonological loop in working memory. The model of working memory consists of a phonological loop, an visuo-spatial sketchpad and a central executive [8]. Cognitive load is the term for the general effort or the interactions that must be processed simultaneously in working memory [9]. The contribution of Broca's area for sentence processing has been further examined by [10] using fMRI for a baseline condition with no secondary task, during a concurrent speech articulation task and during a concurrent finger-tapping task. It was found that concurrent articulation significantly reduces the ability to comprehend object-relative clause sentences compared to subject relative and compared to comprehension during the finger-tapping task. In the baseline condition, they found greater activation in Broca's area for sentences with object relatives than subject relatives but this effect was only found in the in a part of Broca's area during concurrent speech articulation. They interpreted these findings as: *"Under high processing load conditions, such as sentences with object-extracted relative clauses, verbal working memory can be recruited to assist comprehension"*.

By assuming a reasonable degree of cooperation in conversation and the existence of an inverse relationship between cognitive load and the fraction of overlapped speech, it seems reasonable to conclude that the partial overlap of the production and recognition systems contributes to the cognitive load of speaking and listen at the same time to the extent that it can be considered as the first principle of turn-taking.

Working memory can handle two or three of novel interacting elements, while the capacity is higher for non-novel information [11]. What can be remembered is also inversely related to word length and the total span could be predicted on the basis of the number of words which the subject can read in approximately 2 seconds [12]. If this duration effect still exists while one is decoding interlocutors speech, then very short utterances may still be acceptable to comprehend during production since they are less likely to cause excessive cognitive load. This exception to the one at a time principle can explain the common presence of short overlaps found in speaker shifts [3].

In everyday conversation, interlocutors usually give brief vocal feedback like "uhu", "okey" and "Yeah, that's right" while the other is talking. Yngve [2] noticed that feedback is common in overlapped speech. He put forward the idea of a main half-duplex channel in conversation, which meant that overlapped speech including feedback have to be transmitted in a back-channel. The frequent occurrence of feedback in overlapped speech has caught interest among researchers in turn-taking [13] and feedback is sometimes defined as these utterances which do not take the floor or are not full turns [14]. Empirically feedback has indeed shown an over-representation in overlapped speech for English [15] and Swedish [16]. This

cross-speaker context, that feedback often occur in overlapped speech, is a distinct characteristic of feedback.

Very short utterances may still be acceptable to comprehend during production since they are less likely to cause excessive cognitive load. Since feedback segments have short duration, it may explain the over-representation of feedback in overlap. Then short phrases in general are also acceptable and should show the same over-representation in overlap as feedback. Empirically, very short utterances (< 1 sec.) [17] as well as manually labeled feedback [18] have both shown over-representation in overlap.

Another contribution to overlap can be derived from the main function of feedback. The primary function of feedback is to convey affect and attitudes via prosody [19]. Vocal sounds are processed along the auditory "what" processing stream reaching from the auditory cortex to the lateral STG and to the superior temporal sulcus (STS) where an emotional "gestalt" is formed. For each consecutive processing step there is an increasing lateralization to the right hemisphere which processes pitch and segments on a wider temporal scale and an increasing lateralization to the left hemisphere for a finer temporal processing suitable for decoding phonemic structure [20]. Thus, affective-prosodic decoding partially involves different areas in the brain than those used for linguistic decoding. Indeed, decoding of verbal interjections modulated by affective prosody has shown involvement of areas which are primary used for affective decoding [21]. The parallel mechanisms for affective and linguistic decoding may explain why it is not problematic to decode feedback while one is talking. What is likely to pass through this back-channel, to use Yngve's terminology, is non-lexical feedback, i.e. feedback which has low linguistic content with slowly varying spectral flux and high affective content like "uhu". Thus, the proportion of overlapped speech should be higher for non-lexical feedback than for lexical feedback and lowest for syntactical structures of words which excludes laughter and other extra-linguistic sounds. This effect should be additive to the effect of duration. We put the expected proportion of overlap for lexical feedback above non-feedback due to the expected affective loading, the lack of syntactical structure (in case of single words) and the idiomatic/non-novel structure (in case of short phrases).

The predictions are here verified as the fraction of overlapped speech, computed with and without normalizing for duration as well as examine duration alone (Section 3). For this we utilize the DEAL corpus as described in Section 2 and finally the findings are concluded in Section 4.

2. The DEAL corpus

This study uses data from the DEAL corpus [22]. It consists of eight role-playing dialogs recorded as an informal, human-human, face-to-face conversation. The data collection was made with 6 subjects (4 male and 2 female), 2 posing as shop keepers and 4 as potential buyers. Each customer interacted with the same shop-keeper twice, in two different scenarios. The customers were given a task: to buy items at a flea market at the best possible price.

All vocal activity in the DEAL corpus was segmented into Inter Pausal Units which is defined as connected segments of vocalizations bridged by a minimally perceivable pause set to 200 ms. All dialogs were transcribed orthographically including non-lexical entities such as laughter and audible breathing. Filled pauses, repetitions, corrections, restarts and cue phrases were labeled manually. The corpus is rich in feedback tokens. The feedbacks were generally single words or non-lexical tokens and appeared in similar dialog contexts (i.e. as responses to assertions). We divided the feedback tokens into non-lexical, i.e. postulated as the ones which only consists of sonorants [23], and lexical feedback which mainly consists of "okej" (okey), "precis" (exactly), other affirmative words and short phrases. The token counts for the two classes are shown in Table 1. It can be seen that the non-lexical category contains tokens with a more slowly varying spectral flux suitable for processing in the right hemisphere while the lexical category contains phonemes with more spiky spectral flux such as plosives or thrills.

Table 1: *Token counts for non-lexical and lexical feedback in the DEAL corpus.*

non-lexical		
token	count	percentage (%)
ja	448	39.5
m	163	14.4
a	91	8.0
nej	50	4.4
nä	47	4.1
hm	24	2.1
mm	24	2.1
mhm	18	1.6
jaha	15	1.3
jo	14	1.2
Other	62	5.4
lexical		
det	26	2.3
precis	25	2.2
okej	25	2.2
just	17	1.5
jag	7	0.6
förstår	6	0.5
hur	6	0.5
eller	6	0.5
är	5	0.4
då	4	0.4
Other	50	4.4

3. Investigations

The goal is to compute the proportions of overlapped speech for different types of speech segments (Inter Pausal Units). To do this, the provided annotation is quantized into frames. This choice not only simplifies computation but it also makes investigation no. 2 possible, see Section 3.2, where segments of equal length are compared. This choice also makes the results directly applicable to stochastic models for detection, segmentation and turn-taking which operate on frame level [16, 24]. The frame size is chosen to be 50 ms.

3.1. Investigation 1

The first investigation aims to get a rough picture of the interaction between the factors: duration, non-feedback and feedback where the latter is expanded into lexical and non-lexical feedback. To do this, we compute the fraction of overlap and average duration for Non-Lexical Feedback (NLF); Lexical Feedback (LF); Short Non-Feedback segments (SNF), i.e segments with a duration shorter than 1 second (as in [17]); Non-Feedback segments (NF) and Long Non-Feedback segments, i.e segments with a duration longer than 1 second (LNF). The

fraction of overlap metric excludes silence and extra-linguistic sounds, e.g. laughter. The result is shown in Table 2.

3.1.1. Discussion

To interpret the results, the interaction between duration and fraction of overlap has to be considered. As predicted, non-lexical feedback has the highest proportion of overlap but has also the shortest duration. Lexical feedback and short non-feedback segments has almost the same average duration while lexical feedback has much higher fraction of overlap than short non-feedback segments. Non-feedback segments and long non-feedback segments has almost the same fraction of overlap while long non-feedback segments has much longer average duration.

Examining the proportion metric makes it clear that feedback is overrepresented in overlap. Specifically, non-lexical feedback is more over-presented in overlap compared to lexical feedback. However, short non-feedback segments are also more common in overlap than longer non-feedback segments, but less so than feedback. If feedback is excluded from the short segments there is still an over-representation in overlap and it is higher for short segments than for longer segments.

As predicted, these results suggests an interaction between the continuous duration factor and the categorical feedback or non-feedback factor. However, it is not clear whether non-lexical or lexical feedback are separate factors. The effect of duration seems to diminish after 1.0 second. This opt for two follow up investigations, one in which the duration factor is normalized and one in which the duration factor is examined independently.

Table 2: *Fraction of overlap, average duration and number of frames for NLF: Non-Lexical Feedback; LF: Lexical Feedback; SNF: Short Non-Feedback (IPU $\leq$ 1 s.); NF: Non-Feedback segments; LNF: long Non-Feedback (IPU $>$ 1 s.).*

	NLF	LF	SNF	NF	LNF
Ovl.(%)	0.33	0.27	0.18	0.12	0.11
Avg. Dur.	0.36	0.51	0.55	1.16	1.85
N	8002	1343	73633	17577	56056

3.2. Investigation 2

In the second investigation the fraction of overlap is compared for segments of non-lexical feedback, lexical feedback and non-feedback with identical duration. Given that the maximum duration of all segments is N, the segments are collected by creating sets $d = 1 \ldots N$ each containing all segments with a duration equal to d. Then for each set d, a new set $\hat{d}$ is created by collecting equal number of segments for each of the three classes. The segments are picked in chronological order of appearance which means that the most segments are picked from the first dialogs. Every set $\hat{d}$ does now contain an equal number of segments of equal length. This procedure collected 127 segments with 1328 frames from each class. The fraction of overlap is shown in Table 3.

Table 3: *The fraction of overlap computed for equal number of segments with equal duration for each of Non-Lexical Feedback (NLF), Lexical Feedback (LF) and Non-Feedback (NF). There are 127 segments with 1328 frames for each class.*

	NLF	LF	NF
Ovl. (%)	0.32	0.27	0.12

3.2.1. Discussion

The result show that the fraction of overlap is highest for non-lexical feedback, somewhat lower for lexical feedback and lowest for non-feedback. This ordering was predicted by neurocognitive theory. However, while the difference between feedback and non-feedback was rather high, the difference between non-lexical and lexical feedback was smaller.

3.3. Investigation 3

In the third investigation the duration factor is examined separately for non-feedback tokens and only feedback tokens. The total fraction of overlap for each segment is parameterized as a function of the total segment duration in bins of 100 ms. This gives multiple points per tick on the x-axis in the histogram. The fraction of overlap is hypothesized to be smaller for longer segments which made us to chose weighted linear regression as a model. The weights are chosen as the inverse variance per bin computed via the normal distribution approximation of the binomial distribution. This gives a small weight for x-ticks which has a small number of observations or a large spread among the observations, and high weight when the latter two criteria are not satisfied. Two types of scales for the x-axis are tested: linear and logarithmic. The goodness-of-fit metrics and slopes for the models are shown in Table 4.

Table 4: *Goodness-of-fit metrics and slopes for modeling the fraction of overlap via linear regression with duration as an explanatory variable.*

Linear x-scale				
Type	R^2	p(F-test)	p(T-test)	slope
NF	0.39	0.001	0.000	-0.02
F	0.64	0.005	0.987	0.15
Logarithmic x-scale				
NF	0.57	0.000	0.000	-0.04
F	0.72	0.002	0.995	0.09

3.3.1. Discussion

The results for the regression analysis show that a logarithmic transformation of segment duration gives a better fit in in terms of R^2. The F-test for the models was significant for all conditions but the T-test showed that the slope was not significantly different from zero for feedback, but significant for non-feedback. The former finding may be due to lack of enough data-points for feedback since most feedback tokens are shorter than one second or due to that duration does not matter for feedback. The regression line for non-feedback is shown in Figure1. It can be seen that the fraction of overlap decreases with increasing token duration.

4. Conclusion

Based on neurocognitive evidence it is argued that the cognitive load caused by decoding interlocutors speech while one self is talking is dependent on two factors: type of speech, i.e. non-lexical feedback, lexical feedback and other speech; and the duration of the speech segment (Inter Pausal Units). By assuming an inverse relationship of cognitive load and the fraction of overlapped speech, it is predicted that the fraction of overlap is high for non-lexical feedback, medium for lexical feedback and low for non-feedback (excluding extra-linguistic sounds like laughter). In addition, constraints on working mem-

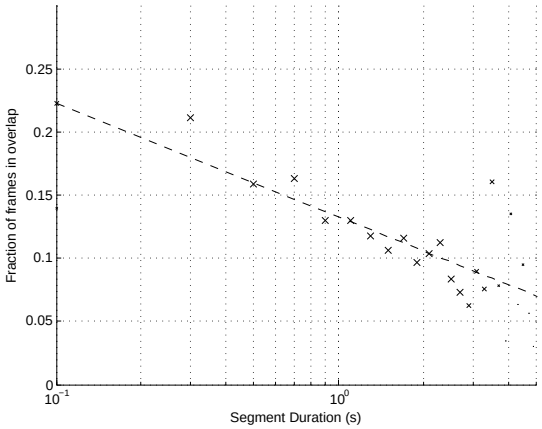

Figure 1: *Explaining the fraction of overlap for non-feedback segments by linear regression as a function of logarithmic transformed segment duration. The size of the markers are proportional on the counts for each point.*

ory predicted that short segments has a higher fraction of overlapped speech than long segments.

By separating the continuous duration factor and the categorical non-lexical/lexical feedback or non-feedback factor, it is shown that the fraction of overlap is 32% for non-lexical feedback, 27% for lexical feedback and 12% for non-feedback. The fraction of overlap for non-feedback can be modeled quite accurately by linear regression and logarithmic transform of duration giving a $R^2 = 0.57$ ($p < 0.01$ for F-test) and a slope $b(2) = -0.04$ ($p < 0.01$ for T-test). However, the fraction of overlap for feedback tokens could not be explained by duration since the T-test failed.

Since the computations are made on frame-level, here chosen to be 50 ms, the results are directly applicable to stochastic models for detection, segmentation and turn-taking which operate on frame level [16, 24]. For example, the computed proportions of overlap for the categorical factor corresponds to maximum likelihood estimates of finding or predicting different types of segments in overlap, and the continuous durational factor corresponds to modeling the probability of overlap with exponential decaying functions.

For turn-taking theory in general, the results gives a neurocognitive motivation for excluding feedback, especially non-lexical feedback, and short segments from the one-at-a-time principle. One possible extension of this work would be to investigate the impact of novel versus non-novel stimuli, since cognitive load is proportional to the novelty of verbal stimuli.

5. Acknowledgements

The authors would like to thank Petri Laukka for discussions. Funding was provided by the Swedish Research Council (VR) projects 2009-4291 and 2009-4599.

6. References

[1] H. Sacks, E. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, pp. 696–735, 1974.

[2] V. H. Yngve, "On getting a word in edgewise," *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pp. 567–577, 1970.

[3] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010.

[4] J. Bonaiuto and K. Thórisson, "Towards a neurocognitive model of turn taking in multimodal dialog," in *Embodied communication in humans and machines*, M. L. I. Wachsmuth and G. Knoblich, Eds. New York: Oxford University Press., 2008, pp. 451–483.

[5] M. Wilson and T. P. Wilson, "An oscillator model of the timing of turn-taking," *Psychonomic bulletin review*, vol. 12, no. 6, pp. 957–968, 2005.

[6] S. K. Scott, C. Mcgettigan, and F. Eisner, "A little more conversation, a little less action - candidate roles for the motor cortex in speech perception," *Nature Reviews Neuroscience*, vol. 10, no. 4, pp. 295–302, March 2009.

[7] B. R. Buchsbaum, G. Hickok, and C. Humphries, "Role of left posterior superior temporal gyrus in phonological processing for speech perception and production," *Cognitive Science*, vol. 25, no. 5, pp. 663–678, 2001.

[8] A. Baddeley and G. Hitch, "Working memory," in *The Psychology of Learning and Motivation*, G. Bower, Ed. Academic Press, 1974, pp. 48–79.

[9] John and Sweller, "Cognitive load during problem solving: Effects on learning," *Cognitive Science"*, vol. 12, no. 2, pp. 257 – 285, 1988.

[10] C. Rogalsky, W. Matchin, and G. Hickok, "Broca's area, sentence comprehension, and working memory: An fMRI study," *Frontiers in human neuroscience*, vol. 2, no. October, p. 13, 2008.

[11] F. Paas, A. Renkl, and J. Sweller, "Cognitive load theory and instructional design: Recent developments," *Educational Psychologist*, vol. 38, no. 1, pp. 1–4, 2003.

[12] A. D. Baddeley, N. Thomson, and M. Buchanan, "Word length and the structure of short-term memory," *Journal of Verbal Learning and Verbal Behavior*, vol. 14, no. 6, pp. 575 – 589, 1975.

[13] E. Schegloff, "Overlapping talk and the organization of turn-taking for conversation," *Language in Society*, vol. 29, pp. 1–63, 2000.

[14] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in English and Japanese," *Journal of Pragmatics*, vol. 32, no. 8, pp. 1177–1207, 2000.

[15] Özgür Çetin and E. Shriberg, "Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: Insights for automatic speech recognition," in *Proc. ICSLP*, Pittsburgh, 2006, pp. 293–296.

[16] D. Neiberg and J. Gustafson, "A dual channel coupled decoder for fillers and feedback," in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association*, Florence, Italy, sep 2011.

[17] M. Heldner, J. Edlund, A. Hjalmarsson, and K. Laskowski, "Very short utterances and timing in turn-taking," in *Proc. of Interspeech 2011*, Florence, Italy., 2011.

[18] D. Reidsma, I. de Kok, D. Neiberg, S. Pammi, B. van Straalen, K. Truong, and H. van Welbergen, "Continuous interaction with a virtual human," *Journal on Multimodal User Interfaces*, vol. 4, no. 2, pp. 97–118, jul 2011.

[19] J. Allwood, J. Nivre, and E. Ahlsen, "On the semantics and pragmatics of linguistic feedback," *Journal of Semantics*, vol. 1, no. 9, pp. 1–26, 1992.

[20] A. Schirmer and S. A. Kotz, "Beyond the right hemisphere: brain mechanisms mediating vocal emotional processing," *Trends Cogn Sci*, vol. 10, no. 1, pp. 24–30, 2006.

[21] S. Dietrich, I. Hertrich, K. Alter, A. Ischebeck, and H. Ackermann, "Understanding the emotional expression of verbal interjections: a functional mri study," *Neuroreport*, vol. 19, no. 18, pp. 1751–5, 2008.

[22] A. Hjalmarsson, "Speaking without knowing what to say... or when to end," in *Proceedings of SIGDial 2008*, Columbus, Ohio, USA, jun 2008.

[23] N. Ward, "The challenge of non-lexical speech sounds," in *In International Conference on Spoken Language Processing*, 2000.

[24] K. Laskowski, J. Edlund, and M. Heldner, "A single-port non-parametric model of turn-taking in multi-party conversation," in *Proc. of ICASSP 2011*, Prague, Czech Republic, may 2011, pp. 5600–5603.