Is it really worth it? Cost-based selection of system responses to speech-in-overlap

Jens Edlund and Anna Hjalmarsson*

KTH Speech, Music and Hearing, Stockholm, Sweden
{edlund, annah}@speech.kth.se

Abstract. For purposes of discussion and feedback, we present a preliminary version of a simple yet powerful cost-based framework for spoken dialogue systems to continuously and incrementally decide whether to speak or not. The framework weighs the cost of producing speech in overlap against the cost of not speaking when something needs saying. Main features include a small number of parameters controlling characteristics that are readily understood, allowing manual tweaking as well as interpretation of trained parameter settings; observation-based estimates of expected overlap which can be adapted dynamically; and a simple and general method for context dependency. No evaluation has yet been undertaken, but the effects of the parameters; the observation-based cost of expected overlap trained on Switchboard data; and the context dependency using inter-speaker intensity differences from the same corpus are demonstrated with generated input data in the context of user barge-ins.

Keywords: Index Terms: barge-in, overlap, generation, dialogue systems

1 Introduction

Ever since the first spoken dialogue systems met with real users in the wild, overlapping speech has been a reality for their designers to cope with [1]. Early on, overlap was viewed purely as an error and therefore discarded by disabling the microphone during system speech. Subsequently, user-initiated overlaps were given the technical label *barge-in*. and sometimes dealt with by the instant cessation of speech [2] - a hard-wired decision with no consideration of the consequences of abandoning a near-complete speech segment or an urgent message. In most dialogue systems that allow barge-ins, overlapping speech from the user is always considered as an attempt to take the turn and the user is always given precedence over the system. Yet, in the Switchboard corpus [3], our analysis found that only 58 % of the overlaps result in a speaker change. This means that a systems whose strategy it is to either *always* or *never* cut itself short will make the wrong decision about half the time, at least if our gold standard is human conversation and our goal human-like interaction.

The current work introduces a preliminary version of a simple yet powerful framework to deal with turn-taking in a sophisticated and flexible manner, with a current focus on managing user-initiated speech-in-overlap. We pit the cost of com-

*Authors in aphabetical order

pleting a planned segment in overlap against the cost of not completing the segment, using a small set of parameters to influence the system's behaviour. Costs and parameters alike are designed to be intuitive and readily understood by humans. This is motivated by a desire to build spoken dialogue systems not solely with the intention of creating good human-machine interfaces, but also in the hopes of learning something about human conversational behaviour [4]. The framework operates on a per-frame basis and exploits two simplifying assumptions: that the number of frames it will take to complete a planned speech segment is known and that the decision is limited to whether to cut the segment short *immediately* or to produce it *in its entirety*. Still, the decision can be re-evaluated repeatedly and incrementally at consecutive frames, allowing for the system to stop at any given point in its speech.

2 Background

Overlap is common, ranging from 6-13 % of the total speaking time in conversational meeting data [5]. While frequently occurring, overlaps are brief in time: we find that 78 % of the overlaps in the Switchboard corpus are 200-500 ms long. Yet, when overlapped, people do not abandon their turn-in-progress immediately. For example, overlapping speech is claimed to be non-disruptive if starting just prior to the current speaker's completion point [6], and some overlaps are continuers, which by definition make no claim to the turn [7]. Hence, a decision to immediately abandon speech-in-progress will often be premature – speech overlap resolution is more versatile. Interlocutors appear to resolve overlaps through negotiation, incrementally [8]. At each point in time, a speaker may drop out, continue as before, or use different strategies to continue speaking unaccompanied [9]. [10] argues that overlaps are resolved based on their prosodic realization and that speakers who "compete for the turn" use increased pitch and loudness. This is supported by a correlation between the number of "successful" interruptions and inter-speaker amplitude difference [11]: the number of overlaps resulting in the incoming speaker taking over the turn appears to be correlated with the difference between the speakers' amplitude during overlap.

Inspired by human-human dialogue, Ström and Seneff [12] presents a system which increases its voice intensity when barge-ins occur at dialogue states where interruptions are undesirable, signalling that barge-ins are disallowed at this stage. When a barge-in occurs at a less critical point in the dialogue, they propose that the system reduce its intensity, but continue to speak, which allows the system to verify that the detected barge-in was indeed speech from the user before cutting itself short.

Decision-theoretic approaches to turn-taking have been presented by [13] and [14]. [13] uses a cost-based finite-state model to minimize gaps and overlaps. The goal is to detect ends of turns early on, increasing the system responsiveness. [14] attempts to minimize costs based on expected outcomes in order to guide turn-taking decisions in multiparty dialogue. Whereas our framework is designed to operate at all times, making a simple binary decision – to speak or not to speak – at each frame of the conversation, the decisions in [14] are limited to end-of-speech, and [13] manipulates the detection of end-of-speech. Another model that takes the urgency of an intended speech segment into consideration is the importance driven turn-bidding presented in [15]. In this general turn-taking framework, the turn is allocated to either the user or the system depending who shows more eager to speak by taking the prominence of different turn-taking cues into account. Reinforcement learning is used to optimize the efficiency of the system's turn-bidding strategies given different user profiles.

3 Framework description

The framework is a general cost-based turn-taking framework for spoken dialogue systems, producing a simple binary decision for each frame: whether to speak or not to speak. Initially, we want to apply the decisions to a subset of possible situations in a dialogue, and narrow the decision down to one of completing or abandoning an ongoing speech segment when overlapped by a user. The decision is based on weighting the cost of incompletion (CoIC) against the cost of expected overlap (CoEO). These costs can be made dependent on for example situational needs or economy principles.

3.1 Cost of incompletion

CoIC represents the cost of not finishing the speech segment that is currently being spoken. CoIC would typically be affected by the urgency of the (remainder of the) segment, but also of the effort already spent on producing the message. If the bulk of the message has already been presented, cutting the turn-in-progress short and risk having to start over in order to get the message through is likely less desirable.

$$CoIC = W_1 * I_r + W_2 * \left(\frac{I_s}{I_r}\right)^{W_3}$$

The formula shows the estimation of CoIC. Here, we represent the information already presented (I_s) as the number of frames already spent and the information yet to be spoken (I_r) as the number of frames remaining. CoIC is the sum of a linear function of the time remaining and an exponential function of the proportion of time spent speaking and the time remaining to complete the segment. The first term, controlled by the weight W_1, approximates the cost of the information content that is lost in case of abandonment and the second makes segments that are relatively close to completion expensive to abandon, with W_2 and W_3 weighting the starting value and slope.

3.2 Cost of expected overlap

CoEO represents the cost of completing a planned speech segment with F_r frames remaining by combining a linear cost that accounts for the effort to produce speech and a cost for producing speech in overlap that grows exponentially with time. The linear part of the cost is motivated by an economy principle where each frame spent speaking comes at a price. The cost is weighted by W_4, as it may increase, for example with ambient noise. The exponential part, modified by the probabilities P_n of overlap of lengths from 0 to F_r, is a cost associated with frames spent speaking in overlap. Its starting value and slope are weighted by W_5 and W_6. This cost of expected overlap is motivated by the assumption that speaking and listening at the same time increase cognitive load, evidenced by the relatively small proportion of overlapping speech found in most dialogues. The cost of overlap is computed as follows:

$$CoEO = W_4 * F_r + \sum_{n=0}^{F_r} P_n(1 + W_5)^{W_6 * n}$$

In our simplified examples, F_r is the same as I_r, since we make the simplifying assumption that each frame holds the same amount of information. The probabilities are learned from a dialogue corpus, and can be conditioned on context. The features used in the initial implementation were chosen because they play a central role in overlaps and they can be expected to be available to the system in real-time.

4 Data

The contextual features used to calculate expected overlap probabilities were based on overlap distributions in the SWITCHBOARD-1 corpus [3]. The corpus consists of 2435 dyadic conversations with 543 different speakers.

We used a speaker-disjoint training set of SWITCHBOARD with 762 conversations (kindly provided by [16]). The test sets are kept aside for future development and evaluation. Instances of overlapping speech were extracted from the transcriptions using a 100 ms frame size. Overlaps shorter than 2 frames were excluded from the analysis since it overlaps of less are often not perceivable by listeners [17]. Furthermore, overlaps including laughter in either channel were excluded (N=7503), since laughter appear to invite overlap and therefore does not follow the norm one-speaker-at-a-time [8]. Finally, we also excluded overlaps where both speakers started up simultaneously just after a preceding overlap (N=2325), since it was difficult to determine who was the original speaker in those cases.

COIC with 3 parameter combinations were used: COIC_1, COIC_2 and COIC_3 with slope, exponential base, and growth rate of 15, 1,5 %; 5, 1, 70 %; and 0, 2, 50 %.

CoEO with 3 parameter combinations were used: CoEO_1, CoEO_2 and CoEO_3 with slope, exponential base, and growth rate of 0, 1, 90 %; 5, 1, 50 %; and 0, 2, 50 %.

The probabilities of overlap given a certain number of intended frames of speech were trained on the corpus data, with observations of the overlapped interlocutor speaking for more than 10 frames after the overlap onset (regardless of overlap duration) grouped into a single category for all intentions to speak for 10 or more frames, 10+, making the maximum expected overlap for any speech segment is 10 frames.

In addition, probabilities conditioned on three inter-speaker intensity differences during overlap, HIGH, MID and LOW, were created. An inter-speaker relation was used to model intensity variation in order to achieve robustness against the continuous and on-going nature of dialogue and to be able to deal with local increases in intensity due to external events such as variation in ambient noise (i.e. Lombard-effects). The difference in inter-speaker normalized intensity (dB) between the overlapping speaker and the overlapped speaker was compared. The intensity measures used were extracted from voiced intervals only. The average intensity in decibel during the first 200 ms of overlap was extracted from both speakers provided that the segment had at least one frame of voiced speech. The training data was then discretized into three equally sized (N = 2326) intensity groups, LOW, MID and HIGH. The labels were termed from the perspective of the overlapped speaker: HIGH intensity overlaps are the ones were the incoming speaker's speech is at least 5.7 dB (decibel) higher than the overlapped speaker's speech during the first 200 ms of overlap; LOW intensity overlaps have the incoming speaker at a minimum of 1.6 dB lower the overlapped speaker's speech, and the remaining overlaps are classified as MID. Three models with the same parameter settings as CoEO were trained on the data from each of these classes: CoEO$_H$, CoEO$_M$, and CoEO$_L$, respectively.

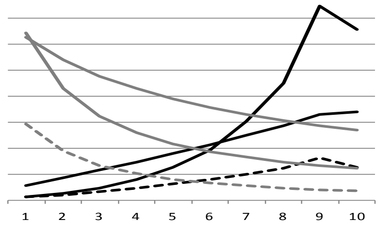

Fig. 1. The development of CoIC (Y axis; falling lines) and CoEO (Y axis, rising lines) over frames to speak (X axis) for different parameter settings: CoIC$_1$, CoIC$_2$, CoIC$_3$ and CoEO$_1$, CoEO$_2$, and CoEO$_3$, respectively, from the top down.

To test that parameter settings and corpus based training have the desired effect, artificial test data containing exactly one instance of each possible input/condition combination was created. The data contained one instance of each combination of number of frames to speak (between 1 and 10), number of frames spoken (between 2 and 10; 2 because we restrict overlaps to instances where exactly two frames of overlap have been observed here), and intensity (HIGH, MID and LOW). This results in a test data of 10*9*3=270 inputs, or 27 for each number of frames to speak. The data is used to illustrate the effects of parameter settings and training.

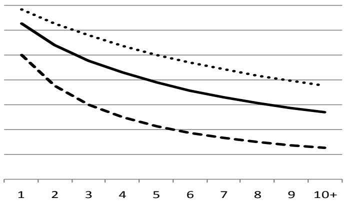

Fig. 2. The development of CoIC1 (Y axis) for different frames to speech (X axis). The solid line includes all data from 2 to 9 frames already spoken. The top line shows CoIC1 for 9 frames already spoken and the bottom line for 2 frames already spoken.

5 Results

5.1 Corpus statistics

The 762 conversations training set contains 20840 overlaps. The mean duration of the overlaps is 450 ms and the median is 400 ms.

5.2 Effects of parameter settings

Fig. 1 shows the effect of linear slope, exponential base and exponential slope of CoIC and CoEO. The CoIC parameters allow designers to increase the overall cost of not completing a speech segment as well as the cost of not completing a speech segment of which little remains, and the CoEO parameters allow designers to control both the overall cost per frame to speak as well as the cost of extended expected overlaps, as intended. CoEO in the figure is based on overall expectations from observations in the corpus. The drop at 10 or more frames to speak is an effect of the fact that in those cases that the overlapped speaker continued for 10 or more frames after the overlap onset, a large proportion were spoken with a very brief overlap, corresponding to backchannels spoken in overlap.

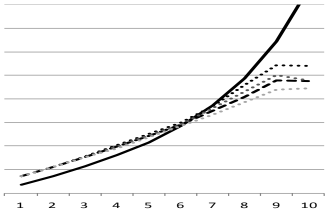

Fig. 3. The development of CoEO (Y axis) for different frames to speak (X axis). The solid line is the cost without training, and the three dotted lines show costs trained on different intensity conditions, $CoEO_H$, $CoEO_M$, and $CoEO_L$, from the top down

5.3 Effects of context dependency

Fig. 2Error! Reference source not found. shows $CoIC_1$ for different numbers of frames already spoken. The function yields higher costs when a lot of effort has already been spent, and lower just at the beginning of speech segments, as intended.

Fig. 3 shows CoEO for the three intensity conditions. The cost of extended speech is higher for the conditions where the incoming speaker is speaks louder, reflecting the increased probability of prolonged overlap, as intended.

5.4 Cost comparison

Fig. 4 shows context dependent version of both CoIC and CoEO, for one parameter setting. The results of a decision based on the difference between CoIC and CoEO depends each of the number of frames left to speak, the number of frames already spoken, and the context in terms of speaker intensity. The parameter settings allow us to vary the influence of these, and to balance exponential and linear cost development.

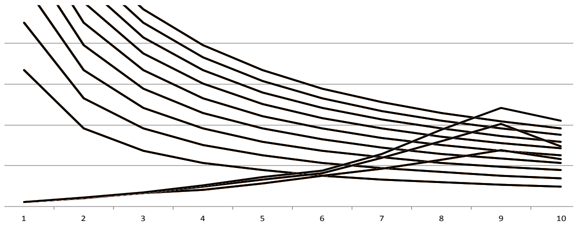

Fig. 4. Development of CoIC (Y axis, falling lines) and CoEO (Y axis, rising lines) over frames to speak (X axis). CoIC is split on frames already spoken, from 2 to 9 (top to bottom) and CoEO on intensity, $\mathrm{CoEO_H}$, $\mathrm{CoEO_M}$, and $\mathrm{CoEO_L}$, top to bottom.

6 Conclusions and future work

For purposes of discussion and feedback, we have presented a cost-based framework to make turn-taking decisions in dialogues. As a first demonstration, we have shown how its binary decision to complete or abandon a planned speech segment after a barge-in varies with different parameters and conditions. We argue that the decision to stop speaking instantly when a barge-in is detected is premature and that the length and the urgency of an incomplete speech segment should influence this decision. The framework allows for context dependency, and we have shown that conditioning its training on the intensity difference between the two speakers during overlap captures the real-world probability of longer overlaps, which is higher when the overlapper speaks with higher intensity: the cost of speaking in loud overlap becomes higher.

The framework poses a higher cost on cutting oneself short when a large proportion of the intended segment is completed. Here, we used a simulated measure for this: the number of frames spoken as compared to the number of frames yet to be said. We intend to exchange this measure for one that is based on the proportion of the information content in the speech segment that has been spoke, rather than the number of frames. A speech segment which is complete with the exception of a time consuming tag line ought to be easy to cut short. We also want to take laughter, praise and greetings that frequently are realized in “chorus” [8] into explicit consideration. Here, we have merely removed laughter from the corpus data.

We believe that our framework is complementary to similar frameworks presented in the past, and that it can be incorporated in a combination that facilitates the exploration of different system strategies with users in an experimental setting. Similarly, we aim to accompany the decision model with a speech synthesis system that can hold and release in addition to simply stopping.

Finally, the decision to speak or not to speak can be made on a frame-by-frame basis, re-evaluating the context as the dialogue progresses, making the system incremental.

7 Acknowledgements

The work was supported by Swedish Research Council (VR) project 2011-615 and the EU project Get Home Safe (EU FP7 STREP 288667).

8 References

1. Rudnicky, A., Hauptmann, A., & Lee, K. (1993). Survey of Current Speech Technology. Communications of the ACM, 37(3), 52-57.
2. Yankelovich, N., Levow, G. A., & Marx, M. (1995). Designing SpeechActs: Issues in speech user interfaces. In Proceedings of the Conference on Human Factors in Computing Systems (pp. 369-376). ACM Press/Addison-Wesley Publ. Co., New York, NY,.
3. Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 517–520). San Francisco.
4. Edlund, J., Gustafson, J., Heldner, M., & Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. Speech Communication, 50(8-9), 630-645.
5. Cetin, Ö., & Shriberg, E. (2006). Overlap in Meetings: ASR Effects and Analysis by Dialog Factors, Speakers, and Collection Site. In Proc. MLMI, 2006 (pp. 212-224). Springer LNCS.
6. Jefferson, G. (1986). Notes on 'latency' in overlap onset. Human Studies, 9(2/3), 153-183.
7. Schegloff, E. (1982). Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In Tannen, D. (Ed.), Analyzing Discourse: Text and Talk (pp. 71-93). Washington, D.C., USA: Georgetown University Press.
8. Schegloff, E. (2000). Overlapping talk and the organization of turn-taking for conversation. Language in Society, 29(1), 1-63.
9. Yang, F., & Heeman, P. (2010). Initiative conflicts in task-oriented dialogue. Computer Speech and Language, 24, 175-189.
10. French, P, & Local, J. (1983). Turn-competitive incoming. Journal of Pragmatics, 7, 17-38.
11. Meltzer, L., Hayes, D., & Morris, M. (1971). Interruption Outcomes and Vocal Amplitude: Explorations in Social Psychophysics. Journal of Personality and Social Psychology, 18(3), 392-402.
12. Ström, N., & Seneff, S. (2000). Intelligent barge-in in conversational systems. In Procedings of ICSLP-00.
13. Raux, A., & Eskenazi, M. (2009). A finite-state turn-taking model for spoken dialog systems. In Proc. of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '09) (pp. 629-637). Boulder, CO, USA.
14. Bohus, D., & Horvitz, E. (2011). Decisions about turns in multiparty conversation: from perception to action. In ICMI '11 Proceedings of the 13th international conference on multimodal interfaces (pp. 153-160).
15. Selfridge, E., & Heeman, P. (2010). Importance-Driven Turn-Bidding for Spoken Dialogue Systems. In Proceeding of ACL. Uppsala, Sweden.
16. Laskowski, K., & Shriberg, E. (2012). Corpus-Independent History Compression for Stochastic Turn-Taking Models. In Proceedings of ICASSP 2012. Kyoto, Japan.
17. Heldner, M. (2011). Detection thresholds for gaps, overlaps and no-gap-no-overlaps. Journal of the Acoustical Society of America, 130(1), 508-513.