Predicting Speaker Changes and Listener Responses With And Without Eye-contact

Daniel Neiberg, Joakim Gustafson

Department of Speech, Music and Hearing, Royal Instituite of Technology (KTH), Sweden
[neiberg|jocke]@speech.kth.se

Abstract

This paper compares turn-taking in terms of timing and prediction in human-human conversations under the conditions when participants has eye-contact versus when there is no eye-contact, as found in the HCRC Map Task corpus. By measuring between speaker intervals it was found that a larger proportion of speaker shifts occurred in overlap for the no eye-contact condition. For prediction we used prosodic and spectral features parametrized by time-varying length-invariant discrete cosine coefficients. With Gaussian Mixture Modeling and variations of classifier fusion schemes, we explored the task of predicting whether there is an upcoming speaker change (SC) or not (HOLD), at the end of an utterance (EOU) with a pause lag of 200 ms. The label SC was further split into LRs (listener responses, e.g. back-channels) and other TURN-SHIFTs. The prediction was found to be somewhat easier for the eye-contact condition, for which the average recall rates were 60.57%, 66.35% and 62.00% for TURN-SHIFTs, LR and SC respectively.

Index Terms: Turn-taking, Back-channels

1. Introduction

In human-human conversation interlocutors take turns based on generalized principles [1]. While such generalized principles are helpful in understanding the essence of the phenomenon, turn taking behavior vary as a function of many factors. Attempts of taking a turn has been found to be proportional to the number of prosodic and syntactical cues [2] given by the interlocutor. In addition, visual cues such as gaze has been found to be an important cue [3]. This result suggests that turn taking is more evident when participants have eye-contact as compared to when participant can not see each other. This has implications for a dialog with a virtual human. Adding the appropriate turn-taking cues to a talking head has been found to elicit turn-taking [4].

To mimic human speaker shifts, a dialog system should be able to reproduce the response times between speaker shifts seen in human-human interaction. While it is common to use a pause duration threshold, usually around 0.5 s for end-of-utterance (EOU) detection, it has been known for a long time that response times (or gaps) are shorter in conversation. The perhaps first study on this [5] showed a mean response time of 410 ms while the mode (the actual peak of the distribution) was 240 ms. A broader study on this [6] also discusses implications for turn-taking and dialog systems. With a perceptually minimum pause length of 200 ms (or minimum response time), they have shown that 40% of all in between-speaker pauses are long enough for the next speaker to react to cues immediately before the silence. These are the speaker changes which is the scope in this paper.

In human-human conversation, participants commonly utters responses such as "yeah", "mhm", "uhu". Fujimoto [7] points out the problem with the terminology for these tokens, especially concerning the term back-channel and proposes to call these short utterances Listener Responses. These are short utterances or vocalizations which are interjected into the speakers' account without causing an interruption, or being perceived as competitive of the floor. The turn-taking cues preceding Listener Responses has been found to be somewhat different than the cues preceding regular turn-shifts [2][8]. While the general prediction task is to determine at the EOU whether there is a speaker change (SC) or not (HOLD), we here distinguish between speaker shifts which involve Listener Responses (LR) and other turn shifts (TURN-SHIFT). This approach expands previous attempts for prediction [9] [10]. We can formulate these research questions as the following classification tasks:

(A) What is the predictability of TURN-SHIFTs vs. HOLDs?

(B) What is the predictability of LR vs. HOLDs?

(C) What is the predictability of SC (general speaker changes) vs. HOLDs?

In this work, we seek a technical solution for on-line use which relies on acoustic cues to achieve fast response times. We rely on *talkspurts* [11] (also called Inter Pausal Units) and predict the next pattern according to the mentioned tasks for the conditions when there is eye-contact and when there is no eye-contact. We explore a length-invariant time varying feature parametrization which is formulated as a modified type II Discrete Cosine Transform (DCT). This parametrization has many useful properties, such as the separation of segment length (or speaking rate) in the classifier or analysis, and the option of modeling the relative shape of the feature trajectory instead of the absolute shape. This parametrization has been successfully used for analysis and classification of listener responses in previous studies [12][13].

2. The MapTask Corpus

The HCRC Map Task Corpus [14] contains 128 dialogs. The task is for one subject to explain a route to another subject. The one who explains the route is denoted as the "giver" and the one who receives the explanation as the "follower". Half of the dialogs were recorded under a face-to-face condition and the other half under a non-visible condition. The speakers in the "follower" role were excluded since their dialog moves mostly consisted of acknowledgments and clarifications. Two conversations, labeled `q3ec1` and `q3ec5`, were discarded due to a buzz in the speech signal, and `q6ec2` was found to be truncated and hence discarded.

We used the official MapTask annotations concerning the distinction between Acknowledgment Moves (MTACK) and other talkspurts (NONMTACK). The precise definition of an Acknowledgment Move is found in [15], which closely resemble the term Listener Response and thus serve our purpose. It is described as '*a verbal response that minimally shows that the speaker has heard the move to which it responds, and often also demonstrates that the move was understood and accepted*'. The inter-label agreement of the Map Task Corpus annotations are good ($\kappa = .83$).

2.1. Talkspurt segmentation

Based on the annotations provided (excluding the speaker noise tokens "noi"), we segmented the corpus into *talkspurts* [11], defined as a minimum voice activity duration of 50 ms separated by a minimum inter-pause of 200 ms. The resulting connected speech segments are referred to as talkspurts, where the latter threshold is approximately equal to the minimum perceptible pause duration. If a talkspurt is comprised of more than one dialog move, the talkspurt is labeled with the label from the first dialog move included in the talkspurt. In 3.16% of the cases, the merging procedure created talkspurts which started as a (MTACK) and ended as a (NONMTACK). The occurrence of these latter talkspurts are considered to be negligible.

2.2. Between Speaker Intervals

The between speaker interval, can be positive (gap) or negative (overlap). To compute the gaps and overlaps, two cases of overlap are first considered. The first is interjection into complete overlap and the second is partial overlap. The between speaker interval was computed from the partial overlap case and the no overlap case which is shown in Figure 1, where the tails are cut at 2000 ms. The parameters for the distributions are given in Table 1. The mean value for the non eye-contact condition is lower than for the eye-contact condition, which is consistent with the findings of [16] and [17]. However, our results for the standard deviation does not resemble the latter study, which might be explained by the different definitions of a turn. To be able to react to incoming speech in overlap one need around 200 ms to decide whether to continue or stop. From the cumulative distributions it was found that 37% of all speaker shifts occurred before 200 ms for the eye-contact condition, while the same proportion was 44% for the no eye-contact condition. This finding is elaborated on in Section 6.

Condition	MEDIAN	MEAN	STD. DEV.
Eye-contact	320	399	540
No eye-contact	240	313	501

Table 1: Parameters for between speaker interval distributions, measured in ms.

2.3. Automatic labeling procedure

As prediction targets, we use two types of non-overlapped observed speaker changes which are extracted by automatic means. At the end of each NONMTACK talk spurt, the speaker change labels are derived by measuring the silence duration for both speakers, measured from the end of the talkspurt to the start of the next talkspurt in respectively channel. Then a speaker change is assigned if the pause is shorter for the interlocutor,

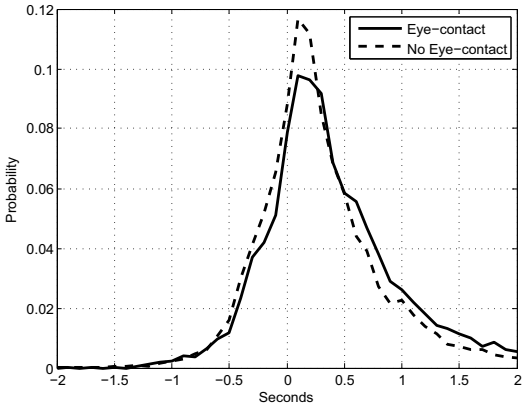

Figure 1: The probability mass function of between speaker intervals using bins of 200 ms. The no eye-contact function is shifted to the left and slightly narrower.

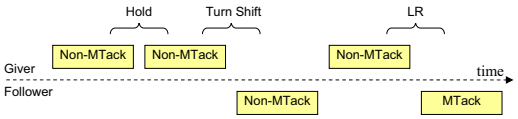

Figure 2: Examples of the three types of labels.

otherwise a non-speaker change is assigned. If the assigned speaker change targets a NONMTACK, it is referred to as TURN SHIFT. If the assigned speaker change targets a MTACK, it is referred to as LR (Listener Response). The joint set of TURN and LR is simply referred to as SC (speaker change) and a non-assigned speaker change is denoted as HOLD. These definitions are illustrated in Figure 2. The non-overlapped labels are chosen as the ones which does not have any overlap in the last 500 ms of each talkspurt. In case the talkspurt is shorter than 500 ms, then the entire talkspurt is checked to ensure that there is no overlap. In addition, if the between-speaker silence is shorter than a minimum response time of 200 ms then it is also considered as an overlapped talkspurt, and is hence ignored. The procedure produces non-overlapped labels for 71%-77% of all talkspurts depending on condition.

3. Feature trajectories as length-invariant Discrete Cosine Coefficients

To parameterize the trajectories of each feature through out a talkspurt, we use DCT coefficients invariant to segment length:

$$X_k = \frac{1}{N}\sum_{n=0}^{N-1} x_n \cos\left(\frac{\pi}{N}(n+\frac{1}{2})k\right) \qquad k = 0,\ldots,N-1$$

where N is the segment length, x_n is the feature value at time n and X_k is the k'th coefficient.

There are several reasons for using this time-varying parameterization. 1) The DCT basis functions are periodic which allows good interpolation of syllabic rhythm in speech. 2) In general, the length-invariance gives a normalization for duration or speaking rate. If duration or speaking rate is added to the final feature vector, then the machine learning algorithm can determine whether it is a salient cue or just speaker variation. For our turn-taking application, sometimes the actual talkspurt is shorter than the look-back window length at the end of each talkspurt, but since the parametrization is length invariant there

is no problem to use the shorter talkspurt instead. 3) These DCT coefficients are also faster to compute than polynomial regression coefficients, since polynomial regression require matrix inversion. 4) The 0'th coefficient is equal to the arithmetic average, which means if it is omitted, then only the relative shape of a trajectory is parametrized. This property is useful for parameterizing features such as F0 (which has a speaker dependent additive bias) , Intensity (which is dependent on the distance to the microphone) or MFCCs (which has an additive channel bias).

Given previous studies on listener response elicitation and turn-taking [2][8], we chose the following feature set: F0 ENVELOPE of the last 500 ms as computed by openSMILE, INTENSITY of the last 1000 ms, SPECTRAL FLUX of the last 1000 ms, which is defined as the L2-norm of energy normalized FTT-bin difference between two adjacent frames. Since spectral flux has been used for estimating tempo in music [18], we hope to capture up-step or down-step in speaking rate via spectral flux, which is much more simple than speaking rate estimation via syllable nucleus detection. Without further motivation, the following two features are also used: DURATION of the previous talkspurt, MFCC of the last 1000 ms.

The acoustic features were extracted by openSMILE [19] at a 10 ms frame rate, where the F0 ENVELOPES are computed by the Sub-harmonic sampling method with octave correction. All F0 ENVELOPES and INTENSITY are first filtered using a moving average filter with a windows size of 3 frames, then the F0 ENVELOPES, SPECTRAL FLUX and DURATION are transformed by the log-operation: $x_{log} = \log_2(1 + x)$. For F0 ENVELOPE the operation gives a perceptually relevant semitone scale, but it also allows for a better fit for Gaussian modeling, which applies to the other features as well. Finally, the features are parameterized in the time dimension using length invariant DCT-coefficients 1-6 (omitting the 0th coefficient, i.e the additive bias), except for DURATION for which the 0'th coefficient (arithmetic average) is used.

4. Experiments

For all experiments, the training set TRAINSET consists of so-called quads 1-4, the development set DEVSET holds quads 5-6 and the evaluation set EVALSET holds quads 7-8. The three sets are speaker independent. By applying the automatic labeling procedure, the counts for the resulting labels are shown in Table 2. For classification, the features were concatenated into a single vector followed by N(0,1) normalization with the mean and variance estimated on training data. Then classification was done using bi-Gaussian Mixture Models (GMM) with diagonal covariances (which was found to work better than Support Vector Machines for this task). The number of DCT coefficients (i.e. the temporal resolution) per feature type are optimized between 1-9 on the development set. Then three ways to combine each feature type is tried out: 1) feature space combination 2) classifier fusion via linear addition of the log-likelihood ratios for each feature type model 3) Linear Discriminant Analysis (LDA) fusion of the log-likelihoods for each class and feature type model, where the LDA prior distribution was set to uniform. Then the three combination schemes are tested for unseen data on the evaluation set.

5. Results

The performance is measured by average recall, which is the average along the diagonal in a confusion matrix. As a rule of thumb, the average recall should be higher than one divided by the number classes.

Eye-contact			
Set	HOLD	TURN SHIFT	LR
TRAINSET	1897	543	496
DEVSET	742	321	282
EVALSET	1606	473	366
No eye-contact			
Set	HOLD	TURN SHIFT	LR
TRAINSET	2412	833	784
DEVSET	983	359	351
EVALSET	865	239	255

Table 2: Counts of automatically extracted labels in the corpus.

The DEVSET results given the optimal number of coefficients is shown in Table 3. For both conditions, the features giving the most contribution are in descending order: DURATION, MFCC and INTENSITY. F0 ENVELOPES was found to more important than SPECTRAL FLUX for the eye-contact condition, while it was found to be the other way around for the non eye-contact condition. Classifier fusion seems to be superior to feature space combination, and the LDA fusion scheme outperforms the linear fusion scheme. Overall, prediction under the no eye-contact condition seems to be more difficult than for the eye-contact condition.

The results for the three feature merging classifiers using optimal time resolution evaluated on unseen data in the EVALSET are shown in Table 4. The same trends as for the DEVSET are observed where classifier fusion is better than feature space combination, and LDA fusion is the best performing classifier (the only exception is for Eye-contact:TURN). In general, prediction seems to be more difficult for the no eye-contact condition while easier for the eye-contact condition, which the exception of no eye-contact: LR. Further investigations may reveal if this latter exception is a fluke or a genuine property of turn-taking when there is no eye-contact.

Eye-contact			
Merging method	TURN	LR	SC
Feat. space	56.12	57.60	56.17
Lin. Fusion	61.45	63.08	61.36
LDA Fusion	60.57	66.35	62.00
No eye-contact			
Merging method	TURN	LR	SC
Feat. space	57.62	56.07	52.76
Lin. Fusion	53.01	58.71	58.13
LDA Fusion	57.82	67.51	60.74

Table 4: Average recall rates in percent for the EVALSET for the three two-class problems: (A) TURN SHIFT vs. HOLD, (B) LR vs. HOLD, (C) SC (speaker change) vs. HOLD.

6. Discussion and Conclusions

This study has compared turn-taking in terms of timing and prediction in human-human conversations under the conditions when participants has eye-contact versus when there is no eye-contact. For prediction we used prosodic and spectral features parametrized by time-varying length-invariant discrete cosine

	Eye-contact			No eye-contact		
	TURN	LR	SC	TURN	LR	SC
F0 ENVELOPES	58.15 (3)	53.59 (1)	52.48 (1)	49.95 (7)	52.86 (4)	53.53 (1)
INTENSITY	61.73 (9)	62.30 (6)	62.17 (9)	54.02 (3)	58.64 (6)	55.62 (3)
SPECTRAL FLUX	55.48 (9)	55.59 (9)	55.59 (8)	52.31 (3)	55.60 (5)	52.37 (3)
DURATION	65.24 (1)	63.44 (1)	64.25 (1)	58.01 (1)	63.40 (1)	61.24 (1)
MFCC	61.76 (6)	64.39 (7)	64.04 (6)	58.35 (9)	59.84 (6)	59.84 (9)
Feat. space	60.76	64.77	62.87	53.60	55.41	54.00
Lin. Fusion	65.14	66.09	65.92	55.58	59.63	59.68
LDA Fusion	67.09	67.52	66.46	59.80	64.98	60.80

Table 3: Average recall in percent for the three types classification tasks on the DEVSET: TURN vs. HOLD, LR vs. HOLD, SC (speaker change) vs. HOLD. The optimal time-resolution specified by the number of DCT coefficients are given in parenthesis.

coefficients. By omitting the 0'th coefficient, which is equal to the arithmetic average, only the relative shape of the feature trajectory is parametrized. With Gaussian Mixture Modeling and variations of classifier fusion schemes, we explored the task of predicting upcoming HOLDs, LRs (listener responses) or TURN-SHIFTs, at a pause lag of 200 ms. The results showed that we can indeed predict upcoming HOLDs from TURN-SHIFTs or LR above chance. The features giving the most contribution are in descending order: Duration, MFCC and Intensity followed by either F0 ENVELOPE or SPECTRAL FLUX depending on condition. The prediction was found to be somewhat easier for the eye-contact condition, which the exception of predicting upcoming LR, which was also the easiest task under both both conditions. From the cumulative distributions of between speaker intervals measured up to 200 ms it was found that 37% of all speaker shifts occurred in overlap for the eye-contact condition, while the same proportion was 44% for the no eye-contact condition. This means that for the no eye-contact condition a larger proportion of of all speaker shifts are either of a non-intrusive floor sharing style or due to interruptions.

The implication of these findings remains to be fully understood, but a possible explanation is that if turn-taking in non-overlap is more difficult in terms prediction under the no eye-contact condition then the same might apply to turn-taking in overlap. Then the larger proportion of speaker shifts in overlap for the no eye-contact condition might be due to unintentional interruptions since one, to a lesser degree, do not know when to talk. Such an interpretation would be consistent with other studies which has found that turn-taking is aided by visual cues.

7. Acknowledgments

Funding was provided by the Swedish Research Council (VR) projects 2009-4291 and 2009-4599. The authors would like to thank Mattias Heldner and members of the eNTERFACE'10 "Continous Interaction for ECAs" team for discussions and comments.

8. 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] A. Gravano and J. Hirschberg, "Turn-yielding cues in task-oriented dialogue," in *Proceedings of SigDial 2009*, 2009, pp. 253–261.

[3] A. Kendon, "Some functions of gaze direction in social interaction," *Acta Psychologica*, vol. 26, pp. 22–63, 1967.

[4] J. Edlund and J. Beskow, "Mushypeek - a framework for online investigation of audiovisual dialogue phenomena," *Language and Speech*, vol. 52, no. 2-3, pp. 351–367.

[5] A. C. Norwine and O. J. Murphy, "Characteristic time intervals in telephonic conversation," *The Bell System Technical Journal*, vol. 17, pp. 281–291, 1938.

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

[7] D. T. Fujimoto, "Listener responses in interaction: A case for abandoning the term, backchannel," *Journal of Osaka Jogakuin 2year College*, vol. 37, pp. 35–54, 2007.

[8] A. Gravano and J. Hirschberg, "Backchannel-inviting cues in task-oriented dialogue," in *Proceedings of Interspeech*, Brighton, 2009, pp. 1019–1022.

[9] K. Laskowski, J. Edlund, and M. Heldner, "An instantaneous vector representation of delta pitch for speaker-change prediction in conversational dialogue systems," in *ICASSP 2008*, 2008, pp. 5041–5044.

[10] L. Ferrer, E. Shriberg, and A. Stolcke, "Is the speaker done yet? faster and more accurate end-of-utterance detection using prosody in human-computer dialog," in *nternational Conference on Spoken Language Processing (ICSLP)*, Denver, Colorado, Sep. 2002.

[11] P. T. Brady, "A statistical analysis of on-off patterns in 16 conversations," *The Bell System Technical Journal*, vol. 47, pp. 73–91, 1968.

[12] J. Gustafson and D. Neiberg, "Prosodic cues to engagement in non-lexical response tokens in Swedish," in *DiSS-LPSS Joint Workshop 2010*, Sep. 2010.

[13] D. Neiberg and K. Truong, "Online detection of vocal listener responses with maximum latency constraints," in *ICASSP 2011*, Czech Republic, 2011.

[14] A. H. Anderson and et al., "The HCRC Map Task corpus," *Language and Speech*, vol. 34, pp. 351–366, 1991.

[15] J. C. Carletta and et al., "The reliability of a dialogue structure coding scheme," *Computational Linguistics*, vol. 23, no. 1, pp. 13–31, 1997.

[16] L. T. Bosch, N. Oostdijk, and J. P. D. Ruiter, "Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues," in *Proceedings 7 th International Conference on Text Speech and Dialogue*. TSD, 2004, pp. 563–570.

[17] M. Bull and M. Aylett, "An analysis of the timing of Turn-Taking in a corpus of Goal-Oriented dialogue," in *ICSLP-1998*, 1998.

[18] M. F. McKinneya, D. Moelants, M. E. P. Davies, and A. Klapuri, "Evaluation of audio beat tracking and music tempo extraction algorithms," *Journal of New Music Research*, vol. 36, no. 1, pp. 1 – 16, Mar. 2007.

[19] F. Eyben, M. Woellmer, and B. Schuller, "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proceedings of ACM Multimedia*, 2010.