Modeling Conversational Interaction Using Coupled Markov Chains

Daniel Neiberg and Joakim Gustafson

Department of Speech, Music and Hearing, KTH, Sweden
jocke@speech.kth.se, neiberg@speech.kth.se

Abstract

This paper presents a series of experiments on automatic transcription and classification of fillers and feedbacks in conversational speech corpora. A feature combination of PCA projected normalized F0 Constant-Q Cepstra and MFCCs has shown to be effective for standard Hidden Markov Models (HMM). We demonstrate how to model both speaker channel with coupled HMMs and show expected improvements. In particular, we explore model topologies which take advantage of predictive cues for fillers and feedback. This is done by initialize the training with special labels located immediately before fillers in the same channel and immediately before feedbacks in the other speaker channel. The average F-score for a standard HMM is 34.1%, for a coupled HMM 36.7% and for a coupled HMM with pre-filler and pre-feedback labels 40.4%. In a pilot study the detectors are found to be useful for semi-automatic transcription of feedback and fillers in socializing conversations.

Index Terms: fillers, feedbacks, coupled hidden markov models, cross-speaker modeling, conversation

1. Introduction

A naturally occurring spontaneous conversation is a process where the participants influence each other. It is reasonable to model all participants in the interaction process to improve automatic detection of communicative vocalizations [1]. Successful examples of this include detecting emotions [2], engagement [3], turn-taking behavior [4] and Dialog Acts (DA) [5][6][7]. Automatic transcription and classification of Dialog Acts (DAs) may be done on a pure lexical level, or by using prosody alone [8], or a combination thereof [9,10]. However, it is not straightforward to train language-models for non-lexical content such as "'mm", "mhm" and "eh" since non-verbal features like speaking rate, pitch slope and voice quality determine their meaning. Furthermore, machine learning of the meaning of these conversational tokens is hampered by the lack of standardized annotation schemes. Non-lexical conversational tokens are usually found in dialog acts which can be roughly divided into those that are interjected into one's own speech (fillers) and those that are interjected into the interlocutor's account (feedback). Different kinds of feedback tokens, such as back-channels, acknowledgments and agreements often share the same type of phonetic content. Because of this it seems to be necessary to use prosodic features and models to construct detectors for these non-lexical conversational tokens. In this paper we aim to construct a detector for semi-automatic annotation which can discriminate between fillers and feedback by modeling the both participants in a conversational interaction.

Using interlocutor (non-target speaker) information to boost back-channel detection, rather than prediction, is not a very common practice. One study [7] use a rule system based on speech durations in dual channel conversational recordings to detect back-channels. Instead of using pitch tracker derived features, the Fundamental Frequency Variation spectrum (FFV) has been used to classify dialog acts [8] which included two types of fillers, back-channels, acknowledgments among others in a multi-participant meeting. This study showed the benefit of using spectral correlates to fundamental frequency change without using a token-based language model, which may be inappropriate for non-verbal tokens. One finding was that non-target prosodic context improves detection of DA interruption.

In the current study we aim to construct a detector for fillers and feedbacks and aims for semi-automatic transcription of corpora. Three specific goals are addressed:

1. Capture the prosodic characteristics of fillers and feedback by using a normalized fundamental frequency cepstrum representation suitable for Hidden Markov Modeling, as well as standard MFCC as auxiliary features
2. Investigate the benefit of modeling the dyadic interaction, by using one Markov chain per speaker and a joint coupled transition matrix.
3. Explore model topologies which take advantage of predictive cues for fillers and feedback. For example, in a study of conversational Japanese and English [11], it was found that back-channels may be predicted by a region of low pitch of the interlocutor. This finding has also been confirmed for Swedish [12]. Another study found cues connected to intonation and both average intensity and F0 [13].

In Section 2, the DEAL corpus is described, in Section 3 a normalized fundamental frequency cepstral representation is outlined, in Section 4 experiments using single and dual chain Hidden Markov Models are reported, which is followed by a pilot study of the semi-automatic annotation of fillers and feedback token in a different kind of dialogue corpus.

2. The DEAL corpus

The current study uses data from the DEAL corpus [14]. It consists of dialog data recorded as informal, human-human, face-to-face task-oriented dialogues. 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 the best possible price from the shop-keeper.

The recordings were done with one microphone per speaker, and recorded at 16 kHz in two channels. All dialogs were first transcribed orthographically including non-lexical entities such as laughter and hawks. Filled pauses, repetitions, corrections, restarts and cue phrases were labeled manually. The DEAL corpus is rich in fillers and feedback tokens. The feedbacks are generally single words (99%) or non-lexical units and appear in similar dialog contexts (i.e. as responses to assertions). The feedbacks are labeled according to attitude; news receiving, dis-preference or general feedback, but in this study the attitude is not addressed.

3. A normalized fundamental frequency cepstral representation

The procedure of finding correlates to pitch starts with calculating a Constant-Q filter bank in a semitone scale. Then the mean F0 is estimated and the filters which are located up to 8 semitones from the estimated mean F0 are retained. These filters are finally used to obtain a normalized fundamental frequency cepstrum. This entire procedure follows the description in [15] were more details are found. Here, an overview of the original approach is given, including a few modifications.

The filter-bank is based on the Constant-Q transform [16] with a corresponding Q factor of $1/(2^{1/12}-1)$ or 16.8 which corresponds to the 12 semitones per octave in a musical scale. The filter-bank spans a total of 81 bins between 60 Hz and 6458 Hz, which is below the Nyquist frequency. Compared to Short-time Fourier Transform (STFT), the constant-Q transform has optimal temporal-spectral resolution for all filters, which means there is no need to optimize the analysis window length for different applications. A standard frame shift rate of 100 Hz is used.

To provide a reference for normalization, a simple method of finding an average F0 within each Inter Pausal Unit (IPU), given by the labels collapsed into speech (as described in Section 4), is proposed. The basic idea is summing harmonics for each filter in the semitone scale per frame. The maximum number of harmonics to sum over is 12 because beyond that consecutive harmonics would fall under the same bin, but here only the first 8 harmonics are considered to give a reasonable frequency resolution for higher order harmonics. An approximation to tone versus noise separation is used which classifies all frequencies with amplitudes below 10 dB from the highest amplitude frequency component as noise. So any filter above this threshold occurring in the output of the filter-bank are considered as tones, which means that the harmonic summing starts at the first index containing non-noise. The per frame estimated F0 is then found by the semitone corresponding to maximum of the per filter harmonic summation. Then the average F0 is found by a weighted average of the per frame estimated F0s using power amplitudes as weights. This is not just motivated by a study which found frequencies at higher intensity levels to be more salient [17], but it also removes the need for voicing decision. If the IPU is marked as non-speech, then the mean frequency is set to 240 Hz. To obtain a normalized F0 spectrum, the range which is within 8 semitones from the mean frequency is retained. While this implies the assumption of a maximum F0 variation of 17 semitones, it reduces the influence of the first overtone which is located at 1 octave (12 semitones) in average.After the log power spectrum is obtained, the cepstrum is calculated by applying a one dimensional discrete cosine transform (DCT).

4. Experiments

For this study we use the six first DEAL dialogs that were labeled at the time. Six-fold cross validation on dialog level is used. The labels for silence, breath and hawks are collapsed into the silence label. Similarly, all speech acts other than fillers and feedback are collapsed into the speech label. With fillers and feedbacks, this gives us four labels in total, but this number may increase by inserting special labels described later on.

Single channel experiments are conducted using standard Hidden Markov Models with emitting distributions modeled as Gaussian Mixtures. As features we use Normalized F0 cepstra and RASTA processed MFCCs, where the RASTA processing removes spikes and channel bias. For the normalized F0 cepstra, the first 6 coefficients are retained. For both the normalized F0 cepstra and the MFCCs, we add the delta along with delta-delta coefficients calculated over a window of 9 frames. A standard 3 state left-to-right topology with 4 Gaussians per state is adopted for each label. The parameters of each left-to-right HMM is estimated using the Baum–Welch algorithm. A global HMM is constructed from the single HMMs, with the help of bi-gram statistics calculated for the labels in the training data. Since the coupled HMMs require high computational effort and put high demands on memory resources, we do not explore higher number of Gaussians and reduce the feature dimension by PCA. The F0 cepstra is reduced from 18 to 10 dimensions and the MFCCs are reduced from 39 to 15 dimensions.

Initial experiments using only cross-validation rotation one and five was conducted and the average F-score was measured on frame level. The score for F0 cepstrum was 24.7%, for RASTA processed MFCC 31.1%, for a combination of the two 37.0% and for PCA projected features 38.1%. Thus, performance is increased in a sequence of steps. While the durations of the collapsed labels have little meaning, the durations of fillers and feedback are shown in Figure 1 and follow two slightly different distributions. To reduce confusion between fillers, feedback and regular speech a shared duration threshold is applied after the recognition pass. Thus, any filler or feedback shorter than 90 ms is classified as speech. This threshold is set such that 5% of fillers and feedbacks segments are lost. This allows us to filter out short schwa-vowels with may be confused with fillers, or too short durations caused by random state-switching.

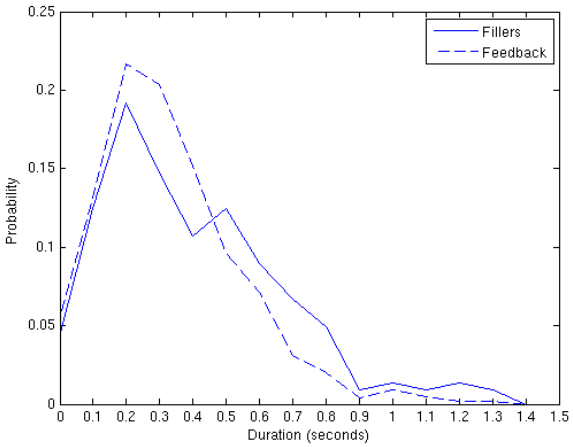

Figure 1: *Distributions durations in fillers and feedbacks*.

A fully dual coupled Hidden Markov Model [18] is basically two standard HMMs where each emitting density is a function of two state variables and the state transition probabilities are conditioned on the previous states in both Markov chains using a joint transition matrix. This type of HMM is a natural extension to capture interaction in dyadic conversations. If the model is supposed to be speaker independent, then the cross chain conditional probabilities and emitting distributions has to be symmetric for the two channels. For pragmatic and computational purposes, we are only retraining the joint transition matrix. First the state sequences for each label is estimated by a Viterbi search for each speaker channel using the single channel HMMs. The necessary statistics for the joint transition matrix is then accumulated symmetrically for the two channels. A joint channel feature space is created by catenating the feature vectors for the two channels. To create joint channel emitting distributions, the Gaussian Mixtures from two states, each belonging to two different Markov

chains, are pooled causing a doubling in the number of parameters. The GMM weights are made sure to sum to one by dividing with the sum after pooling. For all coupled HMM experiments, the PCA projected features are used to reduce the computational burden, and a duration threshold is applied after the recognition pass.

A filler may be preceded by distinct cues which may be captured by a special pre-filler label in the same channel as the filler producing speaker. Initial experiments confirmed that a good duration for this extra label is 500ms. The pre-filler labels are marked backwards in time before each filler, but are terminated as soon as any other label than the speech or silence label is encountered or when the initial duration has passed. Thus, the initial durations are maximum durations. The bi-gram statistics which are used to create the global transition matrix will then force the pre-filler states to be connected to the filler states in sequence.

The pre-feedback labels are marked in the same way as the pre-filler labels, but in the other channel where only speech labels are overwritten. Initial experiments confirmed that a good duration for this extra label is 1000ms. The pre-feedback labels in the other channel will be forced to precede the feedback label in the target channel via the joint transition matrix. Feedback labels may be detected without pre-feedback labels where the feedback is preceded by silence in the non-target channel. It should be noted, that previous studies [11], where back-channels are predicted from interlocutor cues, report only modest accuracy so the performance boost is expected to be small. Examples of this label initialization are shown in Figure 2.

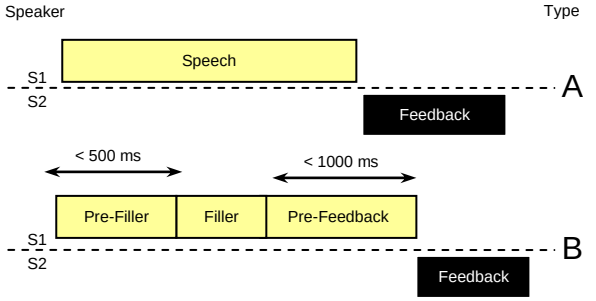

Figure 2: *Examples of initialization of (A) coupled HMM (B) coupled HMM using pre-filler and pre-feedback labels.*

The following configurations are reported for the full evaluation:

- **HMM-Comb-PCA:** F0 and RASTA MFCC combined in the same feature vector with PCA projection;
- **HMM-Comb-D-PCA:** HMM-Comb-PCA with a duration threshold;
- **CHMM-Comb-D-PCA:** Comb-D-PCA using coupled HMMs
- **P-CHMM-Comb-D-PCA:** Coupled-Comb-D-PCA using a pre-filler state with duration of 500 ms a pre-feedback label with duration 1000 ms.

The results of the experiments are shown in Table 1. Performance is measured in F-scores for fillers and feedback on frame level, as well as the average of the two. F-score is defined as the harmonic mean between precision and recall. Improvements are observed in order of appearance, and P-CHMM-Comb-D-PCA is the final and best configuration.

Table 1: *F-scores given in percent for final experiments.*

Configuration	Filler	Feedback	Avg.
HMM-Comb-PCA	36.6	30.6	33.6
HMM-Comb-D-PCA	37.1	31.0	34.1
CHMM-Comb-D-PCA	41.6	31.7	36.7
P-CHMM-Comb-D-PCA	46.5	34.2	40.4

5. Discussion on experiments

Although it seems that coupled HMMs have a clear advantage over regular HMMs for this task, parts of the success may have a simple explanation. There is a constant leakage of speech between the two channels, and sometimes it increases significantly when speakers move during animated discussions. Given the typical conversation style in Swedish where overlapped speech is the exception rather than the rule; the joint transition matrix should suppress errors due to cross-talk. However, other factors may also have contributed. For example, back-channels are often uttered in overlap, fillers are expected to be uttered when the interlocutor remains silent and feedbacks are unlikely when the interlocutor utter a feedback. Also, any systematic overlap may be modeled by the joint transition matrix via the cross channel dependencies for the individual states in each left-to-right HMM.

No attempt has been made to balance precision and recall, but in all experiments the recall rates are 1.8 times higher than the precision. This may not be of any concern if the aim is to use the output for semi-automatic transcription.

6. A semiautomatic annotation pilot

Semi-automatic transcription of fillers and feedback is expected to reduce labor work considerably for large corpora. We have recently collected about 60 hours of audio, video and motion capture data in human-human conversations within the project Spontal [19]. We are encouraged by previous reported attempts [7, 11], where the last study reported semi-automatic annotation of more than 3000 feedbacks in less than 4 hours. It is our intention to investigate if semi-automatic annotation of fillers and feedback in the Spontal corpus is feasible using the P-CHMM-Comb-D-PCA detector. The Spontal corpus consists of recordings of spontaneous face-to-face spontaneous socializing conversations where the participants have received minimal directions for task and topic. For this pilot study we use one recording of 5 minutes between a male and a female speaker. The result of a manual analysis of the automatic detection is summarized in Table 2.

Table 2. *The number occurrences of correctly and wrongly classified fillers and feedbacks, with manually tagged reasons for errors within the parenthesis.*

Feedbacks		**Fillers**	
correct	82	correct	15
missed	7	missed	0
wrong	9	wrong	14
wrong (extraling)	10	wrong (extraling)	3
wrong (cross talk)	20	wrong (prolonged)	7
wrong (feedback within own IPU)	83	wrong (cross talk)	8

This small corpus consisted of 89 feedback tokens and 15 filled pauses. All of the fillers and 92% of the feedbacks were successfully detected. To achieve this high recall rate, we have to accept a low precision rate. However, we want to keep the wrongly detected tokens as few as possible. The overall number of incorrectly detected feedback tokens was 42% and the number of wrongly identified fillers was 68%. We have performed an error analysis on these results. Most of the wrongly detected feedback tokens were due to channel leakage. However, these could be detected and removed automatically in a subsequent filtering step that makes use of cross-talk detection, which of course is a difficult problem by itself. If this is done the number of wrongly identified feedback tokens falls to 18%. Half of these contained extra-linguistic sound like coughs. The number of cross talk related wrongly detected fillers were few, but if these were removed the number of falsely detected fillers dropped to 62%. One third of these were prolongation of other sounds than filled pauses, and 12% were extra-linguistic sounds. The manually judged quality of the segmentation is high. In almost all cases the detected feedbacks and fillers had correct start times and end times.

7. Conclusions

Series of experiments for automatic transcription and classification of fillers and feedbacks have been reported. A feature combination of PCA projected normalized F0 Constant-Q Cepstra and MFCCs has been shown to be effective for standard Hidden Markov Modeling. It is demonstrated how to model each speaker channel with coupled HMMs and expected improvements are confirmed. In particular, model topologies which take advantage of predictive cues for fillers and feedback have been explored. This was done by initializing the training with special labels located immediately before fillers in the same channel and immediately before feedbacks in the other speaker channel. The pre-feedback and pre-filler states in the one channel will be forced to precede the feedback in the other channel via the joint transition matrix.

The feedback and filler detectors were trained on task oriented dialogues. In order to verify the generalization of the detectors we decided to test them on socializing conversations. In this pilot experiment we evaluated the efficiency of our semi-automatic annotation of feedback and fillers. In semi-automatic transcription the initial automatic detector needs to correctly find and segment as many fillers and feedback tokens as possible. Our pilot showed that all fillers and 92% of the feedback were found, with correct segmentation. Despite the fact that we did joint channel modeling there were still some problems with cross talk sections. This is a consequence of the decision to opt for a high recall rate on the expense of lower precision.

8. Acknowledgments

This research is carried out at KTH Speech, Music and Hearing. Funding was provided by the Swedish Research Council (VR) project "*Introducing interactional phenomena in speech synthesis"* (2009-4291). The authors would like to thank Anna Hjalmarsson for proving the DEAL corpus with annotations for fillers and feedback.

9. References

[1] Laskowski, K. "Modeling Norms of Turn-Taking in Multi-Party Conversation", In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 999–1008,Uppsala, Sweden,

[2] C.-C. Lee, C. Busso, S. Lee, and S. S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions", in INTERSPEECH-2009, Brighton, UK, 2009, pp. 1983-1986.

[3] C. Yu, P.M. Aoki, and A. Woodruff, "Detecting user engagement in everyday conversations" in In Proc. 8th Int. Conf. on Spoken Language Processing (ICSLP), 2004, pp. 1-6.

[4] T. Choudhury and S. Basu, "Modeling conversational dynamics as a mixed-memory markov process" in Advances in Neural Information Processing Systems 17, L. K. Saul, Y. Weiss, and L. Bottou, Eds. Cambridge, MA: MIT Press, 2005, pp. 281-288.

[5] K. Laskowski and E. Shriberg, "Modeling other talkers for improved dialog act recognition in meetings" in INTERSPEECH-2009, Brighton, UK, 2009, pp. 2783-2786.

[6] Gravano, S. Benus, J. Hirschberg, S. Mitchell, and I. Vovsha, "Classification of discourse functions of affirmative words in spoken dialogue" in Interspeech, Antwerp, 2007, pp. 1613-1616.

[7] U. Sajjanhar and N.Ward, ."Automatic labeling of back channels", University of Texas at El Paso, Tech. Rep., 2006.

[8] K. Laskowski and E. Shriberg, "Comparing the contributions of context and prosody in text-independent dialog act recognition", in 35th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2010). Dallas TX, USA, March 2010.

[9] V. K. R. Sridhar, S. Bangalore, and S. Narayanan, "Combining lexical, syntactic and prosodic cues for improved online dialog act tagging". Computer Speech and Language, vol. 23, no. 4, pp. 407-422, 2009.

[10] M. Zimmermann "Joint segmentation and classification of dialog acts using conditional random fields" in INTERSPEECH-2009, 2009, pp. 864-867.

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

[12] Edlund, J., Heldner, M., & Pelcé, A. Prosodic features of very short utterances in dialogue. In Vainio, M., Aulanko, R., & Aaltonen, O. (Eds.), Nordic Prosody - Proceedings of the Xth Conference (pp. 57 - 68). Frankfurt am Main: Peter Lang., 2009.

[13] Gravano, A. and Hirschberg, J. "Backchannel-inviting cues in task-oriented dialogue", In INTERSPEECH-2009, 1019-1022, 2009.

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

[15] D. Neiberg, P. Laukka, and G. Ananthakrishnan, "Classification of affective speech using normalized time-frequency cepstra",. In Prosody 2010, May, 2010.

[16] J. Brown, "Calculation of a constant Q spectral transform", J. Acoust Soc of Am, vol. 89, no. 1, pp. 425.434, 1991.

[17] B. C. J. Moore, "An Introduction to the Psychology of Hearing", 3rd ed. Academic Press Limited, 1989.

[18] M. Brand, "Coupled hidden markov models for modeling interacting processes", MIT Media Lab Vision and Modeling, Tech.Rep., 1996.

[19] Edlund, Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., & House, D. Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M., & Tapias, D. (Eds.), Proc. of the Seventh conference on International Language Resources and Evaluation (LREC'10) (pp. 2992 - 2995). Valetta, Malta, 2010.