A Dual Channel Coupled Decoder for Fillers and Feedback.

Conference Paper · January 2011

Source: DBLP

CITATIONS
2

READS
52

2 authors:

Daniel Neiberg
KTH Royal Institute of Technology
33 PUBLICATIONS **457** CITATIONS

SEE PROFILE

Joakim Gustafson
KTH Royal Institute of Technology
154 PUBLICATIONS **1,293** CITATIONS

SEE PROFILE

A Dual Channel Coupled Decoder for Fillers and Feedback

D. Neiberg, J. Gustafson

Department of Speech, Music and Hearing, Royal Institute of Technology (KTH), Sweden

[neiberg,jocke]@speech.kth.se

Abstract

This study presents a dual channel decoder capable of modeling cross-speaker dependencies for segmentation and classification of fillers and feedbacks in conversational speech found in the DEAL corpus. For the same number of Gaussians per state, we have shown improvement in terms of average F-score for the successive addition of 1) increased frame rate from 10 ms to 50 ms 2) Joint Maximum Cross-Correlation (JMXC) features in a single channel decoder 3) a joint transition matrix which captures dependencies symmetrically across the two channels 4) coupled acoustic model retraining symmetrically across the two channels. The final step gives a relative improvement of over 100% for fillers and feedbacks compared to our previous published results. The F-scores are in the range to make it possible to use the decoder as both a voice activity detector and an illucotary act decoder for semi-automatic annotation.

Index Terms: filler, feedback, coupled hidden Markov models, cross-speaker modeling, conversation

1. Introduction

In human-human conversation interlocutors do not speak in isolation, but rather mutually interacts with one another in the discourse. Thus, automatic detection of communicative signals in conversation should benefit from modeling the cross-speaker dependencies in the interaction process. Examples of this include detecting affect [1], engagement [2], turn-taking behavior [3] and Dialog Acts (DA) [4].

The most trivial example of the benefit of modeling cross-speaker dependencies is found in voice-activity detection (VAD), simply because overwhelmingly one speaks at a time. This property of conversation allows for suppression of cross-talk which is a main source of VAD error in conversational corpora with close-in microphones. In [5], interlocutor dependencies were taken into account via a HMM-based framework. Cross-talk suppression was achieved by a Viterbi-search through the joint channel space, governed by joint transition matrix. Alternatively, cross-talk may be suppressed by including cross-talk suppression features in a HMM-based single channel frame-work [6]. The approach by [5] can be directly extended to modeling interlocutor dependencies for text-independent joint segmentation and classification of DAs [7]. The decoder was based on prosodic features, most notably the Fundamental Frequency Variation spectrum (FFV) which is a vector-valued filter-bank with correlates to pitch change. The target DAs included two types of fillers, back-channels, acknowledgments among others in a multi-participant meeting.

In our previous work [8], we constructed a detector for non-lexical tokens with the aim of semi automatic transcription of corpora. These non-lexical tokens are usually found in dialog acts which can be roughly divided into those that are interjected into one's own speech (filled pauses or fillers) and those that are interjected into the interlocutor's account (feedback or back-channels). Feedback often occupy an entire inter-pausal unit (IPU) although exceptions are found [9] while fillers often initiate an IPU. This points towards a frame-based solution, for which a Hidden Markov Model framework is the standard solution. The experiments showed the benefit of capturing 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. Additional improvement was achieved by modeling the dyadic interaction, by using one Markov chain per speaker and a joint coupled transition matrix. Finally, we explored model topologies which take advantage of predictive cues for fillers and feedback. An analysis of the errors of the final decoder, which used all the mentioned techniques, showed that a large proportion of the errors was still due to cross-channel talk.

In this study we expand the task of detecting fillers and feedback by also including silence and other types of speech. This extension allows for using the decoder as a speech activity detector, and also gives the position of both fillers and feedback within an IPU. This information is useful for distinguishing between turn-initial feedback and feedback which occupy an entire IPU (the same applies for fillers). Based on the decoder introduced in [8], the following improvements are reported:

1. The frame-shift rate is increased to 50 ms. This reduces the complexity of the decoder which improves speed and allows longer duration modelling for the same number of states per HMM which leads to improved recognition performance.
2. A cross-correlation feature, computed as the maximum cross-correlation between the channels normalized by the energy in the non-target channel (denoted as NMXC in [6] and JMXC in [10]). This feature is an approximation for the distance to the microphone.
3. Full coupled speaker-independent training of both the joint transition matrix and the acoustic models formed by Cartesian product.

In Section 2, the DEAL corpus is described, in Section 4.1 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 Discussion and Conclusions.

2. The DEAL corpus

This study uses data from the DEAL corpus [11]. It consists of dialog data recorded as informal, human-human, face-to-face conversations. 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 shopkeeper.

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, breath and hawks (the sound of clearing ones throat). 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.

To facilitate comparison with our previous results we use the six first DEAL dialogs. The labels for silence, breath and hawks are collapsed into the SILENCE label. Similarly, all speech acts other than FILLER and FEEDBACK are collapsed into the SPEECH label. With FILLER and FEEDBACK, this gives us four labels in total.

3. Single- and Dual-Channel Statistics

As stated in the introduction, a coupled channel topology can give a richer and more accurate description of a dyadic interaction compared to a single-channel topology. To give a grasp of this, we demonstrate a few obvious statistical examples from the corpus. To facilitate comparison, the labels are quantized to a 50 ms frame rate by rounding the start and end points towards the nearest 50 ms border.

First, the single-channel monogram probabilities are computed. From these, a dual channel cross product is formed, which is compared to the true dual channel monograms. The three type of monogram probabilities are found in Table 1. One can see that the most frequent single-channel labels in descending order are SILENCE, SPEECH, FEEDBACK and FILLER. When the two types of dual channel monograms are compared, the most striking difference is the overestimation of SILENCE-SILENCE for the cross-product. More differences can be observed, but are not addressed since the point is to show the inadequacy of both the single-channel and dual channel cross-product monograms. The estimated dual channel monograms shows that the two most probable events in descending order are SPEECH-SILENCE and SILENCE-SILENCE. Thus, overwhelmingly one speaks at a time or both are mutually silent. We also observe that SPEECH-FEEDBACK is almost as common as SPEECH-SPEECH. In fact, FEEDBACK represent 39% of all kind of SPEECH in overlap.

These trivial observations clearly shows that modeling cross-speaker dependencies should be superior compared to modeling each participant in isolation. A similar study can also be done for the frame bigram statistics, but this not shown due to space restrictions.

Table 1: *Single- and dual-channel monogram probabilities in percent, estimated from frames at a 50 ms frame rate. Each sub-table sums to 100*

Single channel monograms				
	SILENCE	SPEECH	FILLER	FEEDBACK
	62.10	32.38	1.16	4.35
Dual channel monograms by single channel cross-product				
	SILENCE	SPEECH	FILLER	FEEDBACK
SILENCE	51.7	-	-	-
SPEECH	26.9	14.0	-	-
FILLER	0.97	0.50	0.02	-
FEEDBACK	3.62	1.89	0.07	0.25
Dual channel monograms				
	SILENCE	SPEECH	FILLER	FEEDBACK
SILENCE	30.27	-	-	-
SPEECH	55.89	3.15	-	-
FILLER	2.01	0.18	0.01	-
FEEDBACK	5.76	2.39	0.12	0.21

4. Decoder Design

The decoder is an improved version of the one in [8]. In Section 4.1 the acoustic features are briefly described and in Section 4.2 the description of acoustic modeling is given.

4.1. Acoustic features

The following acoustic features are considered:

- RASTA processed MFCCs, where the RASTA processing removes spikes and channel bias.
- Joint Maximum Cross-Correlation (JMXC), i.e the maximum cross-correlation between channels, normalized by the energy of the non-target speaker and computed with a 75 ms window;
- A Normalized Fundamental Frequency Cepstral Representation. The procedure of finding correlates to pitch starts with calculating a Constant-Q filter bank [12] in a semitone scale. The filter-bank spans a total of 81 bins between 60 Hz and 6458 Hz, which is below the Nyquist frequency. The mean F0 is estimated based on harmonic summation and without the need for a voicing decision. Then 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, where the first 6 coefficients are retained. The procedure used is the same as for [8], which is the frame-based one-dimensional version of the two-dimensional segment based approach used in [9], and more details are found in [13].

For both the normalized F0 cepstra and the MFCCs, we add the delta along with delta-delta coefficients calculated over a window of 5 frames (HTK-type delta).

4.2. Acoustic Modeling

Single-channel experiments are conducted using standard Hidden Markov Models with emitting distributions modeled as Gaussian Mixtures. A standard 3-state left-to-right topology is adopted for each label. 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 reduce the feature dimension by principal component analysis. The F0 cepstra is reduced from 18 to 10 dimensions and the MFCCs are reduced from 39 to 15 dimensions.

A fully dual coupled Hidden Markov Model [14] 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, see Fig. 1. This type

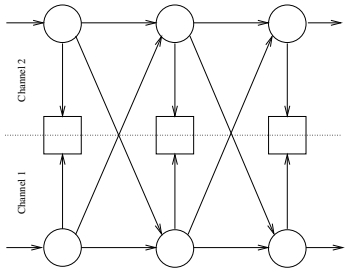

Figure 1: *Inference graph for a coupled double chain HMM with joint feature space.*

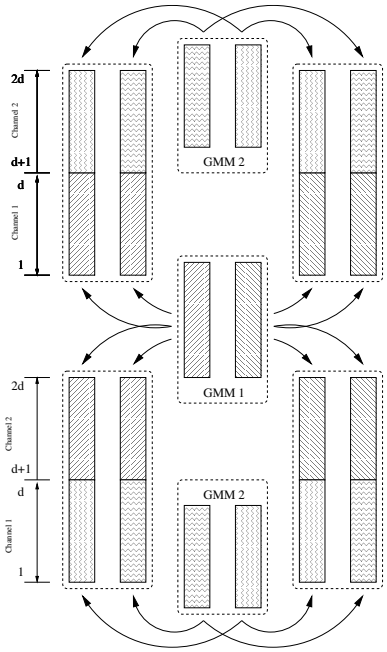

Figure 2: *A Cartesian product between the mean parameters of two states with two Gaussians per state each modeling a d-dimensional feature space, forming $2d$ dimensional Gaussians in a coupled HMM with a joint channel feature space. The diagonal covariances are constructed similarly.*

of HMM is a natural extension to capture interaction in dyadic conversations.

Consider two Markov chains, one for each of the two channels Q and R. Let q be a state in channel Q and let r be a state in channel R. Then let a joint cross-channel emitting distribution $b_{q,r}(o_t)$ connect each state q in channel Q with each state r in channel R. To clarify, there is not a single emitting distribution for each single state, but one emitting distribution for each possible pair of q and r. This is a consequence of forming a Cartesian self product to a joint channel feature space, see Figure 2, so $o_t = [o_t^q, o_t^r]$. The likelihood of the observation sequence $O = \{o_1 o_2 ... o_T\}$ and the two state sequences q and r given the model is

$$P(O, q, r|\lambda) = \pi_{q_1,r_1} b_{q_1,r_1}(o_1) \prod_{t=2}^{T} a_{q_{t-1},r_{t-1},q_t,r_t} b_{q_t,r_t}(o_t)$$

where π is the probability of the occurrence of the state at the first time instant ($t = 1$) and a is the coupled probability of transition from a pair of states in the two channels, q_{t-1} and r_{t-1} at time $t-1$ to another pair of states q_t and r_t at time t. If the model is supposed to be channel-symmetric then the cross-chain conditional probabilities and emitting distributions have to be symmetric for the two channels, i.e. $b_{q_t,r_t} = b_{r_t,q_t}$, $a_{q_{t-1},r_{t-1},q_t,r_t} = a_{r_{t-1},q_{t-1},r_t,q_t}$ and $\pi_{q_1,r_1} = \pi_{r_1,q_1}$.

The coupled training procedure goes as follows: First the state sequences for each label are estimated by a Viterbi search for each speaker channel using the single-channel HMMs. Then the necessary statistics are accumulated symmetrically for the two channels. The Cartesian product is formed. Specifically, the Gaussian Mixtures from two states, each belonging to two different Markov chains, are merged to a joint feature space and the individual Gaussians are combinatorially assembled, squaring the number of parameters, see Figure 2. The GMM weights are normalized such that they sum to one. The joint transition matrices and initial probabilities are formed by cross products. The joint transition matrix and emitting distributions are then updated by using the sufficient statistics from the Viterbi search as an approximation for the expectation step in a single EM-iteration.

The following configurations are considered:

S: Single-channel modeling with vector valued F0 correlates (as computed in 4.1) and RASTA processed MFCC, both computed either at a 10 ms or 50 ms shift-rate (with scaled window sizes for the latter) and combined in the same feature vector with PCA projection creating a 25 dimensional feature space;

CS: Same as the S 50 ms configuration above but with the JMXC as an additional feature, i.e the maximum cross-correlation normalized by the energy of the non-target speaker computed with a 75 ms window, creating a 26 dimensional feature space;

CDT: A Cartesian product of CS creating a 52 dimensional feature space, followed by coupled channel symmetric training of the joint transition matrix

CDTA: Same as CDT, followed by coupled channel symmetric training of the emitting distributions (GMMs)

5. Experiments and Results

The evaluation is done by leave-one-out training on dialog level with round-robin rotation. Performance is measured in F-scores for SPEECH, SILENCE, FILLER and FEEDBACK on frame level, as well as the average of the four. F-score is defined as the harmonic mean between precision and recall. The results are presented as three types of comparisons to balance for fairness, understanding, overview and detail.

In the first comparison, the average F-score of all five configurations given the same number of Gaussians per state before Cartesian product are shown in Table 2. It is clear that performance increases in order of appearance from top to down: increasing frame rate to 50 sm, adding the cross-correlation feature CS, adding a channel symmetric joint transition matrix CDT, and channel symmetric coupled GMM re-training. A saturation at 12 Gaussians per state is observed for CDTA. However, this comparison is not entirely fair, since the CDTA GMMs in fact has the square the number of retrained Gaussians per state as for the other models.

In the second comparison, the single- and dual-channel HMMs CS and CDTA are compared for the same number of trained Gaussians per state. The results are shown in Table 3 and it is clear that coupled retraining outperform single channel models.

Table 2: *Average unweighted F-scores in percent for five configurations and the same number of pre-coupled Gaussians per state.*

	Gaussians per state			
Configuration	4	8	12	16
S (10 ms)	57.2	58.5	59.5	59.7
S	57.8	59.7	60.8	61.1
cS	65.9	67.9	68.9	68.7
cDT	70.7	73.0	74.7	76.2
cDTA	77.7	83.8	86.9	85.7

Table 3: *Average unweighted F-scores in percent for the same number of retrained Gaussians per state.*

	Gaussians per state			
Configuration	16	64	144	256
cS	61.1	69.3	67.1	65.2
cDTA	77.7	83.8	86.9	85.7

In the third and last comparison, the best performing configurations are selected based on average F-score. Then the number of retrained Gaussians per state are varied. The per class and average F-score for the best three different configurations with the cross-correlation feature are shown in 4. While the average F-score increases as one inspect the table from left to right, i.e. from cS, cDT to cDTA, there is little improvement for SPEECH and SILENCE, especially between cS, cDT, while a dramatic improvement for FILLER and FEEDBACK.

Table 4: *F-scores for the three best configurations.*

	Configuration (G.per state)		
Act	cS (64)	cDT (16)	cDTA (144)
SILENCE	93.6	93.4	96.8
SPEECH	85.0	85.6	92.4
FILLER	42.3	68.6	80.9
FEEDBACK	46.2	53.4	76.7
Avg.	69.3	76.2	86.9

6. Conclusions

Analysis and experiments for automatic segmentation and classification of SILENCE, SPEECH, FILLER and FEEDBACK have been reported. We show examples of rather simple cross-speaker dependencies which are possible to model by cross-channel Markov chains. Based on our previous approach, we start presenting a single-channel baseline which uses a feature combination of PCA projected normalized F0 Constant-Q Cepstra and MFCCs for standard Hidden Markov Modeling. For the same number of Gaussians per state, we have shown improvement in terms of average F-score for the successive addition of 1) increased frame rate from 10 ms to 50 ms 2) JMXC features in a single channel decoder 3) a joint transition matrix which capture dependencies symmetrically across the two channels 4) coupled acoustic model retraining symmetrically across the two channels. The improvement between step 2 and 3 is only due to FILLER and FEEDBACK, while SILENCE and SPEECH has the same F-scores for the two configurations. The final fourth step shows improvement for all four classes, and gives a relative improvement of over 100% compared for FILLER and FEEDBACK compared to our previous published results. The F-scores are in the range to make it possible to use the decoder as both a voice activity detector and an illucotary act decoder. The final system is currently used for semi-automatic annotation of a large conversational corpus recored at our lab.

7. Acknowledgments

This research is carried out at KTH Speech, Music and Hearing. Funding was provided by the Swedish Research Council (VR) projects 2009-4291 and 2009-4599. The authors would like to thank Anna Hjalmarsson for proving the DEAL corpus, Ananthakrishnan Gopal for Matlab-code, Kornel Laskowski and Giampero Salvi for comments.

8. References

[1] 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.

[2] 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.

[3] 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.

[4] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP*, 2005, pp. 1061–1064.

[5] K. Laskowski and T. Schultz, "Modeling vocal interaction for segmentation in meeting recognition," in *Proceedings of the 4th international conference on Machine learning for multimodal interaction*, ser. MLMI'07. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 259–270.

[6] K. Boakye and A. Stolcke, "Improved speech activity detection using cross-channelfeatures for recognition of multiparty meetings," in *Proc. ICSLP*, 2006.

[7] 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.

[8] D. Neiberg and J. Gustafson, "Modeling conversational interaction using coupled markov chains," in *DiSS-LPSS Joint Workshop 2010*, Tokyo, Japan, 2010.

[9] ——, "The prosody of swedish conversational grunts," in *Interspeech, Special Session on Social Signals in Speech*, 2010.

[10] K. Laskowski, Q. Jin, and T. Schultz, "Crosscorrelation-based multispeaker speech activity detection," in *Proceedings of the 8th ISCA International Conference on Spoken Language Processing (INTERSPEECH2004)*, Jeju Island, South Korea, 2004.

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

[12] J. Brown, "Calculation of a constant Q spectral transform," *J Acoust Soc of Am*, vol. 89, no. 1, pp. 425–434, 1991.

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

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