ONLINE DETECTION OF VOCAL LISTENER RESPONSES WITH MAXIMUM LATENCY CONSTRAINTS

Daniel Neiberg[1], *Khiet P. Truong*[2]

[1]Department of Speech, Music and Hearing, Royal Institute of Technology (KTH), Sweden
[2] Human Media Interaction, University of Twente, The Netherlands
[1]neiberg@speech.kth.se, [2]k.p.truong@ewi.utwente.nl

ABSTRACT

When human listeners utter Listener Responses (e.g. back-channels or acknowledgments) such as 'yeah' and 'mmhmm', interlocutors commonly continue to speak or resume their speech even before the listener has finished his/her response. This type of speech interactivity results in frequent speech overlap which is common in human-human conversation. To allow for this type of speech interactivity to occur between humans and spoken dialog systems, which will result in more human-like continuous and smoother human-machine interaction, we propose an on-line classifier which can classify incoming speech as Listener Responses. We show that it is possible to detect vocal Listener Responses using maximum latency thresholds of 100-500 ms, thereby obtaining equal error rates ranging from 34% to 28% by using an energy based voice activity detector.

Index Terms— Speech processing, speech analysis

1. INTRODUCTION

In human-human conversations, overwhelmingly one speaks at a time. This does not exclude the presence of short but frequent segments of overlapped speech. These overlapped segments arise from the nature of human-human conversation, which is characterized by *mutual* interaction between each other in the discourse. This means that interlocutors *continuously in coordination* with one another show *attentive speaking* and *active listening* behavior ([1, 2]). By paying attention to the listener, by offering opportunities to the listener to give feedback, by acknowledging the feedback given by the listener etc., the speaker shows attentive speaking behavior. By giving back-channels and acknowledgments such as 'mm-hmm, yeah' or nodding, the listener shows that he/she is actively listening. Unfortunately, most current spoken dialog systems are not able to deal with this 'mutual continuously interactive behavior' and hence, assume a half duplex turn-taking interaction. One way to increase the levels of interactivity and continuity in human-machine interaction is to assess the user's incoming speech (i.e., vocal responses) as quickly as possible such that the system can react even before the user's speech turn has ended, just like humans do in in 50-56 % of all speaker shifts [3].

In the context of developing a route-giving Virtual Agent (VA), we aim for fast classification of vocal responses to allow for a higher level of continuous interactivity. In this context, the VA typically holds the speaker role and the human typically holds the listener role. When the listener's vocal response is merely signaling that he/she is paying attention and following the conversation, then the VA should continue to speak and possibly acknowledge this response. We refer to these types of vocal responses, that do not cause an interruption, or that are not perceived as competitive of the floor, as vocal *Listener Responses* (after Fujimoto [4]). However, when the vocal response is *not* a Listener Response but, for example, an attempt to take the floor, then the VA may decide to stop speaking and let the listener finish his/her vocal response.

In this work, we propose an on-line detector which is able to classify incoming speech as Listener Responses before the speech of the listener has ended. The development of this detector allows for the simulation of two important aspects: 1) detecting incoming speech in overlap as a Listener Response or not - since Listener Responses are per definition not competitive of the floor, the VA can continue to speak without triggering barge-in mechanisms, and 2) detecting incoming speech in silence as a Listener Response or not, allowing for projection-based turn-taking, i.e. the VA starts speaking before the user is finished.

This paper is structured as follows. The corpus used for training and testing is described in Section 2. Section 3 deals with the classification method and acoustic features. The classification experiments and results are presented in Section 4 and 5 respectively. Conclusions and future research are discussed in Section 6.

2. THE HCRC MAP TASK CORPUS

Here, we describe the speech data and annotations used in our classification experiments. In addition, we provide some corpus statistics related to the classification task.

2.1. Speech data and annotations

The HCRC Map Task Corpus [5] contains 128 dialogs. The task is for one subject to explain a route to another subject. We use the 32 dialogs which were recorded under a face-to-face condition. The two conversations labeled as `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 (ACK) and other dialog moves (NONACK). The precise definition of an Acknowledgment Move is found in [6], which closely resembles the term Listener Response and thus serves 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 is $\kappa = .83$, which can be considered good.

This research has partly been supported by the European Community's 7th Framework Programme (FP7/2007-2013) under grant agreement no. 231287 (SSPNet).

Based on the annotations provided, we segmented the corpus into *talkspurts* [7], defined as a minimum voice activity duration of $\alpha = 50$ ms separated by a minimum inter-pause of $\beta = 200$ ms. The resulting connected speech segments are referred to as talkspurts, where the latter threshold β 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 (ACK) and ended as a (NONACK). The occurrence of these latter talkspurts are considered to be negligible.

2.2. Descriptive statistics of Acknowledgment Moves

People use different words to show that they are acknowledging what the speaker has said. To have an idea of what types of words are used for ACK in the HCRC Map Task Corpus, we show the top 20 of most frequent ones in Table 1. We observe that most of the time, typical minimal feedback words are used such as 'right, okay, mmhmm, uh-huh, yeah' that account for about 58% of all ACK in the corpus which still leaves plenty of room for a large variety of other words.

word	%	word	%	word	%	word	%
right	28.2	*oh*	2.7	*got*	0.9	*a*	0.7
okay	14.9	*the*	2.3	*it*	0.9	*to*	0.7
mmhmm	5.3	*that's*	1.6	*you*	0.9	*fine*	0.6
uh-huh	5.3	*no*	1.5	*that*	0.8	*I've*	0.6
yeah	3.9	*I*	1.4	*mm*	0.7	other	26.1

Table 1. Top 20 most frequently occurring words in all Acknowledgment Moves found in the Map Task corpus, as percentages of a total of 9823 words.

Fig. 1 shows the durations of ACK and NONACK segments in the corpus. We observe that ACK segments are usually shorter than NONACK segments; hence, duration could be a useful feature for the classification task ACK vs. NONACK.

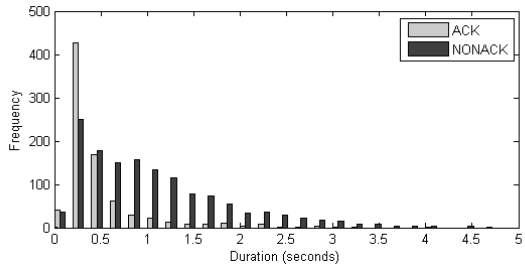

Fig. 1. Histogram of durations of ACK and NONACK segments in the HCRC Map Task corpus.

Listener Responses have also been frequently found in overlapped speech. Given a 10 ms frame discretization of the MapTask talkspurts, the following can be observed: given a speech frame in overlap, there is a 34.9% probability that it is an ACK. Given a speech frame in non-overlap, there is a 5.2% probability that it is an ACK. Thus, ACKs are relatively more common in overlap than in non-overlapped speech.

The between speaker interval, can be positive (gap) or negative (overlap). To compute the gaps and overlaps in ACK context, two cases of overlap are first considered. The first is ACK in complete overlap and the second is ACK in partial overlap. We computed the between speaker interval from the partial overlap case and the no overlap case which is shown in Figure 2. The figure shows that the mode, the actual peak of the distribution, is at 100 ms. By cutting the tails at 2000 ms in order to facilitate comparison with other work, the cumulative distribution is also computed. It shows that 31% of the resumptions are done in overlap, while 48% are done before 200 ms. This latter proportion is slightly lower than the 50-56 % which are reported for between speaker intervals without any particular context [3]. This means that the over-representation of ACK in overlap is mostly due to ACK interjection into complete overlap, and there is nothing special with resuming one's speech while the listener is still uttering a Listener Response. Nevertheless, this leads to the following design constraints: responses in complete overlap has to be classified in the range of a minimally perceivable pause, i.e. around 100-300 ms, while in silence a latency of 300-500 ms from the onset of speech is acceptable. This could be done using a speech recognizer running in incremental mode or by using a specialized detector. Since a speech recognizer will only detect lexical content, the special prosodic characteristics of vocal listener responses cannot be accounted for. In addition, automatic speech recognizers (ASR) frequently miss Listener Responses in spontaneous speech [8]. Hence, we develop an on-line detector specifically tuned to the discrimination between ACK and NONACK.

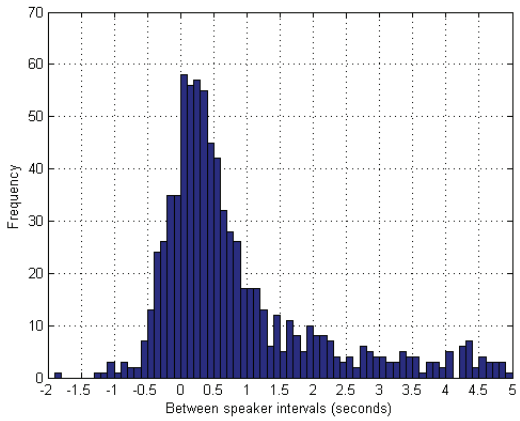

Fig. 2. The between speaker intervals between an ACK response and the interlocutors' continuation using bins of 100 ms.

3. METHOD AND FEATURES

We describe the maximum latency design and the acoustic features used for the development of the detector.

3.1. Maximum latency segmentation

Given the duration constraints derived in the previous Section, we propose a maximum latency implementation, which is illustrated in Figure 3. It is implemented as a voice activity detector which sends an end message after the talkspurt ends, or at a predefined duration threshold τ. If the duration reaches the threshold, it continues to work as a normal voice activity detector internally, otherwise it might trigger again. Note that the detector may trigger before the maximum latency is reached which happens when the talkspurt is

shorter than the threshold subtracted by the minimum inter-pause threshold β. For on-line detection, this maximum latency design was implemented in openSMILE [9].

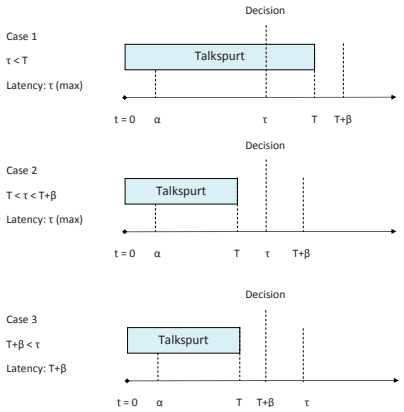

Fig. 3. Operation of a talkspurt segmenter which implements maximum latency decision. α is the minimum voice activity duration, β is the minimum inter-pause duration, T is the talkspurt duration and τ is the maximum latency threshold. In case 1 and 2, the decision is made at τ but in case 3 the decision is made at $T + \beta$

.

3.2. Feature trajectories as length-invariant Discrete Cosine Coefficients

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

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

where N is the segment length, x_n is the feature value at time n, X_k is the k'th coefficient and K is the number of coefficients. Here a fixed and small number of coefficients K is used regardless of the segment length.

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) 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. 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) or MFCCs (which has an additive channel bias).

3.3. Acoustic features

Based on the literature available on the acoustic characteristics of vocal listener responses, we selected a number of features for the ACK vs. NONACK task. It has been shown that Listener Responses have distinct prosodic characteristics - for example, they commonly have a rise in F0 and have distinct intensity contours [10]. Other important Listener Response characteristics are lexical content and short duration [11], and see Fig. 1. These findings have led to the choice of the following feature set: F0 ENVELOPES, INTENSITY, MFCCS, DURATION, and SPECTRAL FLUX (the L2-norm of energy normalized FTT-bin difference between two adjacent frames).

The acoustic features were extracted with a 10 ms frame rate with openSMILE [9], where the F0 ENVELOPES are computed by the Sub-harmonic sampling method with octave correction. The F0 ENVELOPES and SPECTRAL FLUX are transformed by the log-operation $\log_2(1+x)$. For F0 ENVELOPE the operation gives a perceptually relevant semitone scale, but is also makes a better fit for Gaussian or Radial Basis Function (RBF) modeling, which applies for SPECTRAL FLUX as well. The features were extracted up to the maximum latency threshold (in this case, we chose 100, 300, and 500 ms) and subsequently parametrized in the time dimension using length invariant DCT-coefficients 1-6, except SPECTRAL FLUX for which we use 0-5, unless anything else is specified. For Duration, the full talkspurt duration was used during training, while for testing, the duration up to the maximum latency threshold was used. The relevance of the 0'th coefficient is discussed in Sec 3.2.

4. CLASSIFICATION EXPERIMENTS

For the experiments, we divided 64 dialogs (the face-to-face part of the HCRC Map Task corpus) over training, development and evaluation sets. The training set consists of 32 dialogs (using the terminology of the corpus, these dialogs are comprised of so-called quads 1–4), the development set (quads 5–6) and evaluation set (quads 7–8) both hold 16 dialogs each. This results in the following distribution of talkspurts over the different sets as shown in Table 2. Note that these sets are speaker independent.

	TRAINING	DEV	EVAL
ACK	775	482	537
NONACK	1315	677	1138

Table 2. Number of talkspurts used for training, developing and testing the detector.

For classification, we used Support Vector Machines with a Radial Base kernel as implemented in the LibSVM package [12]. The SVM regularization parameters are optimized on the development set, and the best model is then used on the evaluation set.

We simulate two different development and evaluation settings: 'off-line' and 'on-line'. In the off-line setting, the detector is developed and tested using the talkspurt segmentation obtained from processing the manual transcriptions provided with the corpus. In the on-line setting, the detector is developed and tested using a talkspurt segmentation obtained with the thresholded energy based voice activity detector (VAD) available in openSMILE. This is done by clipping out each talkspurt, adding 2 seconds of silence before and after, and the resulting waveform is processed by the VAD to obtain a new talkspurt. By doing this, we test the sensitivity of the time-varying feature parametrization and simulate an on-line setting.

To ensure independence of priors and application, the performance is measured as Equal Error Rate (EER) calculated using the SVM decision values.

5. RESULTS

As expected, we observe in Table 3 that MFCCs followed by Spextral Flux and Intensity, on average, are the main contributors to the distinction between ACK and NONACK, while F0 is the weakest feature. Duration is a strong feature for the 500 ms case but not for the 100/300 ms case. We observe that the information loss caused by omitting the 0'th DCT coefficient for MFCCs, does not hurt performance.

	τ (ms)		
Feature(s)	100	300	500
F0 Envelopes	42.1	43.8	40.3
Intensity	41.1	37.3	35.5
MFCC with 0'th coeff.	31.3	25.9	24.7
MFCC without 0'th coeff.	31.9	25.7	24.5
Spectral flux	36.1	32.5	30.9
Duration	49.9	42.8	28.4
(A) Intensity, Spectral flux, MFCC without 0'th coeff.	28.8	25.7	23.4
(B) Intensity, Spectral flux, MFCC without 0'th coeff., Duration	29.3	25.9	23.3

Table 3. EER for ACK vs. NONACK classification for three maximum latency thresholds (in ms) tested on the development set in an 'off-line' setting.

Hence, we continue to experiment using MFCCs without the 0'th coefficient (i.e., feature sets A and B), and test the developed detectors on the evaluation set in an 'off-line' and 'on-line' simulation setting. The results of these experiments are shown in Table 4. We observe in Table 3 and 4 that, as expected, in general, a higher maximum latency threshold yields better performance. Furthermore, the addition of Duration (as used with maximum latency constraints) to other acoustic features does not hurt performance for lower maximum latency thresholds. Finally, we observe in Table 4 in except on one case, a relative drop of approximately 0-10% when the detector is tested under the on-line condition, i.e., where the classification is obtained by an energy based talkspurt segmentation.

	off-line, τ (ms)			on-line, τ (ms)		
Featureset	100	300	500	100	300	500
(A)	32.5	29.1	26.3	34.3	29.6	28.8
(B)	31.7	29.5	26.2	33.5	28.3	28.3

Table 4. EER for ACK vs. NONACK for three maximum latency thresholds (in milliseconds) - tested on the evaluation set for an 'off-line' and 'on-line' simulation setting.

6. CONCLUSIONS AND DISCUSSION

We have shown that is possible to detect Listener Responses based on only the first 100 ms, 300 ms or 500 ms of a vocal response (following maximum latency constraints) with a performance of EER 34% to 28% in an online setting. Given the finding that speakers commonly quickly resume their speech after roughly 0-400 ms following a vocal response and often even during the listener's vocal response (see Fig. 2), such a Listener Response detector is needed to allow for more continuous and smooth speech interaction between virtual agents and humans. Duration (at least in the 500 ms condition) and MFCCs appear to be strong discriminative features; the duration of an ACK is generally shorter than that of a NONACK (see Fig. 1). To integrate this classifier within a incremental dialog processing framework which is able to handle multiple ongoing plans, we suggest to run three classifiers in parallel at maximum latencies of 100 ms, 300 ms or 500 ms. This would let the dialog manager to prepare decisions at 100 ms, and then execute decisions at 300 ms or 500 ms. For future research, we plan to use this on-line Listener Response detector in an experimental setting in which interactive human-machine dialogs are held to evaluate several strategies on how to deal with Listener Responses. In addition, following a similar maximum latency design, future detectors may be investigated that analyse the vocal response's attitudinal and affective meaning.

7. REFERENCES

[1] H. H. Clark and S. E. Brennan, "Grounding in communication," in *Perspectives on Socially Shared Cognition*, L. B. Resnick, J. M. Levine, and S. D. Teasly, Eds. American Psychological Association, Washington, D.C., 1991.

[2] J. B. Bavelas, L. Coates, and T. Johnson, "Listeners as co-narrators," *Journal of Personality and Social Psychology*, vol. 79, no. 6, pp. 941–952, 2000.

[3] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, pp. 555–568, 2010.

[4] D. T. Fujimoto, "Listener Responses in Interaction: A Case for Abandoning the Term, Backchannel," *Journal of Osaka Jogakuin 2nd year College*, vol. 37, pp. 35–54, 2007.

[5] A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty-Sneddon, S. Garrod, S. Isard, J. C. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. Thompson, and R. Weinert, "The HCRC Map Task corpus," *Language and Speech*, vol. 34, pp. 351–366, 1991.

[6] J. C. Carletta, S. Isard, G. Doherty-Sneddon, A. Isard, J. C. Kowtko, and A. H. Anderson, "The reliability of a dialogue structure coding scheme," *Computational Linguistics*, vol. 23, no. 1, pp. 13–31, 1997.

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

[8] S. Goldwater, D. Jurafsky, and C. D. Manning, "Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, pp. 181–200, 2010.

[9] F. Eyben, M. Woellmer, and B. Schuller, "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proceedings of ACM Multimedia*, 2010, pp. 1459–1462.

[10] S. Benus, A. Gravano, and J. Hirschberg, "The prosody of backchannels in American English," in *Proceedings of the 16th International Congress of Phonetic Sciences (ICSPhS)*, 2007, pp. 1065–1068.

[11] J. Edlund, M. Heldner, S. Al Moubayed, A. Gravano, and J. Hirschberg, "Very short utterances in conversation," in *Proceedings of Fonetik*, 2010, pp. 11–16.

[12] C. C. Chang and C. J. Lin, *LIBSVM: a library for support vector machines*, 2001, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.