Prosodic Feature Generation for Back-channel Prediction

Thamar Solorio, Olac Fuentes, Nigel Ward, Yaffa Al Bayyari

Interactive Systems Group
Computer Science Department
University of Texas at El Paso
El Paso, TX 79968, U.S.A.
tsolorio,ofuentes,nigel,yalbayyar@utep.edu

Abstract

Using prosodic information to predict when back-channels are appropriate in spontaneous dialogs has become somewhat of a reference problem for automatic discovery techniques. Here we present experiments with two ideas: the use of features derived from randomly generated pitch and energy filters, and the use of instance-based learning, specifically the Locally Weighted Linear Regression (LWLR) algorithm. For the task of predicting possible back-channel locations in Iraqi Arabic [6], we obtain 22% precision and 51% recall, which is as good as that obtained using a laboriously developed and hand-tuned rule.

1. Introduction

A good indicator of an attentive listener in spontaneous spoken dialog is the production of back-channels. The speaker is motivated to keep talking when this sign of attention is provided by the audience. Despite the frequent and cross-cultural usage of back-channels in conversations, the problem has not been widely studied; for instance, even the most up-to-date machine translation systems, do not include translations of back-channels. Even more surprising is the fact that in second language instruction the topic is seldom mentioned. Thus, people learning a new language need to discover the correct usage of back-channels by themselves, maybe by engaging in conversations with native speakers of the langauge, or by listening to conversations by native speakers.

Our aim in this work is to develop a model that provides fast and accurate prediction of back-channel opportunities, that is, points in a dialog where a back-channel would seem appropriate to a native speaker of the language being spoken. Such a model has a wide range of potential applications, for example in the generation of more natural-sounding spoken dialog systems and as a component in an intelligent tutoring system for students of foreign languages.

In a previous work, based mainly on perceptual analysis targeted to Egyptian Arabic, Ward and Al Bayyari discovered that a good prosodic feature preceding a back-channel opportunity is a downward slope in the pitch [7]. In their experiments, this prosodic cue leads to a 13% precision and 43% recall. What distinguishes our work from [7] is that, while their work is based on careful human analysis of pitch profiles to discover useful prosodic features, ours is fully automated, relying on machine learning methods. We use a cascaded approach and generate features by applying filters to the energy and the pitch of the speaker's utterance. These features are then input to a machine learning classifier. Our method follows the idea that Viola and Jones have proposed for face and pedestrian detection on images [4, 5]. We have evaluated our method on Iraqi Arabic dialogs, reaching an F-measure of 31.27%.

The next section describes our approach to back-channel prediction. We present the feature generation process and then briefly describe the machine learning algorithm used in this work. Section 3 describes the data used in the experiments. The experimental results are discussed in Section 4. The paper presents conclusions and final remarks in Section 5.

2. The method

Back-channel prediction is a very difficult task, and is even more difficult if the information available is prosody only, without access to speech recognition and syntactic structure analysis. However, since for many languages and dialects of interest, few or no linguistic tools are available, we decided to restrict our method to prosodic features that can be computed quickly and are completely language independent. Below we present pseudo code for our method:

1. Find the set of candidates for back-channel opportunities by averaging the pitch over 0.5 seconds and removing all candidates with an average higher than the threshold.
2. For every back-channel candidate, generate a feature vector that includes the filter responses of the energy and speech signals in a fixed-width window (1 second for energy, 0.5 seconds for pitch) preceding the back-channel candidate timepoint.
3. Input these features to the LWLR classifier.
4. Convert the (real-valued) output of the LWLR to binary. For all the candidates, consider as positives those with predicted values higher than 0.5, while the rest are considered as negative predictions, that is, not candidates for a back-channel opportunity.

We have learned that a good number of back-channels occur at the end of the speaker's utterance, that is, after a pause in the dialog. There are several explanations that fit this observation: the speaker is trying to release the turn to his/her listener and since the listener does not have anything to say he/she just back-channels rejecting the invitation; the listener back-channels to the speaker at the end of the utterance to reassure attention or comprehension of what has just being said. Following this intuition, we develop a

simple criteria for detecting end of utterances, that is, the onset of a pause. The rule uses only presence/absence of pitch. If pitch is not detected over 50% or more of the previous 0.5 seconds, then the time point is considered to be an end-of-utterance. This is the first step in our algorithm; the goal here is to reduce the number of possible opportunities for a back-channel without losing most of the true suitable places for back-channels. The next step is then to generate features for the remaining candidates by applying filters to both the energy and the pitch.

2.1. Feature Generation

The idea of applying filters to the energy spectrum and pitch was inspired by the work of Viola and Jones [4, 5]. They proposed a clever idea, called integral image, that, by means of a preprocessing stage, allows to compute filter responses in constant time, instead of the naive implementation that requires a time that is linear in the size of the filter. In this work we are not using boolean filters as Viola and Jones do; we generate randomly five points for every filter and then perform a linear interpolation. Figure 1 shows examples of our filters.

We are using 10 random filters and we apply each of them to 1 second time frame of the energy, and to 0.5 seconds in the pitch, prior to the candidate timepoint (this is achieved by computing the dot product between a filter and the energy, or pitch). In addition to the random filters we are also using two predefined filters: a downward and an upward slope. These last two are incorporated to simulate what Ward and Al Bayyari found to be a useful cue in Egyptian Arabic. The features extracted in this step are then input to a machine learning algorithm, namely Locally Weighted Linear Regression (LWLR) [1]. LWLR has been used previously in other learning problems showing satisfactory prediction performance [2, 3]. The next subsection describes this algorithm in more detail.

2.2. Locally Weighted Linear Regression

LWLR is a form of instance-based learning. The basic idea is that, given a query point, we classify it by examining how similar points in the training data were classified. As similar points we use the 300 utterance ends in the training data which are most similar to the query point in terms of their prosody and energy patterns. Similarity is computed as a Euclidean distance over the 20 features (filter outputs) described above. The weight given to a data point is proportional to how similar it is to the query point.

More formally, given a query point $\mathbf{x_q}$, to predict its output parameters $\mathbf{y_q}$, we find the k examples in the training set that are closest to it, and assign to each of them a weight given by the inverse of its distance to the query point:

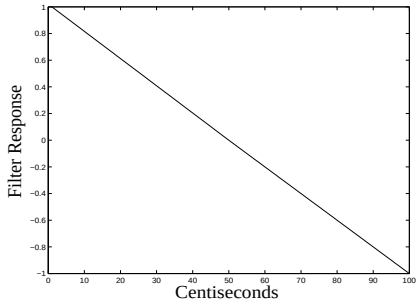
$$w_i = \frac{1}{|\mathbf{x_q} - \mathbf{x_i}|}$$

Let W, the weight matrix, be a diagonal matrix with entries $w_1, \ldots, w_n$. Let X be a matrix whose rows are the vectors $\mathbf{x_1}, \ldots, \mathbf{x_k}$, the input parameters of the examples in the training set that are closest to $\mathbf{x_q}$, with the addition of a "1" in the last column. Let Y be a matrix whose rows are the vectors $\mathbf{y_1}, \ldots, \mathbf{y_k}$, the output parameters of these examples. Then the weighted training data are given by

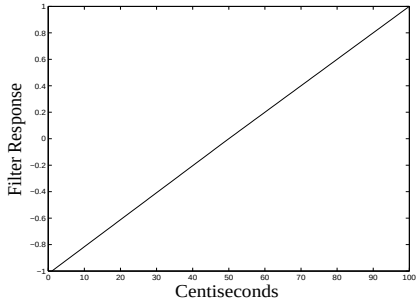
$$Z = WX$$

and the weighted target function is

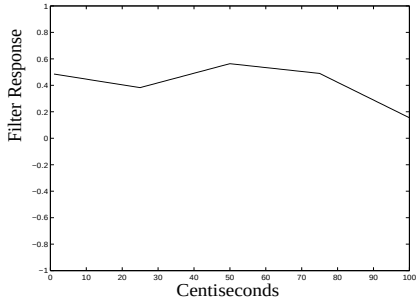
$$V = WY$$



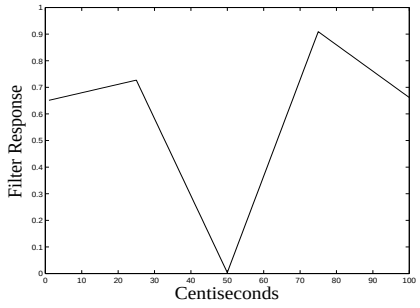
a) Downward slop filter



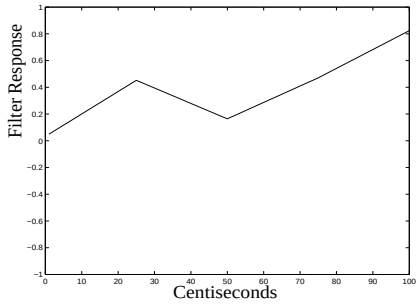
b) Upward slope filter



c) Random filter



d) Random filter



e) Random filter

Figure 1: Examples of the filters used in the feature extraction. Figure (a) shows the downward filter, (b) shows the upward filter and figures (c) to (e) are examples of random filters

Then we use the estimator for the target function

$$\mathbf{y_q} = \mathbf{x_q}^T (Z^T Z)^{-1} Z^T V$$

This algorithm outputs a real value prediction. In the learning task of predicting back-channel opportunities, this value represents the predicted likelihood that this is an appropriate time for a back-channel. To achieve a discrete valued output we apply a threshold to the output of the algorithm as follows: if $\mathbf{y_q} > 0.5$ then $\mathbf{p_q} = 1$, otherwise $\mathbf{p_q} = 0$. We evaluate the F_1 measure based on $\mathbf{p_q}$.

For more information about this algorithm we refer the reader to [1].

3. Data

We used 110 minutes of unrestricted conversations between two native speakers of Iraqi Arabic. The corpus includes face to face dialogs of twelve different male speakers [9]. The dialogs were analyzed and labeled by a native Arabic speaker, the fourth author of this paper. She labeled the places where the listener produced a back-channel, but she also added back-channel labels to places where, according to the speaker's utterance, a back-channel would have been appropriate. We decided to match to this expanded label set as our evaluation criterion because we want our system to learn to predict all times when it is possible for a listener to produce a back-channel, not just the times when the specific speakers in our corpus happened to do so.

The data was divided into two sets, training set and testing set. We used the training set to determine the thresholds, as well as for training the learning algorithm. The test data was only used in the final stage of our evaluation, to measure the performance of our system. For each conversation, 85% of each dialog was used for training and the remaining 15% was used for testing.

4. Experimental Results

In this section we discuss the evaluation of our proposed method. The prediction measures used in our evaluation are the standard ones: precision (p), recall (r), and F-measure (F_1). The definition of each measure is shown below.

$$p = \frac{TP}{TP + FP} \tag{1}$$

$$r = \frac{TP}{TP + FN} \tag{2}$$

where TP are true positives, FP are the number of false positives (errors of commission) and FN are false negatives (errors of omission) for the class.

$$F_1 = \frac{2 \times p \times r}{p + r} \tag{3}$$

In Table 1 we present the results of using our method. The results show that a very simple rule that tries to detect the end of a utterance achieves an impressive F-measure of 18.13%. It should be noted that recall for this rule reaches 77%, this shows that only by looking at pauses from the speaker, we can match over three fourths of the back-channel opportunities. Precision however, is not that good, as the rule achieves only a 10.27%. This first step yields very satisfactory results for our purposes since we are greatly reducing the number of possible candidates without loosing a considerable amount of true positives. But as the table shows,

Table 1: Comparison of prediction results of using the end of utterance rule and using the LWLR algorithm for the test fragments.

Method	Precision	Recall	F-measure
End of utterance rule	10.27	77.40	18.13
LWLR with filter features	22.55	51.00	31.27

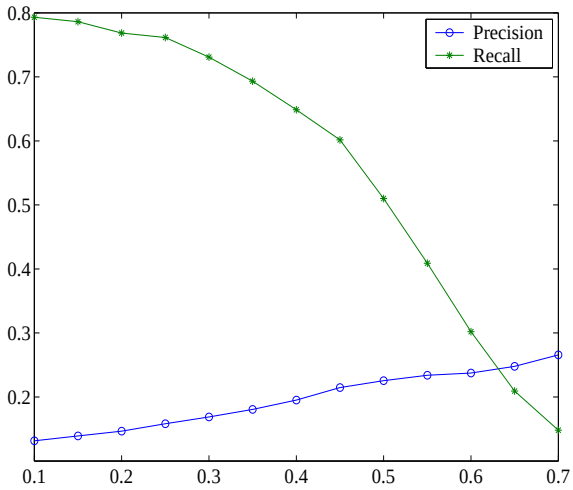


Figure 2: Values for precision and recall for different prediction thresholds in the LWLR algorithm

the filter features are useful and yield a large increase in precision, and a higher f-measure than that of just finding the pauses, although recall suffers a small decrease. Applying the LWLR algorithm we reach an F-measure of 31%. This turns out to be almost exactly the same as that obtained, on this same problem, by the rule reported in Ward and Al Bayyari [6]. As the current model has many more free parameters, this level of performance may not be surprising, however it is noteworthy that this performance was obtained by an almost entirely automatic method.

Overall, this performance is comparable with what we have found for English, where we obtained an F-measure of 26% for the slightly harder problem of predicting only actual back-channels, and lower than that obtained for Japanese, 42% [8]. In terms of the degree to which prosodic cues from the speaker determine when the listener can back-channel, Arabic is more similar to English than to Japanese.

In Figure 2 we show how precision and recall behave for our method when different thresholds are used in the predictions of the LWLR algorithm. As it is expected, recall increases as the threshold decreases, and precision shows the opposite behavior, it increases with larger values of the threshold. Figure 3 shows how the F-measure changes as a function of this threshold. It can be seen that the optimal value in the precision-recall trade-off, with an F-measure value of over 0.3, is reached when we use a threshold of 0.5, although the method is not particularly sensitive to the threshold selection, as values in the 0.3-0.6 range yield similar results.

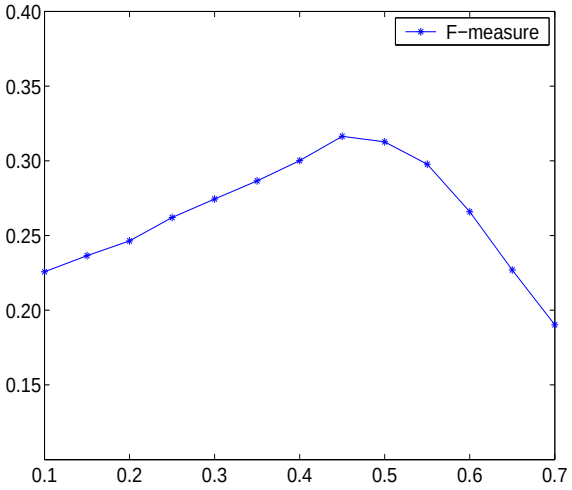


Figure 3: F-measure values using different prediction thresholds in the LWLR algorithm

5. Conclusions

We have presented a method for predicting back-channel opportunities based on machine learning. Our method draws on ideas from a machine vision approach to face detection and previous results on an analysis of Egyptian Arabic. The results presented show that our method gives results as good as a hand-crafted rule. We used a Linear Regression algorithm trained on prosodic features that can be extracted in constant time with a short preprocessing stage.

One potential drawback of methods such us the one presented here is that, even though they have good prediction results, extracting intuitive knowledge that can be exploited in the classroom is not as easy as it is for methods that use human-extracted features. But there are alternatives to solve this lack of readability, for instance, we can use methods for attribute selection to identify the most useful filters. Another strategy can be to combine this method with approaches like the one of Ward and Al Bayyari, that way we can take advantage of the precision of our method and still have a sort of rule that we can teach students learning how to back-channel in a foreign language.

Work in progress includes analyzing the filter features. We believe that a more careful inspection to the filter features can reveal prosodic cues that would be hard to discover by perceptual analysis. We are also working on developing methods for a heuristic exploration of the very large space of potential features. This way we are trying to find sets of features that maximize a combination of precision, recall, and human readability.

6. Acknowledgements

We would like to thank DARPA and NSF grant No. 0415150 for partially supporting this work. We would also like to thank David Novick for useful discussions and the referees of this paper for their helpful insights.

7. References

[1] Christopher G. Atkeson, Andrew W. Moore, and Stefan Schaal. Locally weighted learning. *Artificial Intelligence Review*, 11:11–73, 1997.

[2] Olac Fuentes and Thamar Solorio. An optimization algorithm based on active and instance-based learning. In R. Monroy, G. Arroyo-Figueroa, L. E. Sucar, and H. Sossa, editors, *MICAI 2004: Advances in Artificial Intelligence, Third Mexican International Conference on Artificial Intelligence*, Lecture Notes in Artificial Intelligence 2972, pages 242–251, Mexico City, Mexico, April 2004. Springer.

[3] Thamar Solorio, Olac Fuentes, Roberto Terlevich, and Elena Terlevich. An active instance-based machine learning method for stellar population studies. *Monthly Notices of the Royal Astronomical Society*, 363(2), October 2005.

[4] Paul Viola and Michael Jones. Rapid object detection using a boosting cascade of simple features. *IEEE Computer Society Conferrence on Computer Vision and Pattern Recognition*, 1:511–518, 2001.

[5] Paul Viola and Michael Jones. Robust real-time object detection. In *Second International Workshop on Statistical and Computational Theories of Vision – Modeling, Learning, Computing and Sampling*, Vancouver, Canada, July 2001.

[6] Nigel Ward and Yaffa Al Bayyari. A case study in the identification of prosodic cues to turn-taking: Backchanneling in arabic. In *Interspeech 2006*.

[7] Nigel Ward and Yaffa Al Bayyari. A prosodic feature that invites back-channels in Egyptian Arabic. *Perspectives on Arabic Linguistics XX*, 2006.

[8] Nigel Ward and Wataru Tsukahara. Prosodic features which cue back-channel feedback in English and Japanese. *Journal of Pragmatics*, 32:1177–1207, 2000.

[9] Nigel G. Ward, David G. Novick, and Salamah I. Salamah. The UTEP corpus of Iraqi Arabic. Technical Report UTEP-CS-06-02, University of Texas at El Paso, El Paso, TX, 2006.