Using Model Trees for Evaluating Dialog Error Conditions Based on Acoustic Information

Abe Kazemzadeh
University of Southern California
Los Angeles, CA
kazemzad@usc.edu

Sungbok Lee
University of Southern California
Los Angeles, CA
sungbokl@usc.edu

Shrikanth Narayanan
University of Southern California
Los Angeles, CA
shri@sipi.usc.edu

ABSTRACT

This paper examines the use of model trees for evaluating user utterances for response to system error in dialogs from the Communicator 2000 corpus. The features used by the model trees are limited to those which can be automatically obtained through acoustic measurements. These features are derived from pitch and energy measurements. The curve of the model tree output versus dialog turn is interpreted to be a measure of the level of user activation in the dialog. We test the premise that user response to error at the utterance level is related to user satisfaction at the dialog level. Several different evaluation tasks are investigated: on an utterance level we applied the model tree output to detecting response to error and on the dialog level we analyzed the relation of model tree output to estimating user satisfaction. For the former, we achieve 65% precision and 63% recall and for the latter our predictions show significant .48 correlation with user surveys.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human information processing; H.3.1 [**Content Analysis and Indexing**]: Indexing methods; H.5.2 [**User Interfaces**]: Voice I/O,Natural language,Evaluation/methodology,User-centered design

General Terms

Algorithms, Measurement, Human Factors

Keywords

Evaluation of human-computer dialog systems, Paralinguistic feedback, User response to error

1. INTRODUCTION

The evaluation of Human-Centered Computing (HCC) applications is an important aspect of system design and improvement. This paper focuses on developing an objective, human-centered evaluation metric for human-computer dialog systems. This metric is claimed to be objective in that it uses models trained on speech data from annotated dialog data. The metric can also be termed human-centered because it aims to be used online as a measure of activation in the user's voice during the flow of dialog. Here, this is tested in human-computer interaction in dialog systems, but may also be used for content analysis of human-human dialogs.

1.1 Motivation

From a system development perspective, dialog system evaluation is important in deciding how to optimize system strategies [11, 23]. Deciding that one configuration is better than another is an important, open question. From a commercial perspective, dialog system evaluation can be important in terms of maintaining adequate levels of user satisfaction.

In much of the current research, evaluation has used synthesis of performance measures of separate system components to extrapolate a total evaluation measure, often fit to or tested against user survey results. [25, 21, 9]. For example, automatic speech recognition (ASR) performance measures, task completion measures, number of turns, and user dialog acts may be all incorporated to produce a global evaluation metric of the dialog. However, many of these measures require human annotation, subjective measures, and a task-oriented model of dialog structure. Also, these evaluation methods must be applied on whole dialogs or complete subdialog units. These may be considered human-centered evaluation methods, but if the metrics are based on models that approximate user satisfaction, they cannot be said to be objective. Also, methods that require the analysis of a whole dialog, will not be able to be used in a live system, thus limiting online feedback capabilities, one desirable feature of human-centered computing.

By using only acoustic information, we aim to circumvent some of these issues. Evaluating each user utterance acoustically allows evaluation to take place over arbitrary parts of the dialog, provided there be a neutral utterance for normalization purposes (in this study we use the first utterance of the dialog). Moreover, since errors have been shown to correlate with user hyperarticulation [20], utterance level evaluation can be seen as a means of error detection.

Since the prevalence of errors in a system and the methods a system uses to deal with errors are strongly correlated with overall user satisfaction [9] (or rather, dissatisfaction), we frame our evaluation task in terms of users' response to error

HCM'06, October 27, 2006, Santa Barbara, California, USA.

conditions. In this view, detecting user response to error can be seen as a local version of dialog system evaluation.

In addition, limiting the factors for dialog evaluation to acoustic measurements allows for a means of evaluating the dialog even when recognition may be faulty. In fact, it may even be applied to human-human dialogs. However, our approach, which uses model trees, differs from most methods of detecting user corrections in that the output of our method is not a decision, but real-valued numbers. When plotted throughout the dialog, the resulting curve is intended to represent the level of user activation, which can be used for identifying dialog hotspots, problematic dialog turns where the user is angry or frustrated.

1.2 Background

This work builds on several different areas of dialog and speech research, which we will cover briefly in this section. These areas are dialog error detection, dialog evaluation, and emotion recognition. After first introducing model trees, we will cover the relevant background in these areas. We will conclude with a synopsis of related studies conducted in our group.

1.2.1 Model trees

Model trees [17] are similar to decision trees and regression trees [1] in their tree structure, but differ in their leaf nodes. While the leaves of decision trees contain a classification rule and the leaves of a regression tree have a class mean, each leaf node of model trees contain a linear regression model that applies to the instances that reach it. Therefore, instead of returning a class prediction, model trees, like regression trees, are able to return a real-valued number. However, because regression trees have a fixed mean for each leaf node, the effect is discretization, which can be seen simply as the result of a mapping from class categorization [27].

1.2.2 Error detection

Error detection in human-computer dialog research concerns locating errors that can occur for a number of reasons (e.g., faulty speech recognition, user mistakes, incomplete world knowledge). The detection task may use lexical features, acoustic features, dialog history, and confidence scores, among others, and the errors may be identified in the turn where they occur or from subsequent user corrections [3, 20, 4, 5]. The end goal for this research in dialog systems is to trigger error recovery mechanisms [18].

In this paper, we will look at detecting user response to error by using a measure of user speech activation that has been predicted using model trees based on the users' speech acoustics.

1.2.3 Dialog evaluation

Human-computer dialog evaluation aims to relate the many parameters and configurations dialog systems may have to a metric that describes the quality of the interaction [9]. This evaluation may involve subjective (e.g. user surveys) and/or objective measurements (e.g. counts of turns, repetitions, word error rate). The end goal may be to optimize the system's performance in the case of dialog design [11, 23], or to assess the quality of existing dialog systems in the case of quality assurance.

One major milestone in the development of human-computer dialog evaluation research is the PARADISE framework [24]. This approach fused task accomplishment measures with dialog interaction costs using multi-variable linear regression where the component weights are estimated with respect to survey results.

In this paper we will look at evaluating the Communicator dialogs by testing different ways to use the model tree output as global measures of the dialog and observing the correlations of these metrics with the results of user surveys.

1.2.4 Emotion recognition

Emotion recognition in speech research aims to classify speech into an arbitrary number of emotional classes. Methods used for this task include studying the prosodic features of pitch and energy, using acoustic modeling with hidden Markov models, and analyzing dialog acts and lexical choice [8, 14, 12, 2]. One of the applications of this research in detecting problematic dialog turns. This detection of angry or frustrated emotions in dialog could be used to modify system behavior or to route calls to a human operator [8].

The model tree approach to predict user speech activation is similar to emotion recognition that has uses a binary classification (negative and non-negative). The difference is that the activation level is real-valued measure and not a hard classification.

1.2.5 Our relevant prior work

In the SAIL lab, we have undertaken a series of studies examining the Communicator corpus. In our first study [19] we developed an account of user behavior under error conditions based on dialog strategies of the user and system. By defining error regions within the dialogs (similar to "error chains" in the research of [20]) and manually annotating them, we succeeded in formulating general trends that can determine the successfulness of user and system attempts to correct errors in the dialogs.

In our second study [5], we investigated the acoustic features of user response to error by examining a variety of measurements similar to the ones used in this study. We found salient acoustic features for error responses in general as well as features that are correlated with specific types of error responses and responses that occur in different contexts within error regions.

In addition to this line of research, we have studied the role of emotions in spoken dialog [6, 7, 13].

1.3 Hypothesis

In this paper we hypothesize that by training a model tree to detect user responses to error, we will derive a measure of utterance activation, i.e. "hot spots" [30], on the dialog turn level, as well as an indicator of user satisfaction on a global level.

At a local level, we tested this hypothesis by using the model tree output to detect user response to error. At a global level, we calculated the correlation of the model tree output with different measures of user satisfaction from survey data.

2. DATABASE

2.1 Corpus

We used audio data and tagged transcriptions from the Communicator Travel Planning Systems [22] June 2000 recordings. Each dialog consists of a number of exchanges between a computer travel agent and a human. In this corpus, 85 experimental subjects interacted with 9 different travel agent systems. We worked with a subset of 141 dialogs and the average length of these dialogs was 18 exchanges.

Each transcribed exchange consists of a system utterance, a user utterance (manually transcribed from recordings), and what the ASR system heard and provided as input to the dialog system. Along with each user turn there is a corresponding audio file with the user's recorded utterance, which is recorded in NIST sound file format, encoded in either pcm or mu-law, at a sampling rate of 8000 Hz. These recordings were aligned to the transcriptions and tagged information, with manual corrections made to adjust occasional mismatches that resulted from empty sound files. In all, we analyzed 2586 utterances. The data and the COMMUNICATOR collection procedure are described in detail in [28, 10, 16].

2.2 Tagging

The dialogs in the corpus were manually tagged by two annotators who showed 87% inter-annotator agreement. The tagging scheme was devised with the intention to highlight the ways errors are recognized and dealt with by the user [19].

Briefly, the tag set included (1) SYSTEM tags: explicit confirmation, implicit confirmation, help, system repeat, reject, non sequitur, (2) USER tags: repeat, rephrase, contradict, frustration, change request, start over, scratch, ask, acquiesce, hang up, (3) TASK tags: error, back on track, and task success.

The detailed specifications for the tagging task and examples are found at http://sail.usc.edu/dialog/model_tags.html and also in [13].

An important note about our annotation is that the tagging was done with respect to the text transcripts of the dialogs and not the audio. This is important for tags like frustration, where the tag is not based on the perceived quality of the speech but rather on information from the text transcriptions. This approach has been used in research on detecting corrections [20], while emotional speech research has generally focused on listening tasks [6] or actors [14, 12].

3. METHODOLOGY

The outline of the methodology of this study is as follows: first, features were extracted from speech; second, we trained the model trees using the acoustic features as predictors and tagged data as targets; third, we applied the model trees to evaluate the dialogs on the utterance and global levels; finally, we interpreted the results and tested possible improvements from adding emotion valence in the evaluations.

The former three steps will be discussed in sections 3.1, 3.2, and 3.3 while the results will be presented and discussed in sections 4 and 5.

3.1 Feature Extraction

Figure 1 shows the different stages of feature extraction from the recorded speech utterance for use in model tree training and for our evaluation task. The pitch extraction was done using the Snack sound toolkit [15], which has an implementation of the Entropic ESPS xwaves get_f0 utility. This program returns a stream of four fields for every 10 ms interval: F0, probability voiced, RMS energy, and autocorrelation peak. The F0 and energy values were smoothed using a five point median filter. Also, adjacent voiced regions were merged if separated by less than two unvoiced frames and unvoiced regions were merged if separated by less than three voiced frames. Halving and doubling were corrected by checking for jumps of twice or half the surrounding pitch track.

The features that were extracted from the smoothed pitch track are the following: F0 minimum, maximum, and range; RMS energy minimum, maximum, and range; the minimum and maximum of the F0 ranges of the individual voiced segments of the utterance (which would be equal to the global F0 range if there is only one voiced segment in the utterance); the minimum and maximum F0 velocity (which we defined as the F0 range divided by the length of the voiced segment); the minimum and maximum length of voiced segments (which would be equal if there were only one voiced segment); the minimum and maximum length of unvoiced segments between voiced segments (which is an indicator of pauses; this was set to zero if there was only one voiced segment); and the number of voiced segments (which is an indicator of the length of the utterance).

The F0 and RMS energy features were normalized by dividing the features of each utterance in the dialog by those of the first utterance of the dialog. This was done to minimize inter-speaker variation. The length features were not normalized for several reasons. These features had the possibility of being equal to zero, so the normalization could involve division by zero. Also, the first utterances tended to be longer than the other utterances, except in some of the dialogs where the users gave a task ID. Thus these differences would have skewed the normalization.

3.2 Model Tree Training

To construct the model trees we used the WEKA machine learning toolkit [29]. Every dialog utterance was a training instance. The target of a training instance was set to one if there was a response to an error, or zero if not. Thus, the training of the model tree is very similar to the training of a decision tree to detect error responses except that a real-valued number is output instead of a decision to classify the utterance as error response or not. The intuition of this is that not every error response is hyper-articulated, although there is a tendency for users to do so [20]. We hoped the model tree would learn this. If the model trees could learn this tendency, the resulting plot could be thought of as a topology or landscape of the dialog activation, and elevated regions, as dialog hot spots.

We also balanced the training set so that there were an equal number of instances of error responses and non-error responses. Since there were approximately twice as many neutral utterances as error response utterances, we multiplied the instances of error responses by two so that they were proportionate to the non-error responses. This step

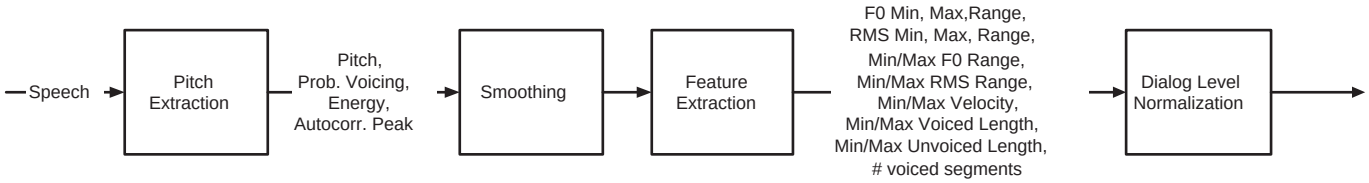

Figure 1: This shows the process of feature extraction

improved our results and shows the importance of trying several different training schemes.

3.3 Evaluation

When the model tree is applied to a dialog, the output is in the form of an abstract measure of activation versus time (measured in turns). To see whether this was actually useful to evaluate dialogs, we tried several different tests on both the utterance level and dialog level.

On the utterance level, we used the model tree output to determine whether the utterance was a response to error or not by evaluating the precision and recall using different thresholds. This showed the tradeoffs between strictness and laxness tuning the error response detection threshold.

On the dialog level, we tested the use of the model tree output to evaluate the dialog with respect to user satisfaction. The simplest way was to sum up the model tree outputs for each utterance in the dialog and to compute the correlation with the user survey data. Other possibilities were to sum model outputs that exceeded a certain threshold or to average the summed output by the number of user turns (thereby correcting for the factor of dialog length). In addition, there were several user satisfaction metrics from the surveys, so we considered which metric our evaluation correlated to the most.

4. RESULTS

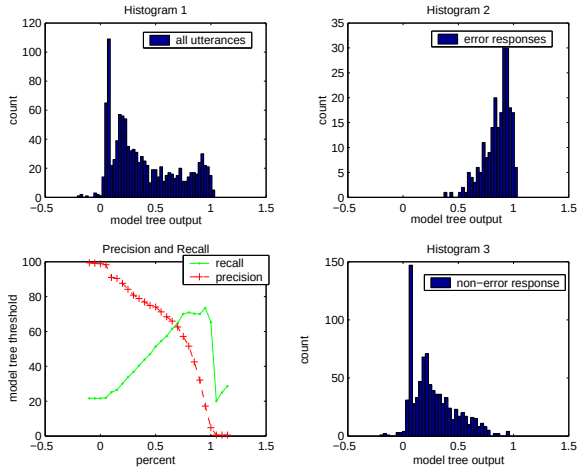

Figure 2: This shows histograms of model tree output and how the threshold function relates to precision and recall

Using model trees to detect user corrections gives 65% precision and 63% recall at the optimal threshold of .7, as can be seen in figure 2. The threshold can be seen as separating the two humps of the histogram distribution and adjusting the threshold will result in either a preference toward precision or recall.

The correlation between user satisfaction and model tree output has a negative relationship, because when the model tree output is high the user is predicted to be dealing with dialog error. However since in the survey one represented most satisfied and five as least satisfied, these figures have been presented with positive correlation figures. The correlation between the sum of the model tree output for a dialog and all user satisfaction metrics combined was low, .282, but still comparable to the correlation when using the tagged error response data directly instead of the model tree, .345. However, when the sum of the outputs is averaged by the length of the dialog, the correlation is stronger (.401). When the model tree outputs that are in the lower two thirds of the output range ($< .66$) are thresholded out, the correlation of the model tree with all user satisfaction rises to .480. These correlations are significant at the .01 level.

The specific user satisfaction measure that had the highest correlation with the model tree output (.498) was the prompt, "The system worked the way I expected". This user satisfaction metric also had the highest correlation with the tagged information about user response to error (.412). The correlations with the other survey prompts can be seen in figure 3.

5. DISCUSSION

The results show that our approach of using model trees to plot user activation levels using only acoustic features has reasonable performance results when used to detect user error response at an utterance level. One benefit of this approach is that the thresholding provides a fuzzy classification as opposed to the hard classification of decision trees. The sensitivity of the error response detection can be modified to improve either the precision or recall.

For evaluating the dialog on a global level, our approach gave modest results when evaluated with respect to user survey results, at best predicting approximately 23% of the variance (Pearson's r^2). For comparison [26] reported 37% of the variation of the user satisfaction could be predicted by combining task completion judgment, sentence accuracy, on task dialog duration, and system turn duration. Though the current study's results are not as good, only acoustic features were used, showing that speech acoustics alone can give important information about user satisfaction. One other interesting difference between the results of [26, 5] and this study is that our predictions improved by averaging out the effects of dialog length, while in [26, 5], the number of turns on task was one of the useful predictors of user satisfaction.

Survey Question	Correlation with tagged data	Correlation with Model tree output
It was easy to get the info I wanted	.412	.389
I found it easy to understand what the system said	.035	.092
I knew what I could say or do at each point in the dialog	.269	.311
The system worked the way I expected it to	.365	.498
I would like to use the system regularly	.332	.409

Figure 3: This table shows the user survey correlation with user response to error (counts from tag information) and model tree output (thresholded and normalized for dialog length, as described above.)

6. CONCLUSION

In this paper we presented a method for using model trees to evaluate human computer dialog systems. By using this method to carry out evaluation tasks at the utterance and dialog level, we tested the premise that by acoustically detecting user response to error, one may thereby derive a measure of user satisfaction. To represent this user state, we made use of the notions of dialog activation landscape and hotspots.

One avenue for further work on this approach is in the training of the model tree. In our research we trained the model tree using the tagged error responses as targets equal to one and non-error responses to be targets equal to zero. Also, training with balanced instances of error responses and non-error response gave us better results. Using different and more complex training schemes and weighting of the targets is one way that might further improve the results.

Another possible direction for further research would be to consider user specific deviations from the general tendency shown in the correlation between the model tree output and user satisfaction levels. This may help explain the large amount of variation shown in the correlation.

Furthermore, the notion of dialog activation landscape begs the question of a more dynamic and higher resolution measure to detect emphasis within an utterance. This could be done at the word, syllable, segment, or even the feature level.

Another avenue for future research that we will pursue is testing the model tree approach on different corpora. Currently we are working on studying the dialog of a recorded radio performance of Arthur Miller's *All My Sons*, which has many examples of emotionally charged speech. Another example of possible corpora to use could be recorded meeting transactions, in which identifying "hotspots" could help in identifying important topics.

7. ACKNOWLEDGMENTS

This section is optional; it is a location for you to acknowledge grants, funding, editing assistance and what have you. In the present case, for example, the authors would like to thank Gerald Murray of ACM for his help in codifying this *Author's Guide* and the **.cls** and **.tex** files that it describes.

8. REFERENCES

[1] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Sone. *Classification and Regression Trees*. Chapman and Hall, Boca Raton, 1984.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor. Emotion recognition in human-computer interaction. *IEEE Sig. Proc. Mag., vol. 18(1), pp. 3280, Jan*, 2001.

[3] J. Hirschberg, D. Litman, and M. Swerts. Prosodic cues to recognition errors. In *ASRU*, 1999.

[4] J. Hirschberg, D. Litman, and M. Swerts. Identifying user corrections automatically in spoken dialogue systems. In *NAACL*, 2001.

[5] A. Kazemzadeh, S. Lee, and S. Narayanan. Acoustic correlates of user response to errors in human-computer dialogues. In *ASRU*, St. Thomas, U.S. Virgin Islands, 2003.

[6] C. Lee and S. Narayanan. Towards detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 2004. (in press).

[7] C. M. Lee, S. Narayanan, and R. Pieraccini. Classifying emotions in human-machine spoken dialogs. In *ICME*, Lusanne, Switzerland, 2002.

[8] C. M. Lee, S. Narayanan, and R. Pieraccini. Combining acoustic and language information for emotion recognition. In *ICSLP*, Denver, CO, 2002.

[9] S. Lee, E. Ammicht, E. Fosler-Lussier, J. Kuo, and A. Potamianos. Spoken dialogue evaluation for the bell labs communicator system. In *HLT*, San Diego, California, 2002.

[10] E. Levin, S. Narayanan, R. Pieraccini, K. Biatov, E. Bocchieri, G. di Fabbrizio, W. Eckert, S. Lee, A. Pokrovsky, M. Rahim, P. Ruscitti, and M. Walker. The at&t-darpa communicator mixed initiative spoken dialog system. In *Proc. of ICSLP*, Beijing, 2000.

[11] D. J. Litman, M. S. Kearns, S. Singh, and M. Walker. Automatic optimization of dialog management. In *COLING*, Saarbruken, Germany, 2000.

[12] I. Murray and J. Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *J. Acoust. Soc. Am., 93 (2), Feb*, 1993.

[13] S. Narayanan. Toward modeling user behavior in human-machine interactions: Effects of errors and emotions. In *ISLE Workshop on Multimodal Dialog Tagging*, Edinburgh, UK, 2002.

[14] J. Noad, S. Whiteside, and P. Green. A macroscopic analysis of an emotional speech corpus. In *Eurospeech*, Rhodes, Greece, 1997.

[15] R. I. of Technology in Stockholm. The snack sound toolkit. http://www.speech.kth.se/snack/. Viewed 6/26/2005.

[16] A. Potamianos, E. Ammicht, and H.-K. J. Kuo. Dialogue management in the bell labs communicator system. In *ICSLP*, Beijing, China, 2000.

[17] J. R. Quinlan. Learning with continuous classes. In *Proc. Fifth Australian Joint Conference on Artificial*

Inteligence, Hobart, Tasmania, 1992. World Scientific, Singapore.

[18] H. Sagawa, T. Mitamura, and E. Nyberg. Correction grammars for error handling in a speech dialog system. In *HLT/NAACL*, Boston, 2004.

[19] J. Shin, S. Narayanan, L. Gerber, A. Kazemzadeh, and D. Byrd. Analysis of user behavior under error conditions in spoken dialogues. In *ICSLP*, Denver, 2002.

[20] M. Swerts, D. Litman, and J. Hirshberg. Corrections in spoken dialogue systems. In *ICLSP*, Beijing, 2000.

[21] Walker, R. Passonneau, J. Aberdeen, J. Boland, E. Bratt, J. Garofolo, L. Hirschman, A. Le, S. Lee, S. Narayanan, K. Papineni, B. Pellom, J. Polifroni, A. Potamianos, P. Prabhu, A. Rudnicky, G. Sanders, S. Seneff, D. Stallard, and S. Whittaker. Cross-site evaluation in darpa communicator: The june 2000 data collection. *Computer Speech and Language*, 2002.

[22] M. Walker, J. Aberdeen, J. Boland, E. Bratt, J. Garofolo, L. Hirschman, A. Le, S. Lee, S. Narayanan, K. Papineni, B. Pellom, J. Polifroni, A. Potamianos, P. Prabhu, A. Rudnicky, G. Sanders, S. Seneff, D. Stallard, and S. Whittaker. Darpa communicator dialog travel planning systems: The june 2000 data collection. In *Proc. Eurospeech*, Aalborg, Sweden, 2001.

[23] M. Walker, J. Fromer, and S. Narayanan. Learning optimal dialogue strategies: A case study of a spoken dialogue agent for email. In *ACL/COLING*, Hamburg, 1998.

[24] M. Walker, D. Litman, C. Kamm, and A. Abella. Paradise: A framework for evaluating spoken dialogue agents. In *Association of Computational Linguistics , ACL 97*, Madrid, 1997.

[25] M. Walker, D. J. Litman, C. A. Kamm, and A. Abella. Evaluating spoken dialogue agents with paradise: Two case studies. *Computer Speech and Language, 12(3)*, 1998.

[26] M. A. Walker, R. Passonneau, and J. E. Boland. Quantitative and qualitative evaluation of darpa communicator spoken dialogue systems. In *ACL*, Toulouse, France, 2001.

[27] Y. Wang and I. H. Witten. Inducing model trees for continuous classes. In *9th European Conference on Machine Learning*, Prague, April 1997.

[28] W. Ward and B. Pellom. The cu communicator system. In *IEEE ASRU*, Keystone, CO, 1999.

[29] I. H. Witten and E. Frank. *Data Mining*. Morgan Kaufmann, San Francisco, 2000.

[30] B. Wrede and E. Shriberg. The relationship between dialog acts and hot spots in meetings. In *ASRU*, St. Thomas, US Virgin Islands, 2003.