NOVEL ACOUSTIC FEATURES FOR AUTOMATIC DIALOG-ACT TAGGING

Harish Arsikere *Arunasish Sen* *Prathosh A. P.* *Vivek Tyagi*

Speech and Signal Processing Group, Xerox Research Center–India, Bangalore, Karnataka, India
{Harish.Arsikere, Arunasish.Sen, Prathosh.AP, Vivek.Tyagi}@xerox.com

ABSTRACT

This paper presents 57 new acoustic features for automatic dialog-act tagging. The features are intended to be richer than and complementary to the traditional cumulative statistics of intonation. Some of our novel contributions include feature normalization with respect to neighboring utterances, incorporation of periodicity and formant features, modeling of cognitive phenomena such as hesitations, and utterance-level aggregation of short-term acoustic effects. The proposed features are applied to 3-way dialog-act tagging and question detection using two databases (British-English call-center conversations and Switchboard), and compared with a popular cumulative-statistics baseline using logistic-regression models. Our features are found to be significantly better than and complementary to the baseline, on average, achieving an *absolute* performance gain of ∼5–6%. Combined feature ranking reveals that about 75% of the top 20 features belong to the proposed feature set, and that the two corpora differ in their feature preferences despite similar overall performance.

Index Terms— acoustic features, dialog acts, call-center conversations, British English, question detection

1. INTRODUCTION

Dialog act (DA) tags represent the communicative acts that comprise natural conversations and task-oriented dialogs. Identifying whether an utterance is a question, statement, acknowledgment, and so forth can be useful in several speech and language processing applications such as spoken language understanding [1], annotation of archived conversations and meetings [2], DA-constrained speech recognition in human-machine dialogs [3], etc. Our interest in DA tagging stems from its application to post-call customer-care analytics: processing call-center conversations to automatically assess agent performance, track agent-customer alignment, and so on.

Previous studies have used a combination of lexical, syntactic, prosodic and discourse-structure cues for automatic DA tagging [1–8]. While lexico-syntactic features are clearly important for achieving good DA-tagging performance, prosodic features provide valuable complementary information [1,4]. This is particularly true when lexico-syntactic cues are computed using noisy transcripts from automatic speech recognition (ASR) systems. Additionally, some utterances (e.g., "okay") can be inherently ambiguous given just lexical information. In the context of customer care, acoustic cues also offer potential domain and language independence.

Previous work on exploiting acoustics/prosody for DA tagging—and the related task of question detection—has been mainly through cumulative and utterance-final (the final 200 ms) statistics (means, extremes, slopes, least-square polynomial-fit coefficients and derivatives) of pitch, intensity, speaking rate and duration [1, 2, 7, 9–12]. A comprehensive description of such features can be found in [1]. One exception to this popular cumulative-statistics approach is the n-gram approach of [4], wherein bi-grams of quantized pitch and intensity contours were shown to yield better DA-tagging performance than the pitch and intensity features of [1]. However, such n-gram features lack theoretical motivation and require large databases to be effective. Information about intonational events (pitch accents and prosodic phrase boundaries) could also benefit DA tagging [1], but such features are not entirely signal based because they require an event detector trained on manually-labeled data.

This paper proposes a rich set of acoustic features that goes beyond the traditional cumulative statistics of prosody. Our features are novel in several ways: (1) We often adopt the *divide-and-aggregate* approach: divide the given utterance into short segments, compute an acoustic parameter for each, and aggregate all the information into one feature value. Cumulative features, in contrast, either model utterance-final effects or treat the entire utterance as one segment. (2) To account for the fact that DAs are produced in the context of their neighboring utterances, several of our features are computed in a *turn-relative* manner—normalizing the features of a given DA with respect to those of its neighbors. (3) We sometimes model the *temporal location* of important acoustic events (location of the first pause, for example). Such cues have not been explored before, to our knowledge. (4) Some of our features are inspired by *cognitive phenomena*. One such feature is an estimate of utterance-initial speaking rate, which is expected to be higher for DAs that are better 'planned'. (5) We also model the degree of periodicity (a measure of voice quality) and formant dispersion (a measure of clarity); such cues are hypothesized to carry DA-specific information.

The proposed features are conceptually different from—and are hence expected to complement—existing feature sets (such as those of [1] and [4]). We apply our feature set to two tasks (three-way DA tagging and question detection) using two databases (conversations from a UK-based call center and a subset of the Switchboard corpus), and compare it with the popular feature set of [1]. Few studies have applied DA tagging to call-center or British-English corpora, which makes the present study relevant from a data perspective as well. This paper does not attempt to find the best model or classifier for the above tasks. It also does not attempt the automatic segmentation of dialogs into turns or turns into DAs.

2. DATA AND TASK DEFINITIONS

Our interest in DA tagging stems from its application to large-scale call-center analytics. For the present study, data from a UK-based call center were made available to us. This data will be referred to as the BCC (British Call Center) corpus in the rest of this paper. A subset of the Switchboard corpus [13] was also used for this study, the motivation for which was to (1) test the efficacy of the proposed features with more than one corpus, and (2) compare the feature usage for different conversation styles (task oriented versus free flowing) and accents of English (British versus American).

Fine label (Proportion)	Example(s)
Information Provision (40.3%)	"i can see that you called us a few times before"; "five eight one three"
Continuer (17.6%)	"right"; "okay"; "yeah no worries at all"
Information Request (17.2%)	"can i start by taking a contact telephone number"; "where do i get that from"
Reciprocity (5.1%)	"hello"; "hello there"; "alright cheers"; "bye bye"
Report Activity Status (3.2%)	"i put the sim card in it"; "hang on it's it's got a green light on it but that's all it's doing"
Confirmation (2.9%)	"<it's xyz> . . . xyz"; "<i've got the abc> . . . an abc okay"
Action Request (2.7%)	"and then pull your finger down the screen"; "yeah just uhm just press that"
Future Event Report (2.1%)	"and i'm going to get them to have a very in depth look at it"
Clarification Request (1.9%)	"oh okay you said it was uhm it was just restarting wasn't it"
Future Action Request (1.6%)	"uh i need you to print that one off and stick it to the front of the package"

Table 1. Examples of the 10 most-frequent fine labels in the BCC corpus. For *Confirmation*, <> indicates previous utterances.

The BCC corpus consists of 158 task-oriented dialogs regarding common cell-phone problems and their solutions. Each dialog was first segmented into agent and customer turns and then transcribed at the word level. Then, using the transcriptions and the audio files, each turn was divided into a maximum of three utterances (the most common number being one); turns were divided only when a change of intent was detected. Each utterance was assigned a "fine" label that captured the participant's action or intent at a local level. These fine labels were designed to be independent of syntax, semantics and the application domain (telecom services, bank transactions, etc.). A team of in-house annotators was responsible for all the segmentation, transcription and labeling tasks mentioned above.

In all, 17 unique fine labels were used to annotate the BCC corpus (a total of 10513 utterances). Examples of the 10 most-frequent fine labels are shown in Table 1—these fine labels account for $\sim$95% of the corpus. *Information Provision*, *Continuer* and *Information Request* emerge as the dominant labels, accounting for almost 75% of the utterances in the corpus.

It is probably clear from Table 1 that our fine labels do not conform to standard DA taxonomies (see [14] for examples)—they are in fact more indicative of social actions than speech or dialog acts. Therefore, for the purposes of this paper, the fine labels are binned into 3 widely-used DA categories: Questions (*Information Request*, *Action Request*, etc.), Statements (*Information Provision*, *Report Activity Status*, etc.) and Backchannels (*Continuer*, *Reciprocity*, etc.), based on which we attempt 3-way DA tagging and question detection. A 17-way fine-label classification is also of interest to us; this will be attempted once a larger corpus (with sufficient data points for each fine label) becomes available.

The SWBD-DAMSL tag set is the most popular DA taxonomy for Switchboard [15, 16]. The original tag set contains 60 labels, but most studies use reduced tag sets for classification purposes. The following 7 DA categories were identified in [1] and [4]: Questions, Statements, Backchannels, Agreements, Incomplete Utterances, Appreciations and Others. To enable comparisons with the BCC corpus, only the first three DAs were considered for this study. A subset (8129 utterances in total) of the Switchboard corpus was used for all experiments (3-way DA tagging and question detection).

3. ACOUSTIC FEATURES

We implemented 57 new acoustic features in all; owing to space constraints, however, only the more novel ones will be discussed here. For comparison purposes, we also implemented 42 baseline features from [1] (all features except those based a trained prosodic-event detector). To aid analysis, the features are binned into the following four categories: Pitch and Voicing ($\mathcal{P}$), Duration and Pausing ($\mathcal{D}$), Intensity ($\mathcal{I}$), and Speaking Rate and Rhythm ($\mathcal{S}$).

To model the context in which DAs are produced, we included the utterance-relative variants of some of our features. The utterance-relative variants of a feature f_n (n denotes the utterance number) are defined as $\log(f_n/f_{n-1}+1)$ and $\log(f_n/f_{n+1}+1)$.

To minimize gender- and speaker-dependent effects, some of our features were normalized with respect to a 'floor' value. While the baseline features use the entire conversation to compute such normalization parameters, the proposed features use just the given utterance, thus lending themselves better to online DA tagging.

The **Pitch and Voicing** category has 27 features, of which 8 are listed in Table 2. Given an utterance, the Snack sound toolkit (Version 2.2.10 [17]) was used to obtain the following three parameters at 10 ms intervals: (1) pitch (F0) frequency, (2) voiced/unvoiced decision (binary), and (3) normalized cross correlation (a real number between 0 and 1 that is proportional to the degree of voicing). The F0 floor of a given utterance was computed as the mode of its F0 histogram (after excluding unvoiced segments). Given all this information, features $\mathcal{P}1$–$\mathcal{P}8$ in Table 2 can be readily computed. Note that the chosen features highlight some of our novel ideas: $\mathcal{P}2$ and $\mathcal{P}3$ encode location information; $\mathcal{P}6$ and $\mathcal{P}8$ use the divide-and-aggregate approach; and $\mathcal{P}7$ provides an average voice-quality measure. As an example, Figure 1(a) shows that the location of maximum F0 occurrence can effectively distinguish between questions and statements; this is intuitive since questions are often associated with a terminal F0 rise [9–11].

The **Duration and Pausing** category has 8 features, of which 5 are listed in Table 2. Given speech/non-speech and voiced/unvoiced decisions at 10 ms intervals, $\mathcal{D}1$–$\mathcal{D}5$ can be readily computed. The algorithm in [18] was used to obtain the required speech-activity information. Some of our key ideas were employed in $\mathcal{D}1$ (modeling location information), $\mathcal{D}2$ (utterance-relative computation) and $\mathcal{D}4$ (modeling response time—a cognitive parameter). Another interesting feature (whose distributions are shown in Figure 1(b)) is $\mathcal{D}5$: the higher proportion of voicing in statements could be due to the fact that they typically have more frequent occurrences of "uh"s and "uhm"s (owing to limited pre-planning).

The **Intensity** category has 8 features, of which 4 are listed in Table 2. $\mathcal{I}1$–$\mathcal{I}4$ can be readily computed using intensity contours and speech-activity information. Intensity contours were obtained simply by computing the short-term energies of 20 ms frames spaced at 10 ms intervals. For $\mathcal{I}2$ and $\mathcal{I}3$, the contours were smoothed using a 5-point moving-average filter. For $\mathcal{I}1$ and $\mathcal{I}4$, intensity floor was computed as the mode of the intensity histogram (excluding non-speech segments). $\mathcal{I}2$ captures short-term intensity fluctuations over a syllable-like duration of 300 ms, while $\mathcal{I}3$ captures long-term intensity modulation. Figure 1(c) shows the feature distributions corresponding to $\mathcal{I}3$, suggesting that questions are, on an average, prosodically less 'lively' than statements.

Feature Name	Description
Pitch and Voicing ($\mathcal{P}$)	
1. `prcF0_extremes*`	percentage of F0 values that are less than 0.75*(F0 floor) or more than 1.5*(F0 floor)
2. `loc_`σ`F0_gt_Xprc`Σ`F0`	normalized location in utterance ($\in [0, 1]$) where the cumulative F0 sum surpasses X% of total F0
3. `loc_maxF0`	normalized location in utterance ($\in [0, 1]$) where F0 attains its maximum value
4. `avg_vSegDur*`	average duration of continuous voicing
5. `num_vSegs*`	number of continuously-voiced segments
6. `min_dct123_F0Segs`	F0 contour $\rightarrow$ continuously-voiced segments $\rightarrow$ [for each segment, percentage energy in the first three DCT coefficients of F0] $\rightarrow$ minimum over all segments
7. `prcNcorr_gt_0.9`	percentage of voiced frames having a normalized cross-correlation greater than 0.9
8. `min_range_F0Segs`†	F0 contour $\rightarrow$ continuously-voiced segments $\rightarrow$ [for each segment, (max F0 $-$ min F0) normalized by F0 floor] $\rightarrow$ minimum over all segments
Duration and Pausing ($\mathcal{D}$)	
1. `loc_firstPause`	normalized location in utterance ($\in [0, 1]$) where the first pause occurs
2. `dur_by_dur{P/N}`	ratio: duration of the given utterance to duration of the previous/next utterance
3. `num_spSegs`	number of segments having continuous speech activity
4. `pause_dur{P/N}`	duration of silence between the given utterance and the previous/next utterance
5. `prc_vSpeech`	percentage of speech frames that are voiced
Intensity ($\mathcal{I}$)	
1. `98prcInts_spRegs*`†	98th percentile of intensity (excluding non-speech frames), normalized by intensity floor
2. `avg_dct123_intSegs`	intensity contour $\rightarrow$ 300 ms chunks $\rightarrow$ [for each chunk, percentage energy in the first three DCT coefficients of intensity] $\rightarrow$ average over all chunks
3. `stdev_intPks`	standard deviation of intensity peaks that are higher than 0.5 times the mean peak value
4. `stdevInts_spRegs*`†	standard deviation of intensity (excluding non-speech frames), normalized by intensity floor
Speaking Rate and Rhythm ($\mathcal{S}$)	
1. `95prcEntr_utt*`	given utterance $\rightarrow$ 200 ms chunks $\rightarrow$ [for each chunk, spectral entropy of Mel filter-bank output] $\rightarrow$ 95th percentile over all chunks
2. `modn_dct2to10_beg`	first one second of the given utterance $\rightarrow$ Mel filter-bank output $\rightarrow$ [for each filter-bank channel, percentage energy in DCT coefficients 2–10] $\rightarrow$ average over filter-bank channels
3. `avg_modn_dct2to10`	given utterance $\rightarrow$ 1 sec. chunks $\rightarrow$ [`modn_dct2to10_beg` for each chunk] $\rightarrow$ average over chunks
4. `rangeF1_utt`†	(max F1 $-$ min F1) normalized by median value of F1
5. `min_stdevF1_200`	F1 contour $\rightarrow$ 200 ms chunks $\rightarrow$ [for each chunk, standard deviation of F1] $\rightarrow$ minimum over chunks

Table 2. A list of some of our key acoustic features. Features marked with a '*' have utterance-relative variants, and features marked with a '†' are normalized with respect to a 'floor' value. "DCT" in all cases stands for Discrete Cosine Transform.

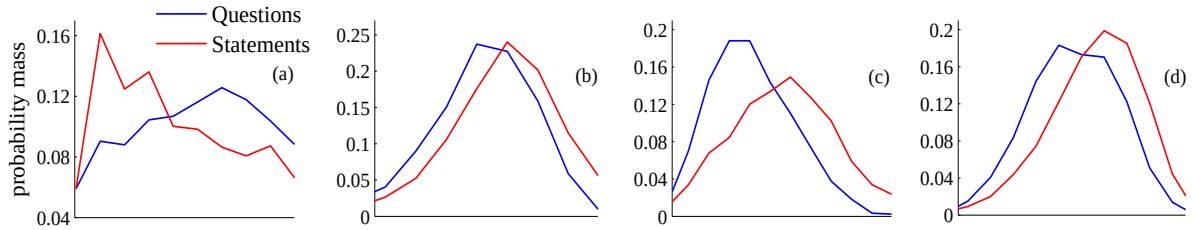

Fig. 1. Histograms of (a) `loc_maxF0`, (b) `prc_vSpeech`, (c) `stdev_intPks`, and (d) `modn_dct2to10_beg` (from the BCC corpus).

This **Speaking Rate and Rhythm** category has 14 features, of which 5 are listed in Table 2. To compute $\mathcal{S}4$ and $\mathcal{S}5$, contours of the first formant (F1) were obtained using the Snack toolkit. While $\mathcal{S}4$ is an indicator of spoken clarity (inspired by [19]), $\mathcal{S}5$ is a rough measure of short-term vocal-tract constancy. Similar features were computed using contours of the second formant. Note that all formant-based features were obtained using voiced frames only. To compute $\mathcal{S}1$–$\mathcal{S}3$, a standard 26-channel Mel filter bank was used. While $\mathcal{S}1$ uses a divide-and-aggregate approach to model short-term spectral entropy (the definition of spectral entropy was borrowed from [20]), $\mathcal{S}2$ and $\mathcal{S}3$ rely on the modulation spectra of long speech segments. For example, $\mathcal{S}2$ considers the first one second of the given utterance and computes a speaking-rate estimate based on the amount of low-frequency modulation energy. DAs that are not completely planned before production (e.g. statements in response to questions) have a tendency to begin slowly; consequently, they are expected to have a higher value of $\mathcal{S}2$. The feature distributions in Figure 1(d) show that this hypothesis is indeed true.

Most of our features were transformed using a non-linear function such as log(), cube root, fifth root, etc., in order to reduce their dynamic ranges and Gaussianize their distributions. No further normalization (e.g. z-score normalization) was applied.

4. RESULTS AND DISCUSSION

Five-fold cross validation was used for all of our experiments (3-way DA tagging and question detection on the BCC and Switchboard corpora); there were no speaker overlaps between the train and test folds. Certain feature values could not be computed on rare occasions (owing to very short utterance durations, for instance); such missing values were imputed with class-specific means during training, and class-independent means (obtained from training data) during testing. Logistic regression with cross-entropy loss (MATLAB's `mnrfit` function) was used for classification.

For 3-way DA tagging, we chose the average rate of correct detection (ACD) as our performance metric:

$$\text{ACD} = \frac{1}{3}\sum_{i=1}^{3} P(\hat{y} = i \mid y = i) \times 100, \quad (1)$$

where y and $\hat{y}$ are the actual and estimated class labels, respectively. For question detection, we chose the area under the ROC (receiver operating characteristic) curve (AUC) as the performance metric. Both ACD and AUC were averaged over the 5 cross-validation folds.

	$\mathcal{P}$	$\mathcal{D}$	$\mathcal{I}$	$\mathcal{S}$	**All**
Baseline [1]	56.0	53.6	37.8	46.3	60.5
Proposed	59.1	54.4	52.2	54.7	65.2
Combined	61.9	58.6	53.8	53.8	66.7

Table 3. Average ACD results for DA tagging with the BCC corpus.

	$\mathcal{P}$	$\mathcal{D}$	$\mathcal{I}$	$\mathcal{S}$	**All**
Baseline [1]	60.4	58.9	46.1	39.1	63.7
Proposed	61.8	60.7	54.2	56.9	66.4
Combined	65.1	61.6	55.7	54.0	68.3

Table 4. Average ACD results for DA tagging with Switchboard.

	BCC corpus	**Switchboard**
Baseline [1]	0.793	0.750
Proposed	0.820	0.765
Combined	0.844	0.795

Table 5. Average AUC values for question detection.

	$\mathcal{P}$	$\mathcal{D}$	$\mathcal{I}$	$\mathcal{S}$
BCC corpus	7 (3)	5 (1)	3 (0)	5 (4)
Switchboard	7	4	2	7

Table 6. Number of features from each category in the list of top 20 features (features were ranked using the MRMR criterion). For each category, the number of features that are common to the two corpora is indicated in parentheses.

Tables 3 and 4 show 3-way DA tagging results corresponding to each feature category and feature set (baseline, proposed and combined), for the BCC and Switchboard corpora, respectively. The proposed features outperform the baseline features in every category, but the improvements are most significant in the case of Intensity and Speaking Rate features. The proposed features also complement the baseline features in most cases, notably the Pitch and Duration categories. Overall, for the BCC corpus, the proposed and the combined feature sets yield absolute performance gains of 4.7% and 6.2%, respectively, over the baseline. The improvements are smaller, but significant, for the Switchboard corpus. Note that the proposed features, despite their superior performance, cannot be used by themselves in a practical DA tagging system; they must be used in conjunction with lexico-syntactic cues to achieve higher performance levels.

Table 5 shows the results of question detection obtained for the BCC and Switchboard databases. Our features are again better than and complementary to the baseline features, but the gains are slightly smaller compared to 3-way DA classification. The combined feature set provides absolute performance gains of 5.1% and 4.5% over the baseline, for the BCC and Switchboard corpora, respectively.

4.1. Feature Analysis

To analyze feature importance and to understand the differences in feature usage between the two databases, we combined the baseline and the proposed features and ranked them together using the minimum redundancy maximum relevance (MRMR) criterion [21]. The top 20 features led us to the following observations.

• The number of features belonging to the proposed feature set is significantly high—15/20 features for the BCC corpus and 14/20 features for Switchboard, thus clearly showing that the performance of the combined feature set (last row and column of Tables 3 and 4) is largely due to the proposed features.

• All four feature categories are important. For example, the top 20 features for the BCC corpus include (cf. Table 2): $\mathcal{P}5$, $\mathcal{P}5$*, $\mathcal{S}3$, $\mathcal{I}3$, $\mathcal{D}2$, $\mathcal{I}1$*, $\mathcal{D}1$, $\mathcal{D}5$, $\mathcal{I}2$, $\mathcal{P}6$, and a variant of $\mathcal{S}4$.

• Only 8 of the top 20 features are common to the two corpora. This difference in feature usage could be due to the differences in conversation styles (informal in Switchboard but task oriented in the BCC corpus) and accents of English (American versus British). Table 6 shows the number of features used in each category. While the relative contributions of the four feature categories are similar for the two corpora, the difference appears mainly in the kind of Duration and Intensity features used.

• Of all the novel ideas introduced in this paper (see Section 1), division-and-aggregation and utterance-relative normalization seem to be the most useful. For example, of the 15 new features that appear in the top-20 list of the BCC corpus, 5 are based on the first idea and 4 are based on the second.

5. CONCLUSION

The proposed acoustic features are conceptually different from conventional DA-tagging cues—they are based on fundamentally new concepts such as feature normalization with respect to neighboring utterances and the utterance-level aggregation of short-term acoustic effects. Our features were designed around the following acoustic-prosodic parameters: pitch, voicing, duration, pausing, intensity and speaking rate; a total of 57 features were implemented. Experimental evaluation using two databases (call-center conversations in British English and a small subset of the Switchboard corpus) and two tasks (question detection and 3-way classification of questions, statements and backchannels) showed the proposed features to be significantly better than and/or complementary to a popular cumulative-statistics baseline [1]. Feature-usage analysis revealed that the speakers in the above two corpora have different feature preferences, particularly in the duration and intensity categories; this can be attributed, at least in part, to task and accent differences.

While the results in this paper are encouraging, they must be validated further by combining our features with lexico-syntactic cues, experimenting with more fine-grained DA tags, and using automatic methods for segmenting dialogs into turns and turns into DAs. These ideas constitute our future work.

6. REFERENCES

[1] E. Shriberg, A. Stolcke, D. Jurafsky, N. Coccaro, M. Meteer, R. Bates, P. Taylor, K. Ries, R. Martin, and C. Van Ess-Dykema, "Can prosody aid the automatic classification of dialog acts in conversational speech?" *Language and Speech*, vol. 41, pp. 443–492, 1998.

[2] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *Proceedings of ICASSP*, 2005, pp. 1061–1064.

[3] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, pp. 339–373, 2000.

[4] V. K. R. Sridhar, S. Bangalore, and S. Narayanan, "Combining lexical, syntactic and prosodic cues for improved online dialog act tagging," *Computer Speech & Language*, vol. 23, pp. 407–422, 2009.

[5] N. Webb, M. Hepple, and Y. Wilks, "Dialogue act classification based on intra-utterance features," in *Proceedings of the AAAI Workshop on Spoken Language Understanding*, 2005.

[6] S. Rosset and L. Lamel, "Automatic detection of dialog acts based on multi-level information," in *Proceedings of ICSLP*, 2004, pp. 540–543.

[7] R. Fernandez and R. W. Picard, "Dialog act classification from prosodic features using support vector machines," in *International Conference on Speech Prosody*, 2002.

[8] D. Surendran and G.-A. Levow, "Dialog act tagging with support vector machines and hidden markov models," in *Proceedings of Interspeech*, 2006.

[9] S. Ananthakrishnan, P. Ghosh, and S. Narayanan, "Automatic classification of question turns in spontaneous speech using lexical and prosodic evidence," in *Proceedings of ICASSP*, 2008, pp. 5005–5008.

[10] V. M. Quang, L. Besacier, and E. Castelli, "Automatic question detection: prosodic-lexical features and cross-lingual experiments," in *Proceedings of Interspeech*, 2007, pp. 2257–2260.

[11] K. Boakye, B. Favre, and D. Hakkani-Tür, "Any questions? Automatic question detection in meetings," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 2009, pp. 485–489.

[12] J. Liscombe, J. J. Venditti, and J. B. Hirschberg, "Detecting question-bearing turns in spoken tutorial dialogues," in *Proceedings of Interspeech*, 2006.

[13] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proceedings of ICASSP*, 1992, pp. 517–520.

[14] D. R. Traum, "20 questions on dialogue act taxonomies," *Journal of Semantics*, vol. 17, pp. 7–30, 2000.

[15] D. Jurafsky, E. Shriberg, and D. Biasca, "Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13," *Institute of Cognitive Science Technical Report*, pp. 97–102, 1997.

[16] D. Jurafsky, E. Shriberg, and D. Biasca, "Switchboard Dialog Act Corpus," http://web.stanford.edu/~jurafsky/ws97/, last accessed: 2015-01-20.

[17] K. Sjölander *et al.*, "The Snack sound toolkit," http://www.speech.kth.se/snack/, last accessed: 2015-01-20.

[18] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, pp. 1–3, 1999.

[19] E. Godoy, M. Koutsogiannaki, and Y. Stylianou, "Assessing the intelligibility impact of vowel space expansion via clear speech-inspired frequency warping," in *Proceedings of Interspeech*, 2013, pp. 1169–1173.

[20] H. You, Q. Zhu, and A. Alwan, "Entropy-based variable frame rate analysis of speech signals and its application to ASR," in *Proceedings of ICASSP*, 2004, pp. 546–549.

[21] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1226–1238, 2005.