What Speech Tells Us About Discourse: The Role of Prosodic and Discourse Features in Dialogue Act Classification

Mohammed E. Hoque, Mohammad S. Sorower, Mohammed Yeasin, Max M. Louwerse

***Abstract*— This paper explores the relative importance of discourse features, prosodic features and their fusion in robust classification of speech acts. Five different feature selection algorithms were used to select set of features to improve the robustness of the classification. The natural synergy between subset of prosodic and discourse features was then used to model the speech acts using different categories of classifiers.**

I. Introduction

Understanding and producing multimodal communication in humans and agents requires an understanding not only of the semantic meaning of an utterance, but also of the intended meaning behind that utterance. Take for instance an utterance like "go between those". This utterance could be interpreted as an instruction ("you should go between those!"), as a yes/no question ("should I go between those?"), as an acknowledgment (speaker just stated "go between those" and the respondent confirms acknowledging the utterance by repeating "got it, go between those"). In all three cases the semantic meaning of the utterance is the same (there is an event of going and an implied patient is undergoing this event). What differs is the pragmatic meaning behind each of these utterances. This pragmatic meaning can be captured in speech acts, like questions, commands, promises, warnings. Knowing what speech act an utterance can be classified in can help reveal its pragmatic meaning [1][2][3][4].

Speech acts are known to shape the structure of the dialogue and can often be helpful in predicting the intonational patterns for a dialogue. Studies [5][6] have shown that the sequence of speech acts and the association between such acts and observed intonational contours can significantly help the performance of speech recognition engines. Speech acts have even proven to be useful in predicting eye-brow movements [7] and modalities like eye gaze, facial expressions and route drawings [8].

Successfully classifying utterances into speech acts is a research challenge for computational linguists, engineers, computer scientists and psychologists alike. Most speech act classification systems rely on discourse features to assign utterances to a speech act [1][9][10][11]. The problem with systems like these is that they solely focus on discourse features, making it unclear whether, and to what extent, other modalities, like speech features, also contribute to speech act classification. Moreover, only relying on discourse features makes online classification problematic. Whereas offline classification may work well because of the availability of discourse features that were carefully transcribed from speech, online classification does not have access to these features due to the far from optimal performances of speech recognition systems.

It seems easy and logical to consider prosodic features in speech act classification. In the example used earlier ("go between those") by analyzing the intonation pattern (e.g. rising or falling pitch), the utterance can be classified as a question or an instruction. Natural conversations, however, turn out to have little variation in pitch contour and intonation pattern for many speech acts.

Let us illustrate this with an example from a large corpus of natural multimodal communication, to be discussed below. Figure 1(a) shows the pitch contour of a small segment of a conversation between two dialogue partners, where one initially asks a question ("in between those?") and the other reaffirms by responding ("uh-huh, in between those."). Figure 1(b) and 1(c) shows the pitch contour of the same statement (*in between those*), used in two different ways, a question and statement. From Figure 1, it is evident that there are a few noticeable differences between the pitch contours, despite the fact that the two utterances mark different speech acts (instruction and yes/no question).

The little variation in pitch contours perhaps explains the relatively low accuracy in speech act classification obtained through prosody only, ranging from 40-43% [12][13].

The performance of speech act classification systems can perhaps be improved by fusing prosody and discourse information together. The classifier should not only be capable of disambiguating discourse information, but should also compensate for the low word recognition rate of the speech engines by using prosody, thus minimizing syntactic and semantic search complexities [14][15]. Also, it is important to study the relative importance of the features from discourse and speech data by identifying features that are more important into the classification decision. This feature selection framework not only provides useful cues regarding which features are more relevant for a particular dialogue act, but also helps to reduce the dimensionality of the feature set by eliminating collinear features. In the past, speech act classifications have been performed with one or two classifiers, like support vector machines and hidden markov model [12][13]. However, we argue that effectiveness of a feature set is dependent on the characteristics of classifiers. While a certain feature set may

M. E. Hoque is with the Department of Electrical and Computer Engineering / Institute for Intelligent Systems, Memphis, TN 38152 USA (e-mail: mhoque@memphis.edu).

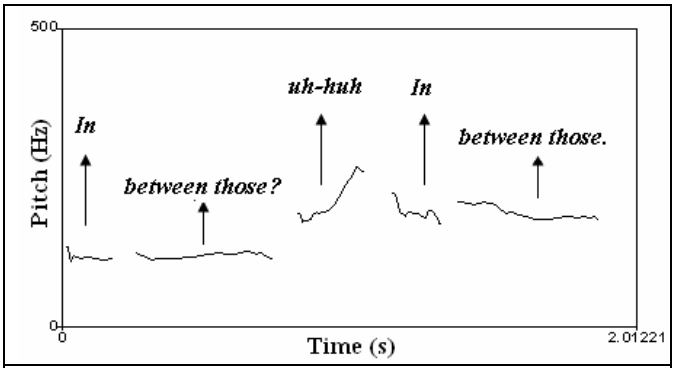
M. S. Sorower is with the Department of Electrical and Computer Engineering / Institute for Intelligent Systems, Memphis, TN 38152 USA (e-mail: msorower@memphis.edu).

M. Yeasin is with the Department of Electrical and Computer Engineering / Institute for Intelligent Systems, Memphis, TN 38152 USA (e-mail: myeasin@memphis.edu).

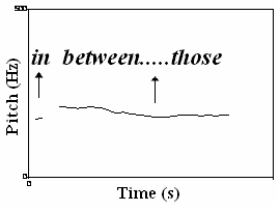
M. M. Louwerse is with the Department of Psychology / Institute for Intelligent Systems, Memphis, TN 38152 USA (e-mail: mlouwerse@memphis.edu).

work well with one classifier, it may fail for others. Therefore, it is important to create models of features using a variety of feature selection algorithms and test those models across a diverse set of classifiers. Through this approach, it may be possible to identify discourse and speech feature sets that are robust across all the diverse set of classifiers.

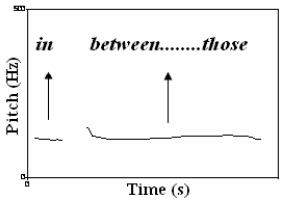
This paper addresses the questions of relative importance of discourse features, speech features and their fusion in speech act classification. To address such questions and use the synergy, a subset of speech and discourse features have been identified using 5 different feature selection algorithms and then tested with 7 sets of classifiers allowing for a comparison of features, classifiers and modalities.



(a) A segment of a map-task conversation
Speaker A: in between those?
Speaker B: uh-huh..in between those



(b) Speaker A: in between those?



(c) Speaker B: in between those.

Figure 1: Pictorial description (pitch) of a case where prosody fails to distinguish between a question and statement
(a) The overall conversation in context, (b) Question made by Speaker A; (c) Response made by Speaker B.

II. MEMPHIS MULTIMODAL MAP TASK CORPUS

Our interest in speech acts stems from a large multimodal communication project [16][17]. This research project explores how different modalities in face-to-face dialogues align with each other and tries to implement those rules extracted from human experiments in an artificial conversational agent (ECA). The ECA is expected to interact with humans more naturally as a validation of the study. In order to engage human participants into a natural task oriented conversation, the Map Task scenario [18] has been chosen as the general setup for study.

The Map Task is a map-oriented experimental setting in which two participants work together to achieve a common goal through conversation. One of the participants is arbitrarily denoted as Instruction Giver (IG) who collaborates with the other partner, known as Instruction Follower (IF), to reproduce on IF's map a route printed on IG's (Figure 2). However, the maps of the IG and IF are not identical. Different landmarks or features of landmarks are used. Moreover the color of some landmarks on IF's map, are obscured by an ink blot. The differences are intentionally designed to elicit dialogue in a controlled environment based on common ground and differences in their maps. These inconsistencies between the maps are expected to be resolved through multimodal communication between the IG and IF.

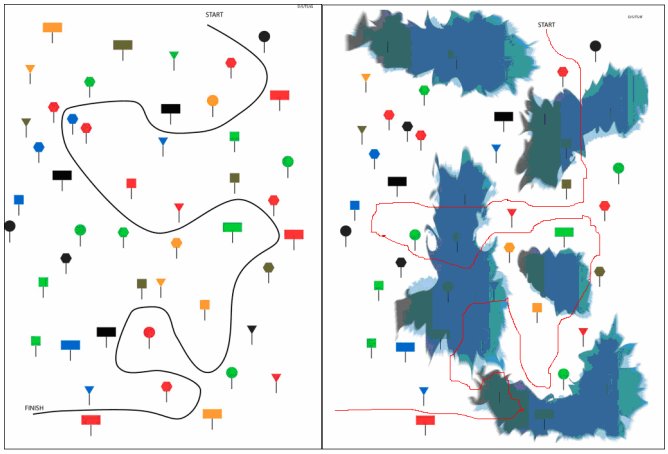


Figure 2. Example of maps. IG map presented on left, IF's map (with route drawn by IF) on right.

The current corpus consists of 256 conversations from 64 participants. All the participants were instructed to perform the role of IG (4 conversations) and the role of IF (4 conversations). Different maps that varied in terms of homogeneity of objects were used in each conversation. Figure 2 demonstrates an example of the maps for the IG and IF. The participants included 62% of female, and 39% of African American and 57% of Caucasian. For this experiment, 16 conversations were randomly sampled totaling 72 minutes of dialogue with different participants and different maps for each conversation. Below we focus on those aspects of the corpus relevant for this study.

Thirty-two participants performed the multimodal Map task, 21 of them females and 11 males. All of the participants were native speakers of English. A Marantz PMD670 speech recorder was used to record speech of IG and IF on two separate (left and right) channels using two AKG C420 headset microphones. Participants, seated in front of each other, were separated by a divider to prevent any direct communication between them. They could only communicate through microphones and headphones, while they could view both the upper torso of the dialogue partner and the map on a computer monitor in front of them. A colored map was presented to IG with a route drawn on it (similar to the one presented in Figure 2). The IG was supposed to communicate the route information to the IF as accurately as possible. The 12 dialogue acts that are typically used for Map Task coding were used [1][2]. Table 1 presents an overview of these dialogue acts with necessary descriptions and examples. The utterances of half of the

conversations were manually coded as one of the twelve dialogue acts. Inter-rater reliability between the coders in terms of Cohen's Kappa was satisfactory at .67. Coders resolved the conflicts, primarily relating to the acknowledgment dialogue act, and coded the remaining transcripts for dialogue acts.

III. Proposed Approach

The proposed approach consists of five main components as shown in Figure 3, namely, I) segment the conversation automatically into turns, II) extract (manually) dialogue acts from turns (if applicable) and then label them using human experts, III) mine the feature-space to select novel prosodic and discourse features from the audio-visual corpus, IV) fusion of prosodic and discourse features, and V) use various machine learning techniques for classification of dialogue acts. Subsequent subsections briefly discuss each module of the proposed speech act classification system.

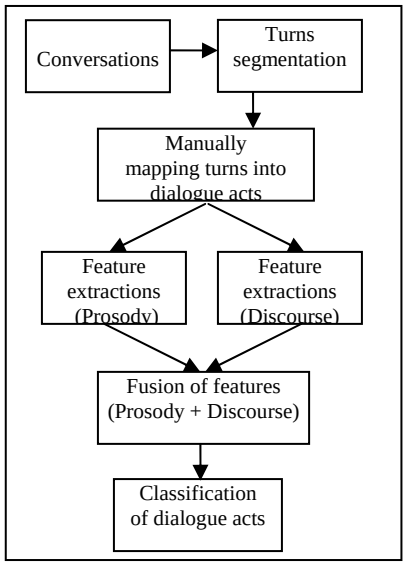


Figure 3: High Level diagram of the dialogue act classification system.

A. Turn segmentation

To detect the turn, the speech signal from the IG and IF have been considered simultaneously. The pauses in spoken words were used as the feature to detect the beginning and end of a turn in a natural conversation. Pauses were detected on each audio channel using the upper intensity limit and minimum duration of silences. In measurement of intensity, minimum pitch specifies the minimum periodicity frequency in any signal. In this case, 75 Hz for minimum pitch yielded a sharp contour for the intensity. Audio segments with intensity values less than its mean intensity were classified as pauses. Thereby, mean intensity for each channel rather than a pre-set threshold was used, enabling our pause detection system to properly adapt to the diverse set of voice properties of the participants. Any audio segments with silences more than 0.4 seconds were denoted as pauses. The speech processing software *Praat* [19] was used to perform all calculations to identify these pause regions.

Figure 4 gives an example of how a conversation gets segmented into turns. Note that turn 3 contains more than one speech act and thus, needs to be segmented further. Therefore, some manual inspection was needed to segment the conversation into speech acts from turn levels.

IG: Go right……...okay….....then..go straight

turn1 | *pause* | *turn3*| *pause*| *cont. of turn3*

IF: ……..okay……..

pause |*turn2*| *pause*

Segmented turns:
turn 1- IG - *Go right*
turn 2- IF - *Okay*
turn 3- IG - *Okay……then...go straight*

Segmented speech acts:
turn 1- IG - *Go right* (Instruction)
turn 2- IF - *Okay* (Acknowledgment)
turn 3- IG - *Okay* (Acknowledgment)
turn 3- IG - *Then...go straight* (Instruction)

Figure 4: Example of how turns are segmented from conversations.

B. Features Extraction

Prosodic features related to pitch, intensity, formant, duration, pauses, rhythm were extracted (details are provided in Table 2). Discourse features that were extracted included parts of speech tagging and sequence, dialogue history, probability of one utterance belonging to 13 different categories using Probabilistic Latent Semantic Analysis (PLSA) [20], as shown in Table 2.

C. Features Selection and Classification

To boost the performance of speech act classification, extracted discourse or speech features are often projected onto the low dimensional subspace [13][21] (for instance using principle components analysis and linear discriminant analysis). While the subspace projection adds values in improving the performance of model, but often fails to answer important questions such as which set of features carry most information. To solve this problem, a few feature mining algorithms are used for the selection of features. The selected set of features is used as input to various machine learning techniques (for example, Bayes, Functions, Meta, Trees, and Rule based classifiers) to model different speech acts.

IV. Results and Discussions

Five different feature sets were created for prosody and discourse using five different feature selection algorithms, such as, Subset Evaluator (Best First), Chi Squared Attribute Evaluator (Ranker), Consistency Subset Evaluator (Greedy

TABLE 1. SPEECH ACT CATEGORIES (DESCRIPTION AND EXAMPLES)

DialogAct	Description
INSTRUCT	Commands partner to carry out action *Go between the two green houses.*
EXPLAIN	States information not directly elicited by partner *I have a set of four hours*
CHECK	Requests partner to confirm information *So, between the green and blue one?*
ALIGN	Checks attention, agreement, readiness of partner *Ok, do you see those two blue cars?*
QUERY-YN	Yes/no question that is not CHECK or ALIGN *Do you see the house?*
QUERY-W	Any query not covered by the other categories *What do I do after I cross the house?*
ACKNOWL	Verbal response minimally showing understanding *Uh huh.*
REPLY-Y	Reply to any yes/no query with yes-response *Yeah, I see the house.*
REPLY-N	Reply to any yes/no query with no-response *No, no house there.*
REPLY-W	Reply to any type of query other than 'yes or 'no' *I see a car.*
CLARIFY	Reply to question over and above what was asked *Cross the car and there is a house*
READY	Preparing conversation for new dialog game *Alright, you are going to start at the top.*
UNCODBL	e.g. laughing

step wise), and Gain Ratio Attribute Evaluator (Ranker)[23]. Seven different categories of classifiers - Bayes, Function, Meta, Tree and Rule were used to model 12 speech acts using the feature sets. This helped to identify the robust set of prosodic and discourse features that can be used to model the speech acts using diverse classifiers. For example, combination of four prosodic features, such as, *role (IG or IF), duration of the speech act, average value of the second formant, and speaking rate* were found to be the most important features using SubSet evaluator and its performance was comparable to the model which employed 50 prosodic features. Chi Squared Attribute evaluator yielded features such as *speaking rate, duration of the speech act, ε_{time}, role, number of voice breaks in a dialogue act* as the optimal feature with reasonable accuracy rate. Feature sets generated using Consistency Subset Evaluator were able to classify 13 different speech acts 48% of the time in average, using features such as *role, energy, F0 related statistics, statistics related to second and third formant, number of voice breaks, pauses, and number of rising and falling edges* in a dialogue act.

A similar procedure was employed to identify the optimal discourse feature sets and test their accuracy. For SubsetEvaluator feature selection algorithm, features such as, *role (IG or IF), number of words in each speech act, previous speech act, the first three sequences of the parts of speech of the speech act*, yielded comparable accuracy in compare to another model with more than 100 discourse features. Consistency Subset Evaluator, however, yielded the highest accuracy of distinguishing any of the 13 speech acts 65% of the time, using features such *role, number of parts of speech (cardinal number, determiner, noun, verb, and adjective), number of words in each speech act, the first 5 sequence of the parts of the speech, previous two speech acts*.

Finally, the prosodic features were fused with the discourse features to boost overall classification accuracy. A simple feature level fusion of the discourse and prosodic features yielded an average of 65.60% accuracy, with the highest of 70.56% obtained using the meta based classifier. Based on the observation, it can be inferred that adding prosody with discourse in MapTask corpus does boost the overall accuracy as demonstrated in Figure 4. For example, for Function based classifiers, the performance increase was up to 8.33%.

Table 3 and 4 also show that the classification performance of LogitBoost, a log based classifier, is noticeably consistent on average yielding the highest performance across the three categories (speech, discourse speech and discourse). The fusion and normalization of data from the two linguistic modalities speech and text, is a difficult problem. LogitBoost shrinks the dynamic range of the prosodic and discourse features. The monotonic logarithmic mapping thus makes LogitBoost more consistent and robust than the other classifiers used in this paper. To improve the performances of other classifiers, a better normalization approach is required and will be explored in future work.

It is noteworthy that the two linguistic modalities, prosody and discourse features, correlate in the mistakes they make in the speech act classification process ($r = .88$, $p < .001$, $N = 156$). Whereas we had expected that prosody would capture differences between speech acts like instruction and query-yn, and discourse would capture differences like reply-y and reply-n, this is not what the current results show. Explanations need to be further investigated, but sample size for some of the dialogue acts, ambiguity in the coding system, and limited lexical and prosodic variation in natural speech are some of the tentative explanations.

It is also surprising that the probability values for a given utterance belonging to a certain class obtained from PLSA turned out to be very insignificant. One potential explanation is that PLSA is suited for larger paragraphs, whereas most of the current corpus consists of smaller utterances (5-6 words per utterance on an average). The low performance of the PLSA may also be explained by the Maximum Likelihood (ML) that is used to estimate the model parameters (distribution of words per speech acts and the distribution of speech acts in the corpus). Given the sparse nature of the term-dialogue matrix it is hard to estimate the model parameters using the classic ML.

Future effort on speech act classification will include fusion of classifiers by utilizing their diversity in the decision process. In this experiment, it was evident that

certain classifiers work best under certain conditions with different kinds of feature sets. For example, from Table 4, it can be inferred that function based classifier SMO provides the highest average performance enhancement by fusion of prosody and discourse. But Table 4 also shows that meta-based classifiers, Bagging and LogitBoost, provide the highest average accuracy for prosody and discourse, respectively. Therefore, future efforts will include fusion of those classifiers considering the diversity of their decision process.

TABLE 2: THE PROSODY AND DISCOURSE FEATURES EXTRACTED FROM DIALOGUE ACTS AND THEN OPTIMAL FEATURES WERE IDENTIFIED

		Features	Optimal features
Prosody	Pitch	Minimum (pMin), Maximum (pMax), Mean (pMean), Standard Deviation (pSD), Absolute Value (pAV), Quantile (pQ), Unvoiced/Voiced frames of pitch (pUV).	pMin, pMax, pMean, pAB, pQ, pUV
	Intensity	Minimum (iMin), Maximum (iMax), Mean (iMean), Standard Deviation (iSD), Quantile (iQ)	iMin, iMax, iQ
	Formant	Average value of first formant (fVal1), second formant (fVal2), third formant (fVal3). Average bandwidth of first formant (fBand1) , second bandwidth (fBand2), third bandwidth (fBand3), Mean of first formant (fMean1), second formant (fMean2), third formant (fMean3), fMean2/fMean1, fMeanf3/fMean1, Standard deviation of first formant (f1STD), second formant (f2STD), third formant (f3STD), f2STD/f1STD, f3STD/f1STD	fVal2, fVal3, fBand1, fmean3, fMean3/fMean1, f1STD, f3STD
	Duration	duration of the dialogue act (d1), ε_{time}, ε_{height} [21]	D1, ε_{time}
	Pauses	percent of Unvoiced Frames (pUF), # of Voice Breaks (#OVB), percent of Voice Breaks (pVOB), # of Pauses (nP), maximum duration of Pauses (mdp), average duration of pauses (adp) , total duration of Pauses (tdp)	#OVB, pVOB, nP, adp, mdp, tdp
	Rhythm	speaking Rate (SR).	SR
	Edges	Magnitude of the highest rising edge (mhre), magnitude of the highest falling edge (mhfe), average magnitude of all the rising edges (amare) average magnitude of all the falling edges (amafe), # of rising edges (#re), # of falling edges (#fe).	#re, #fe
	Misc.	jitter (jt), shimmer (sh), energy (e), power (p), role	Role, energy
Discourse	Parts of speech sequence in each utterance [P1-P30]		P1, P2, P3, P4, P4, P5, P6, P7
	Number of words in an utterance [WC]		WC
	Probabilistic Latent Semantic Analysis (PLSA) [20] values for each utterance belonging to a cluster (Model fitting with the EM algorithm) [cluster1-cluster12]		
	Previous speech acts (prev1, prev2)		Prev1, Prev2
	Parts of speech tagging [22]		CD, DT, EX, IN, JJ, VB, VBN, WP, VBP

TABLE 3: ACCURACY OF CLASSIFYING 13 DIALOGUE ACTS WITH PROSODY ONLY, DISCOURSE ONLY, AND BOTH (PROSODY + DISCOURSE), M0= ALL THE FEATURES, M1= SUBSET EVALUATOR (BESTFIRST), M2= CHI SQUARED ATTRIBUTE EVALUATOR (RANKER), M3= CONSISTENCY SUBSET EVALUATOR (GREEDY STEP WISE), M4= GAIN RATIO ATTRIBUTE EVALUATOR (RANKER).

	Feature Selection Algorithm	Accuracy to classify 13 dialogue acts (%)							Avg. Accuracy (%)
		Bayes based classifier	Function based classifier	Meta based classifier	Meta based classifier	Tree based classifier	Tree based classifier	Rules based classifier	
		Bayes Net	SMO	LogitBoost	Bagging	Random Forest	J48	Decision table	
Prosody	M0	42.67	49.22	50.06	51.72	50.36	40.43	49.26	47.67
	M1	50.28	47.95	48.01	49.38	42.75	43.08	50.50	47.42
	M2	46.67	48.07	47.65	48.64	37.56	47.14	50.36	46.58
	M3	46.68	48.60	48.73	51.84	49.42	41.36	49.05	47.95
	M4	46.32	47.99	47.89	48.85	42.34	42.06	50	46.49
Discourse	M0	61.66	68.16	67.02	64.62	65.5	66.03	64.36	65.33
	M1	66.48	60.88	64.98	61.97	60.52	63.17	61.25	62.75
	M2	64.25	59.56	67.74	62.7	65.13	66.03	62.45	63.98
	M3	64.73	64.60	66.54	63.37	64.6	66.45	64.15	64.92
	M4	61.35	54.7	62.57	64.11	62.76	63.41	62.97	61.69
Both	M0	58	69.21	70.56	65	64.04	66.35	66.10	65.60
	M3	63.46	64.93	68.2	63.09	63.95	65.46	64.15	64.74

TABLE 4: COMPARISON OF CLASSIFIERS IN SPEECH ACT CLASSIFICATION

Classifier	Accuracy Prosody (%)	Accuracy Discourse (%)	Accuracy Prosody & Discourse (%)	Boost by fusion (%)
BayesNet	46.524	63.70	60.73	-4.67
SMO	**48.37**	**61.58**	**66.71**	**8.33**
LogitBoost	48.47	**65.78**	**69.38**	5.47
Bagging	**50.09**	63.354	63.05	-.47
Random Forest	44.49	63.70	63.96	.41
J48	42.81	65.09	65.25	.24
Decision Table	49.83	63.036	65.13	3.32

V. CONCLUSION

This paper has addressed the questions of relative importance of discourse features, speech features and their fusion in speech act classification. Five different feature selection algorithms and tests with seven sets of classifiers on a natural multimodal communication corpus showed that certain classifiers work best under certain conditions with different kinds of feature sets. That is, a one-size-fits-all approach for algorithms and classifiers does not yield optimal performance. Instead, a synthesis of algorithms and classifiers is needed. Similarly, the results presented here show that discourse and prosodic features are intrinsically related, whereby for speech act classification speech says as much about discourse, as discourse about speech.

ACKNOWLEDGMENT

This research was supported by grant NSF-IIS-0416128. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding institution. We would like to thank Ellen Bard and Markus Guhe for their help in developing the maps, and Nick Benesh, Gwyneth Lewis, Patrick Jeuniaux, Jie Wu and Megan Zirnstein for their help in the data collection and analyses.

REFERENCES

[1] M. M Louwerse, & S. Crossley, Dialog Act Classification using n-gram Algorithms, *Proceedings of the Florida Artificial Intelligence Research Society International Conference (FLAIRS), Florida, USA,* 2006. Menlo Park, CA: AAAI Press.

[2] J. Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon, and A. Anderson, *The Reliability of a Dialogue Structure Coding Scheme*, Computational Linguistics, vol. 23, 1997, pp. 13-31.

[3] J. Searle, A Taxonomy of Illocutionary Acts., Minnesota Studies in the Philosophy of Language, ed. K. Gunderson, 1975, pp. 334-369. Minnesota: Univ. of Minnesota Press.

[4] J. L. Austin. How to Do Things with Words. Oxford: Oxford University Press. , 1962.

[5] P. Taylor, S. King, S. Isard and H. Wright, Intonation and Dialogue Context as Constraints for Speech Recognition, *Language and Speech*, vol. 41, 1998, pp. 493-512.

[6] H. Hastie-Wright, M. Poesio and S. Isard. Automatically Predicting Dialogue Structure Using Prosodic Features, *Speech-Communication*, vol. 36, pp. 63-79.

[7] M. L. Flecha-Garcia, Eyebrow Raising and Communication in Map Task Dialogues, *Proceedings of the 1st Congress of the International Society for Gesture Studies*, Univ. of Texas at Austin, TX, USA, 2002.

[8] M. M. Louwerse, N. Benesh, M. E. Hoque, P. Jeuniaux, G. Lewis, J. Wu, M. Zirnstein (under review), Multimodal Communication in Face-to-face Conversations, *the 29th meeting of Cognitive Science Society*, Nashville, TN, 2007.

[9] E. Shriberg, R. Bates, A. Stolcke, P. Taylor, D. Jurafsky, K. N. C. Ries, R. Martin, M. Meteer, and C. Van Ess-Dykema, Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech, *Language and Speech*, vol. 41, 1998, pp. 439–487.

[10] J. Ang, Y. Liu and E. Shriberg. Automatic Dialog Act Segmentation and Classification in Multiparty Meetings, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, PA, USA, 2005.

[11] D. Jurafsky, R. Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, A. Stolcke, P. Taylor and C. Van Ess-Dykema, Automatic Detection of Discourse Structure for Speech Recognition and Understanding, *Proceedings of the 1997 IEEE Workshop on Speech Recognition and Understanding*, pp. 88–95, Santa Barbara, CA, USA.

[12] D. Surendran and G. Levow, Dialog Act Tagging with Support Vector Machines and Hidden Markov Models, *Proceedings of Interspeech*, Pittsburgh, PA, September, 2006.

[13] R. Fernandez and R. W. Picard, Dialog Act Classification from Prosodic Features Using Support Vector Machines, *Proceedings of Speech Prosody 2002*. Aix-en-Provence, France. April 2002.

[14] R. Kompe, Prosody in Speech Understanding Systems, Springer-Verlag New York, Inc. Secaucus, NJ, USA, 1997.

[15] C. W. Wightman and M. Ostendorf, Automatic Labeling of Prosodic Patterns, *IEEE Transactions on Speech and Audio Processing*, vol. 2, 1994, pp. 469–481.

[16] M. M. Louwerse, E. G. Bard, M. Steedman, X. Hu, and A. C. Graesser, Tracking Multimodal Communication in Humans and Agents. *Technical report*, Institute for Intelligent Systems, University of Memphis, Memphis, TN, 2004.

[17] M. M. Louwerse, P. Jeuniaux, M. E. Hoque, J. Wu, G. Lewis, Multimodal Communication in Computer-Mediated Map Task Scenarios. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pp. 1717-1722. Mahwah, NJ: Erlbaum, 2006.

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

[19] P. Boersma and D. Weenink, 2006, *Praat: doing phonetics by computer* (Version 4.5.15) [Computer program]. Retrieved February 12, 2007, from http://www.praat.org/.

[20] T. Hofmann, EECS Department, Computer Science Division, University of California, Berkeley & International Computer Science Institute, Berkley, CA, Probabilistic Latent Semantic Analysis – Uncertainty in Artificial Intelligence, UAI'99, Stockholm, 1999.

[21] M. E. Hoque, M. Yeasin, M. M. Louwerse, Robust Recognition of Emotion from Speech, *6th International Conference on Intelligent Virtual Agents*, Marina Del Rey, CA, August 2006.

[22] E. Brill, A Simple Rule-based Part of Speech Tagger, *Proceedings of the Third Annual Conference on Applied Natural Language Processing*, ACL, Morristown, NJ, USA, 1992.

[23] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Edition ed. San Francisco: Morgan Kaufmann, 2005.