Robust Recognition of Emotion from Speech

Mohammed E. Hoque[1], Mohammed Yeasin[1], Max M. Louwerse[2]

Department of Electrical and Computer Engineering / Institute for Intelligent Systems[1]
Department of Psychology / Institute for Intelligent Systems[2]
The University of Memphis
Memphis, TN 38152 USA
{mhoque, myeasin, mlouwerse}@memphis.edu

Abstract. This paper presents robust recognition of selected emotions from salient spoken words. The prosodic and acoustic features were used to extract the intonation patterns and correlates of emotion from speech samples in order to develop and evaluate models of emotion. The computed features are projected using a combination of linear projection techniques for compact and clustered representation of features. The projected features are used to build models of emotions using a set of classifiers organized in hierarchical fashion. The performances of the models were obtained using number of classifiers from WEKA tools. Results showed that the lexical information computed from both the prosodic and acoustic features at word level yielded robust classification of emotions.

Keywords: emotion recognition, prosody, speech, machine learning

1. Introduction

Animated conversational agents allow for natural multimodal human-computer interaction and have shown to be effective in intelligent tutoring systems [1, 2]. Agents used in intelligent tutoring are designed to articulate difficult concepts in a well paced, adaptive and responsive atmosphere based on the learners' affective emotional state of minds. Expert educators, both human and artificial, are expected to identify the cognitive states of mind of the learners' and take appropriate pedagogical actions [3]. Because of the realization that monitoring cognitive states in the student through the student's verbal feedback alone is not enough, research that focuses on monitoring of other modalities like speech has become more common [4, 5].

Recognizing emotion from speech has been an ongoing area of investigation by researchers mainly in domains such as call center environments. Dellaert *et al.* [6] attained an accuracy of 60-65% on distinguishing patterns among sadness, anger, happiness, and fear in the general domain of Human-Computer Interaction (HCI). The results were cross validated with three classifiers: the maximum likelihood Bayes classification, kernel regression, and k-nearest neighbor (K-NN) methods using the pitch contour features. Lee *et. al.* [7] tried to distinguish between two emotions: negative and positive, in call center environment, using linear discrimination, k-NN classifiers, and support vector machines (SVM) and achieved a maximum accuracy

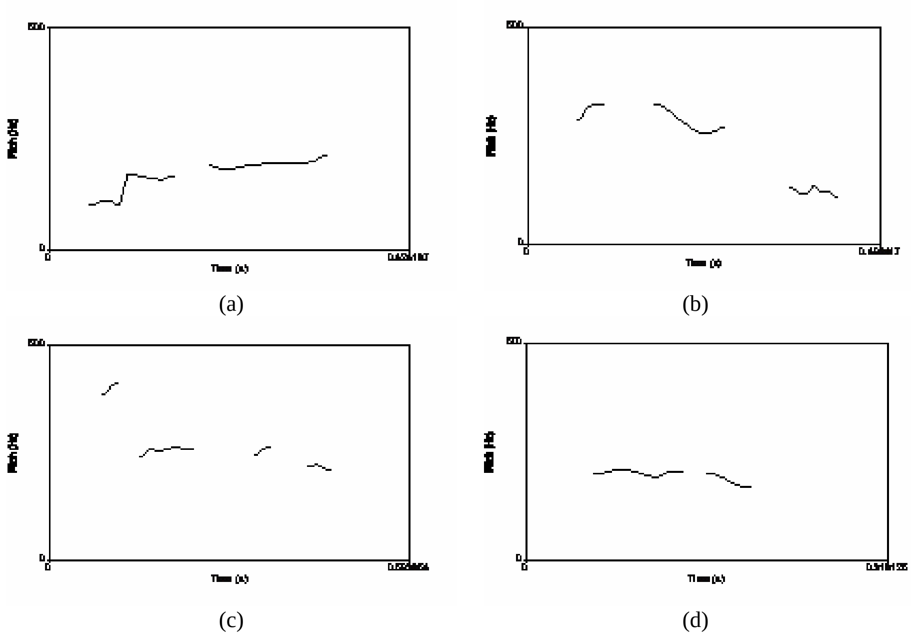

Figure 1. Pictorial depiction of the word "okay" uttered with different intonations to express different emotions. The pitch accent of various emotion such (a) confusion, (b) flow, (c) delight and (d) neutral.

rate of 75%. Paeschke [8] used a real time emotion recognizer using neural networks adopted for call center applications and reported 77% classification accuracy in two emotions: agitation and calm. It has been experimented and showed in [9], [10], how "quality features" (based on formant analysis) are used in addition to "prosody features", (particularly pitch and energy) to improve the classification of multiple emotions. This technique is known to exploit emotional dimension other than prosody. Yu *et. al.* [11] used SVMs, which are binary classifiers, to detect one emotion versus the rest. On four distinct emotions such as anger, happy, sadness, and neutral, they achieved an accuracy of 73%.

Robust recognition of emotion expressed in speech requires a thorough understanding of the lexical aspects of speech [12]. Lee et. al. hypothesized that a group of positive and negative words, which were confined to a call center environment, were related to different emotions. The occurrences of such predefined words were used to infer the emotional reaction of a caller using a probabilistic framework. Lee et al. argued that there is one-to-one correspondence between a word and a positive or negative emotion.

Though this may be true for some words, more commonly a word can convey different emotions by the use of different intonational pattern to. For example, the frequently used discourse marker "okay", is often used to express affirmation (S1 "Ready?" S2 "Okay"), but can also be used to express delight (S1 "So and that's how the procedure works" S2 "Okay!"), confidence (S1 "You're ready for the jump?"

“Okay”), or confusion (S1 “You just multiply by the divider” S2 “Okay…?”). The meaning of these different uses of “okay” may be guessed by the context, but their emotional value become clear in the intonational patterns that are used to express the word. Figure 1 shows that despite the fact that the word is the same, the intonational patterns are very different for different emotions. We therefore predict that lexical information extracted from combined prosodic and acoustic features that correspond to intonation pattern of “salient words” will yield robust recognition of emotion from speech, providing a framework for signal level analysis of speech for emotion.

To test this hypothesis, we selected word-level utterances from video data, from which features related to fundamental frequency (F0), energy, rhythm, pause and duration were extracted.

2. Proposed Approach

Our proposed approach consists of five major components (see Figure 2), namely, (i) collection of suitable data sets for training and testing, (ii) extraction of feature, (iii) projection of feature to lower dimensional space, (iv) learning the models using machine learning techniques and (v) evaluation of models. This paper presents a holistic approach in robust recognition of emotion from speech.

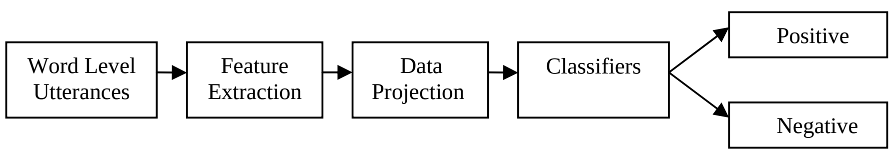

Figure 2. The high level description of the overall emotion recognition process.

First, a suitable database is captured for building and evaluating the models. Second, intonation patterns from spoken “salient words” are extracted with a combination of prosodic and acoustic features. Third, the extracted features are projected onto the lower dimensional space using combined Principle Component Analysis (PCA) [13] and Linear Discriminant Analysis (LDA) for a compact and clustered representation of computed features. Fourth, a set of machine learning techniques from the WEKA [14] toolbox are used to learn the models from the training samples. Finally, testing samples are used to evaluate the performances of models. Subsequent subsection describes the details of various components of robust recognition of emotion from speech.

2.1. Database and Preparation

Collecting large databases of natural and unbiased emotion is challenging. One needs a representative data set to infer various emotions from speech using machine learning technique to establish the hypothesis and to obtain meaningful results. The performance of a classifier that can distinguish different emotional patterns ultimately depends on the completeness of the training and testing samples and how similar it is compared to the real world data.

The data captured to perform experiments can be categorized into three methods depending on how they are captured. The first method employs actors to utter various or similar sentences in various feigned emotional patterns. The second method utilizes a system that interacts with a human subject and draws him/her to an emotional point and records the response. The third approach is to extract real life human utterances, which express various natural emotions.

The main drawback of having actors expressing emotional utterance is that the utterances are generally independently acted out in a laboratory setting. These data may converge very well, but may not suitable for real life human-computer interaction settings. On the other hand, setting up an experiment where individuals interact with computers or other individuals is expensive and time consuming. In our experiment, emotional utterances were clipped from movies. Though it is true that emotions are still "acted out", the discourse context and the absence of a lab setting makes it more natural than the first method. Three movies from which the utterances were taken were "Fahrenheit 911", "Bowling for Columbine" and "Before Sunset". "Fahrenheit 911" and "Bowling for Columbine" are political documentaries with many cases of positive and negative emotions. "Before Sunset" is a chatty romantic movie with delightful, frustrating and confusing expressions with minimal background music. Fifteen utterances were selected for four emotion categories: confusion/uncertain, delight, flow (confident, encouragement), and frustration [3]. Utterances that were selected were stand-alone expressions in conversations that had an ambiguous meaning, dependent on the context. Examples are "Great", "Yes", "Yeah", "No", "Ok", "Good", "Right", "Really", "What", "God". Three graduate students listened to the audio clips and successfully distinguished between the positive and negative emotions 65% of the time without specific instructions as to what intonation patterns to listen to. A hierarchical classifier was designed to first distinguish between positive (delight and flow) and negative (confusion and frustration) emotions. The same set of classifiers were applied again on positive and negative emotions separately to differentiate between delight and flow under positive emotion and confusion and frustration under negative emotion as shown in a Figure 3.

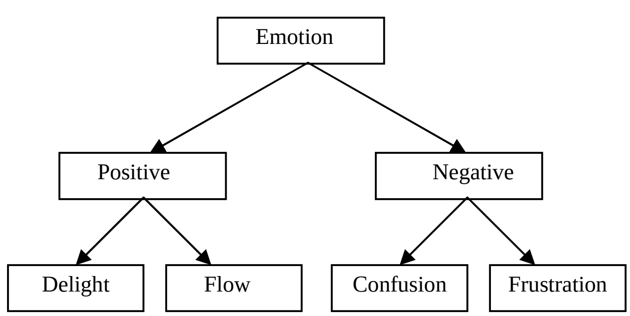

Figure 3. The design of the hierarchical binary classifiers.

2.2. Emotion Models using Lexical Information

To compute the lexical information from spoken salient words, 22 acoustic and prosodic features related to segmental and suprasegmental information believed to be correlates of emotion were calculated. Computed features were utterance level statistics related to fundamental frequency (F0) [15-17]. Other features were related to duration, intensity, and formants. In particular, the following features were computed for developing the models.

1. **Pitch:** Minimum, maximum, mean, standard deviation, absolute value, quantile, ratio between voiced and unvoiced frames.
2. **Duration**: ε_{time} ε_{height}
3. **Intensity**: Minimum, maximum, mean, standard deviation, quantile.
4. **Formant**: First formant, second formant, third formant, fourth formant, fifth formant, second formant / first formant, third formant / first formant
5. **Rhythm**: Speaking rate.

The speech processing software Praat [18] was used to calculate the features in batch mode. ε_{time}, ε_{height} features, which are part of duration, are prominence measures.

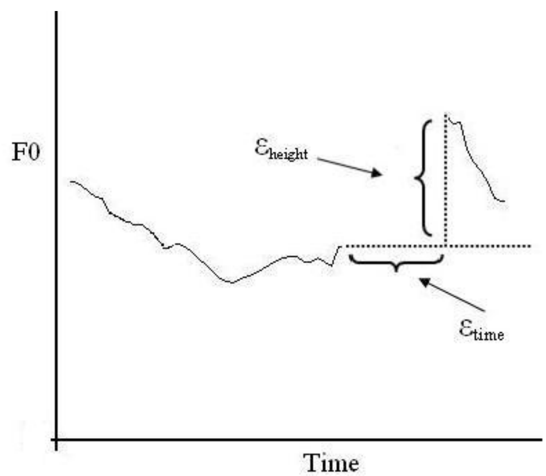

Figure 4. Measures of F0 for computing parameters (εtime, εheight) which corresponds to rising and lowering of intonation.

ε_{height} and ε_{time} features are related to phenomenon when fundamental frequency breaks down in word levels. ε_{time} refers to the pause time between two disjoint segments of F0 (often referred as Pitch), whereas ε_{height} refers to the vertical distance between the segments symbolizing voice breaks as shown in Figure 4. Inclusion of *height* and *time* accounts for possible low or high pitch accents. The frequency shift between the segments was selected rather than absolute measures to take into account the discourse [19].

Empirical studies [12] have demonstrated that not all base acoustics correlates mentioned above are equally useful in emotion recognition. Therefore, there is a need to reduce the feature space to get rid of the redundancies. This may in fact work better as the de-correlated data are projected into lower dimension to maximize the separation between emotion classes. In this experiment the combination of data projection techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) were used to de-correlate the data and then project them into lower dimensions. Based on the order and combination of data projection techniques, five stand-alone models were created which were separately tested against a set of 21 different classifiers as shown in Table 1. The first model fed the raw 22 features directly into the classifier. The second and the third model applied PCA on the raw features and took the first 15 (F15) and 20 (F20) eigenvectors respectively to de-correlate the base features. In the fourth model, LDA is directly used on the raw features to project them directly onto lower dimension. The fifth model consisted of the combination of PCA (F15) and LDA. The sequence of how the feature extraction, data projection and classification took place is shown in Figure 2. The following classifiers, shown in table 1, were carefully selected from the WEKA toolbox [14] to recognize the emotional patterns and then validate them. A 10-fold cross validation technique was used whereby the training data was randomly split into ten sets, 9 of which were used in training and the 10^{th} for validations. Then iteratively another nine was picked and so forth

TABLE 1. THE LIST OF CLASSIFIERS USED TO VALIDATE THE ROBUSTNESS OF THE ALGORITHM USING WEKA TOOLBOX.

Types of Classifiers				
Rules	**Trees**	**Meta**	**Functions**	**Bayes**
Part	RandomForrest	AdaBoostM1	Logistic	Naïve Bayes
NNge	J48	Bagging	Multi-layer Perceptron	Naïve Bayes Simple
Ridor	Logistic Model Tree	Classification via Regression	RBF Network	Naïve Bayes Updateable
-	-	LogitBoost	Simple Logistics	-
-	-	Multi Class Classifier	SMO	-
-	-	Ordinal Class Classifier	-	-
-	-	Threshold Selector	-	-

2.3. Results and Discussion

Results showed that the combination of data projection techniques such as PCA and LDA yielded better performance as opposed to using raw features or using LDA or PCA alone (Table 2). An average of 83.33 % accuracy was achieved using the combination of PCA and LDA. On the other hand, features like PCA (F15), PCA (F20), LDA resulted in accuracy rates of respectively 50.79%, 57.1%, 61%, and 52.01% on average. The performance of combining PCA and LDA is higher than PCA or LDA itself mainly because PCA de-correlates the data, whereas LDA projects the data onto lower dimension. Therefore, the combination of PCA and LDA is expected to work better.

When the same models were applied to positive emotions and negative emotions even more impressive results emerged (Table 3). Positive emotions were collapsed into delight and flow and negative emotions were collapsed into confusion and frustration using the learning phases introduced by [3]. The performance of the diverse set of classifiers to recognize negative emotions is better than the performance to recognize positive emotions. One potential explanation for this is that negative emotions may deviate more from the standard than positive emotions. In other words, positive emotions may in general less likely be recognized as an emotion, because they map onto the default. Negative emotions on the other hand deviate from that default, thereby facilitating recognition, both in humans and computers.

TABLE 2. SUMMARY OF CLASSIFICATION RESULTS FOR 21 SELECTED CLASSIFIERS

Category	**Classifiers**	**Accuracy (%)**				
		Features (a)	**PCA (b)**		**LDA (c)**	**PCA+LDA (d)**
			F15 (b1)	**F20 (b2)**		
Rules	Part	50	66.67	66.67	47.61	83.33
	NNge	33.33	33.33	38.09	38.09	83.33
	Ridor	66.67	83.33	100	47.20	66.67
Trees	Random Forrest	50	50	50	66.67	83.33
	J48	50	66.67	66.67	47.61	83.33
	Logistic Model Tree	33.33	47.61	83.33	66.67	71.67
Meta	AdaBoostM1	61.90	71.42	71.42	42.85	61.90
	Bagging	33.33	66.67	83.33	42.85	66.67
	Classification via Regression	50	66.67	66.67	47.61	83.33
	Logit Boost	50	50	61.90	52.38	83.33
	Multi Class Classifier	50	42.85	52.38	57.14	83.33
	Ordinal Class Classifier	50	66.67	66.67	47.62	83.33
	Threshold Selector	50	66.67	66.67	61.90	100
	Logistic	50	42.85	57.38	57.14	83.33

Functions	Multi-layer Perceptron	50	57.14	52.38	50	83.33
	RBF Network	33.33	66.67	52.38	38.09	83.33
	Simple Logistics	33.33	47.61	83.33	66.67	66.67
	SMO	71.42	57.14	61.90	52.38	71.42
Bayes	Naïve Bayes	66.67	50	33.33	52.38	66.67
	Naïve Bayes Simple	66.67	50	33.33	57.14	66.67
	Naïve Bayes Updateable	66.67	50	33.33	52.38	66.67

Note. (a) raw features are used into classifiers, (b1) using the first 15 (f15) eigenvectors of PCA into the classifiers, (b2) using the first 20 (f20) eigenvectors of PCA into the classifiers. (c) using LDA to project the data into lower dimension and then use them into the classifiers. (d) combination of both PCA and LDA to not only de-correlate the data redundant feature space, but also to project them into lower dimension and then use them into the classifiers.

TABLE 3. SUMMARY OF CLASSIFICATION RESULTS FOR 21 CLASSIFIERS ON POSITIVE AND NEGATIVE EMOTIONS.

Category	**Classifiers**	Accuracy (%)	
		Delight + Flow	Confusion + Frustration
Rules	Part	72.72	100
	NNge	80	100
	Ridor	66.67	100
Trees	RandomForrest	63.63	66.67
	J48	72.72	100
	LMT	72.72	100
Meta	AdaBoostM1	54.44	100
	Bagging	63.64	66.67
	Classification via Regression	72.72	100
	LogitBoost	63.64	100
	Multi Class Classifier	72.72	100
	Ordinal Class Classifier	72.72	100
	Threshold Selector	83.33	100
Functions	Logistic	72.72	100
	Multi-layer Perceptron	66.67	100
	RBF Network	66.67	100
	Simple Logistics	72.72	100
	SMO	72.72	100
Bayes	Naïve Bayes	72.72	100
	Naïve Bayes Simple	72.72	100
	Native Bayes Updateable	72.72	100

Note. Results with the combination of PCA + LDA were only recorded as they comparatively produce better results as shown in Table 2.

3. Conclusion

Automatic recognition of emotion is gaining attention due to the widespread applications into various domains, including those with animated conversational agents. Automated recognizing emotion with high accuracy still remains an elusive goal due to the lack of complete understanding and agreement of emotion in human minds. The experiment presented in this paper achieved an average of 83.33% success rate of defining positive and negative emotion using a varied set of classifiers confined to learning environment. Lexical and prosodic features were used on word level emotional utterances to improve the performance the emotion recognition system. Our results indicate that using a proper set of projection techniques on word level lexical and prosodic features yields accuracy rate of 80 to 100%. It is worth noting that the datasets were tested by three graduate students who were able to classify the emotions into correct bins 65% of the time. This supports our hypothesis that word level prosodic and lexical features provide useful clues about positive and negative emotions. This hypothesis also enables us to have a framework for signal level analysis.

We are of course aware of the risk that clipping arbitrary words from a conversation may be ineffective at various cases as some words may convey more in context only. Therefore, our goal for the immediate future is to look at meaningful words in a sequence while introducing context in our analysis as well. A research project that investigates multimodal communication (prosody, dialog structure, eye gaze and facial expressions) in Map Task scenarios will thereby generate the needed data [5, 20]. In the second phase of this project the results of the data analysis will allow us to develop an animated conversational agent that uses the right intonational contours in the right contexts, expressing the right emotions.

Psychologists have argued that visual information modifies the perception of speech [21]. Also, combination of visual and audio information provides robust performance when modalities are captured in noisy environment [22]. Therefore, in order for our agent to be successful in learning environment, it is imperative that the agent should be able to fuse the audio and video data to reach a decision regarding the emotional states of the learners. Therefore, our future efforts will include fusion of video and audio data in a signal level framework to boast the performance of our existing emotion recognition system.

4. Acknowledgements

This research was partially supported by grant NSF-IIS-0416128 awarded to the third author. 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.

References

[1] A. C. Graesser, K. VanLehn, C. Rose, P. Jordan, and D. Harter, "Intelligent tutoring systems with conversational dialogue.," *AI Magazine*, vol. 22, pp. 39-51, 2001.

[2] M. M. Louwerse, A. C. Graesser, S. Lu, and H. H. Mitchell, "Social cues in animated conversational agents," *Applied Cognitive Psychology*, vol. 19, pp. 1-12, 2005.

[3] B. Kort, R. Reilly, and R. W. Picard, "An Affective Model of Interplay Between Emotions and Learning: Reengineering Educational Pedagogy-Building a Learning Companion.," presented at In Proceedings of International Conference on Advanced Learning Technologies (ICALT 2001), Madison, Wisconsin, August 2001.

[4] S. K. D'Mello, S. D. Craig, A. Witherspoon, J. Sullins, B. McDaniel, B. Gholson, and A. C. Graesser, "The relationship between affective states and dialog patterns during interactions with AutoTutor.," presented at Proceedings of the World Conference on E-learning in Corporate, Government, Health Care, and Higher Education, Chesapeake, VA, 2005.

[5] M. Louwerse, P. Jeuniaux, M. Hoque, J. Wu, and G. Lewis, "Multimodal Communication in Computer-Mediated Map Task Scenarios.," presented at The 28th Annual Conference of the Cognitive Science Society, Vancouver, Canada, July 2006.

[6] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing Emotion in Speech.," presented at Proceedings of the ICSLP, 1996.

[7] C. Lee, S. Narayanan, and R. Pieraccini, "Classifying Emotions in Human-Machine Spoken Dialogs," presented at Proc. of International Conference on Multimedia and Expo, Lausanne, Switzerland, August 2002.

[8] A. Paeschke and W. F. Sendlmeier, "Prosodic Characteristics of Emotional Speech: Measurements of Fundamental Frequency Movements," presented at Proceedings of the ISCA-Workshop on Speech and Emotion, 2000.

[9] R. Tato, R. Santos, R. Kompe, and J. M. Pardo, "Emotional Space Improves Emotion Recognition," presented at Proc. Of ICSLP-2002, Denver, Colorado, September 2002.

[10] S. Yacoub, S. Simske, X. Lin, and J. Burns, "Recognition of Emotions in Interactive Voice Response Systems.," presented at the 8th European Conference on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland, 1-4 September 2003.

[11] F. Yu, E. Chang, Y. Q. Xu, and H. Y. Shum, "Emotion Detection From Speech To Enrich Multimedia Content.," presented at the Second IEEE Pacific-Rim Conference on Multimedia, Beijing, China, October 24-26, 2001.

[12] C. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE transaction on speech and audio processing*, vol.13, 2005.

[13] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[14] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, 2nd Edition ed. San Francisco: Morgan Kaufmann, 2005.

[15] C. E. Williams and K. N. Stevens, "Emotions and speech : Some acoustical correlates.," *JASA*, vol. 52, pp. 1238-1250, 1972.

[16] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression.," *J. Personality and Social Psychology*, vol. 70, pp. 614–636, 1996.

[17] S. Mozziconacci, "The expression of emotion considered in the framework of an intonational model," *Proc. ISCA Wrksp. Speech and Emotion*, pp. 45-52, 2000.

[18] P. Boersma and D. Weenink, " Praat: doing phonetics by computer," Version 4.4.16 ed, 2006.

[19] S. Kettebekov, M. Yeasin, and R. Sharma, "Prosody-based Audio Visual co-analysis for co-verbal gesture recognition," *IEEE transaction on Multimedia*, vol. 7, pp. 234-242, 2005.

[20] M. M. Louwerse, E. G. Bard, M. Steedman, X. Hu, and A. C. Graesser, "Tracking multimodal communication in humans and agents.," Institute for Intelligent Systems, University of Memphis, Memphis, TN., 2004.

[21] D. W. Massaro, "Illusions and Issues in Bimodal SpeechPerception.," presented at Proceedings of Auditory Visual Speech Perception ’98., Terrigal-Sydney Australia, December, 1998.

[22] M. Pantic and L. J. M. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction.," *Proceedings of the IEEE*, vol. 91, pp. 1370 – 1390, Sept. 2003.