Toward Detecting Emotions in Spoken Dialogs

Chul Min Lee, *Student Member, IEEE,* and Shrikanth S. Narayanan, *Senior Member, IEEE*

***Abstract*—The importance of automatically recognizing emotions from human speech has grown with the increasing role of spoken language interfaces in human-computer interaction applications. This paper explores the detection of domain-specific emotions using language and discourse information in conjunction with acoustic correlates of emotion in speech signals. The specific focus is on a case study of detecting *negative* and *non-negative* emotions using spoken language data obtained from a call center application. Most previous studies in emotion recognition have used only the acoustic information contained in speech. In this paper, a combination of three sources of information—acoustic, lexical, and discourse—is used for emotion recognition. To capture emotion information at the language level, an information-theoretic notion of *emotional salience* is introduced. Optimization of the acoustic correlates of emotion with respect to classification error was accomplished by investigating different feature sets obtained from feature selection, followed by principal component analysis. Experimental results on our call center data show that the best results are obtained when acoustic and language information are combined. Results show that combining all the information, rather than using only acoustic information, improves emotion classification by 40.7% for males and 36.4% for females (linear discriminant classifier used for acoustic information).**

***Index Terms*—Acoustic correlates, dialog systems, emotion recognition, emotional salience, feature selection, information fusion, principal component analysis, spoken language processing.**

I. Introduction

RESEARCH in understanding and modeling human emotions, a topic that has been predominantly dealt with in psychology and linguistics, is increasingly attracting attention within the engineering community. A major motivation comes from the desire to develop human machine interfaces that are more adaptive and responsive to a user's behavior. There is an increasing need to know not only what information a user conveys but also how it is being conveyed. Research by psychologists and neuroscientists has shown that emotion is closely related to decision-making [1] and thus, emotion plays a significant role in the rational actions of human beings. Given the importance of emotions in human communication and decision-making, it is desirable that intelligent human-machine interfaces be able to accommodate human emotions in an appropriate way. Researching emotion, however, is extremely challenging in several respects. One of the main difficulties results from the fact that it is difficult to define what emotion means in a precise way. There are ongoing debates concerning how many emotion categories exist, how to reconcile long-term properties such as moods with short-term emotional states such as full blown emotions, and debate as to how to seek measurable correlates of emotions. Hence, an engineering approach to emotion invariably has to rely on a number of assumptions so as to bound the problem for tractability.

This paper focuses on recognizing emotions from spoken language. The importance of emotion recognition from human speech has increased significantly with the need to improve both the naturalness and efficiency of spoken language human-machine interfaces [2]. For instance, automatic dialog systems with the ability to recognize emotions can respond to callers according to the detected emotional state or they can pass control over to human operators. Automatic emotion recognizers can be viewed as systems that assign category labels to emotional states. Emotion recognition in spoken dialogs not only requires signal processing and analysis techniques, but also incorporates psychological and linguistic analyzes of emotion.

While, in general, cognitive theory in psychology argues against categorical labeling from just physiological features [3]–[5], it provides a pragmatic starting point, especially from an engineering perspective. The primary reasons for this are: 1) the lack of a definite description and agreement on a set of basic emotions [3], [6], 2) the lack of consistency in the definitions of emotions where the same emotional category tends to be described in the literature in different ways [2]. Hence, focusing on the archetypal emotions—happiness, sadness, fear, anger, surprise, and disgust—is typically justified as a way to arrive at finer distinctions. For example, Scherer explored the existence of a universal psychobiological mechanism of emotion in speech by studying the recognition of five emotions in nine languages, obtaining 66% of overall accuracy [7].

While the ability to recognize a large variety of emotions is attractive, it may not be necessary or practical in the context of developing algorithms for conversational interfaces. Based on this assumption, in this paper, we favor the notion of application-dependent emotions and thus, examine what is a reduced space of emotions. In particular, we focus on recognizing *negative* and *non-negative* emotions from speech signals using data derived from a commercially deployed automatic call center dialog system. The reason for this is that the detection of negative emotions can be used as a strategy to improve the quality of the service in automated call center applications. Most previous efforts involving emotion recognition from speech have been limited to acoustic information [8]–[11]. In the context of a conversational interface, it is, however, possible to combine lexical, semantic and discourse information in such a way that emotion recognition is maximized. This paper attempts to combine various aspects of spoken language information—acoustic, lexical, and discourse—so as to detect the user's emotional state.

Manuscript received November 22, 2003; revised February 26, 2004. This research was supported in part by funds from the NSF through the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, Cooperative Agreement under Contract EEC-9529152, a CAREER award, and the Department of the Army under Contract DAAD 19-99-D-0046. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. C.-C. Jay Kuo.

The authors are with the Department of Electrical Engineering and IMSC, University of Southern California, Los Angeles, CA 90089 USA (e-mail: cml@sipi.usc.edu; shri@sipi.usc.edu).

Digital Object Identifier 10.1109/TSA.2004.838534

Several pattern recognition methods, using the acoustic signal information, have been explored for automatic emotion recognition from speech. For example, Dellaert *et al.* used maximum likelihood Bayes classification, kernel regression, and k-nearest neighborhood methods [10], whereas, Roy *et al.* used Fisher linear discrimination methods [11]. Petrushin developed a real-time emotion recognizer using an ensemble of neural networks for a call center application [9] and was able to achieve 77% classification accuracy for two emotional states, *agitation* and *calm*, with eight features chosen by a feature selection method.

A variety of acoustic features have also been explored. For example, Banse *et al.* examined acoustic profiles or vocal cues for emotion expression using actors voices for fourteen emotion categories [12]. The acoustic parameters Banse *et al.* used were related to fundamental frequency/pitch (F0), energy, speech rate, and spectral information in voiced and unvoiced portions. Murray *et al.* provided a summary of acoustic correlates for archetypal emotions from the viewpoint of speech synthesis [13]. These acoustic correlates include the following:

1) pitch-related features;
2) formant frequencies;
3) timing features;
4) voice-quality parameters;
5) articulation parameters.

A comprehensive summary of qualitative acoustic correlates for the six archetypal emotions (anger, happiness, sadness, fear, grief, surprise) is provided in Cowie *et al.* drawn from a vast body of literature. Cowie *et al.* summarized as follows:

1) voice level: signal amplitude, energy and duration have been shown to be reliable measures of voice level;
2) voice pitch;
3) phrase, word, phoneme, and feature boundaries;
4) temporal structure which refers to measures at the pitch contour level and related structures in the intensity domain.

Also, McGilloway *et al.* studied thirty two different acoustic features for the classification of five emotion states [14]. The acoustic features included those related to F0 (usually regarded as pitch), energy, duration, and tune (segments of the pitch contour bounded at either end by a pause of 180 ms or more). Benchmark classification was done on those features and highly contributing acoustic correlates were found using linear discriminant analysis. Data was recorded by forty readers, and "number of F0 points recovered," "inter-quartile range for intensities at minima," and "median of silence durations" were found to be the best among thirty two features that were considered.

Most of the previously reported studies used speech recorded from actors who were asked to express (or feign) prescribed emotions. Furthermore, most of these utterances were produced in isolation, i.e., not in any conversational context. One notable exception is the study by Batliner *et al.* [15]. In this work, a *Wizard-of-Oz* scenario was used to collect data. Subjects assumed that they were communicating with a real computer although it was actually operated by a human behind the scenes. The study reported classification of utterances into two categories: *emotional* and *neutral*. The authors used details about topic repetition as their *discourse* information to improve emotion recognition accuracy. A more detailed analysis of discourse markers related to frustration and politeness (for example, the use of swear words, negation, and the repetition of the same sub-dialog) using a child-machine interaction corpus was reported in [16]. More recently, a study by Ang *et al.* [17] explored the detection of user frustration using a number of dialog level features based on the DARPA Communicator corpus. In our paper, a corpus of utterances obtained from a commercially-deployed human-machine spoken dialog application was used. In summary, these investigations show the promise of using higher level linguistic information for emotion recognition.

The availability of a constrained-domain dialog application, such as an automated call center, provides the possibility of utilizing spoken language information along a number of dimensions such as through the use of acoustic, lexical (word choices), and discourse correlates of emotions. Leveraging previously published results, a number of acoustic correlates are considered in this paper and are reconciled in a systematic way using feature selection and feature reduction methods. For *language* information, a novel method for estimating the emotion information conveyed by words (and by sequences of words) is proposed. It is well known that people tend to use specific words to express their emotions based on associations they have learned between certain words and the corresponding emotions. In this regard, for example, psychologists have tried to identify the language of emotions through careful experimentation by asking people to list words that describe specific emotions [6]. Such results are useful for identifying emotional keywords in a generic way but do not translate directly to the problem at hand. Our interest is in associating emotions to words in spoken language and it is highly domain and situation dependent. We obtained the emotional keywords in our data by automatically calculating *emotional salience* of the words in the data corpus. Emotional salience is a measure of how much information a word provides about a given emotion category. The salience of a word in emotion recognition can be defined as mutual information between a specific word and emotion category. Similar ideas have been used in natural language acquisition for associating words to meanings [18].

Discourse information of emotion recognition has been combined with acoustic correlates to improve the overall performance of emotion classification [15], [17]. In [17], repetition or correction information was used for the discourse information. Batliner *et al.* also adopted repetition as their discourse information [15]. In this work, we separated users' responses to the system in human-computer communication into five possible categories: *rejection*, *repetition*, *rephrase*, *ask-start over*, and *none of the above*. Under the label *none of the above*, most utterances corresponded to speech acts of the response to system information requests such as providing person or place names (in this corpus). As expected, the occurrence of each discourse label is biased according to the emotion categories. For example, *rejection* has been found more in negative emotion utterances.

Given these various information sources, the question that remains is how to combine these information sources so as to enable emotion recognition. One strategy is information fusion at the feature level by building large dimensional input features [17]. This method suffers from potential dimensionality issues in regards to classification with increasing feature sizes. An alternative strategy, and the one adopted in this paper, is to combine the various sources of information (acoustic, lexical, and discourse information in our case) at the decision level. For the classification of the individual data streams, we investigated both linear discriminant classifiers (LDC) with Gaussian class-conditional probability and k-nearest neighborhood classifiers (k-NN).

The rest of the paper is organized as follows: Section II describes the speech data corpus adopted for the experiments. Section III discusses the acoustic features used, and methods for feature selection, including principal component analysis (PCA). The discussion on language and discourse information is provided in Section IV, and the scheme for combining information sources is in Section V. Experimental results are in Section VI and conclusions are given in Section VII.

II. Database and Preparation

Developments in automatic speech recognition have benefited immensely from real-world data. Given the complexity of the definition and the range of emotion categories, the problem related to data concerns how to obtain the required amount of realistic data to do research that yields meaningful results and algorithms. Most studies in emotion recognition in speech have used actors' voices where actors are asked to read/speak given sentences usually designed to have emotionally-neutral semantic content, with pre-specified emotions. Since those data sets are limited to utterances for archetypal emotions not in a dialog context, results based on them may not be generalized to human-machine interaction scenarios. On the other hand, real data suffers from potential coverage problem, i.e., we need vast amounts of data characterizing various emotion types, and from a large number of users and contexts, to design valid models and algorithms. Our limited-domain approach allows in-depth focus on a finite set of emotions using significant amounts of data obtained from realistic human-machine interactions.

A. Database

The speech data (8 kHz, 8 bit mu-law compressed) used in the experiments were obtained from real users engaged in spoken dialog with a machine agent over the telephone using a commercially-deployed call center application. Since most dialog turns in this corpus had one utterance, each turn is considered to be an utterance in this work. The speech database used for our experiments contained 1187 calls, each having an average of six utterances; the total number of utterances was approximately 7200.

B. Data Preparation

The original usage data corpus was comprised of calls in the order of thousands with only a fraction representing potentially negative emotions. Hence, this required some automatic preprocessing to narrow down data of interest for emotion recognition research and development. The first step was to mine this data using objective measures such as ASR accuracy, total number of dialog turns, and rejection rate to narrow down the inventory to potentially useful dialogs for our experiments. This was followed by subjective tagging of the data into one of two possible emotion categories—negative and non-negative—by four different human listeners. In our study, *negative* emotions represent anger and frustration in human speech, whereas *non-negative* emotions represent its complement, i.e., they represent neutral or positive emotions such as happiness or delight. The order of utterances was randomly chosen in order for listeners not to be influenced to guess the emotions by the situation in the dialogs (thus minimizing the effect of discourse context). After administering the human listening tests, it turned out that most *non-negative* emotion utterances were neutral in nature, i.e., they had no apparent display of emotions.

TABLE I
Kappa Statistics and Normal Test (Z-Test) for Female and Male Data

	Female	**Male**
Kappa statistic, K	0.454	0.477
Normal test, Z	2.48	2.01

To measure the amount of agreement among the taggers, the *kappa statistic* was used [19], [20]. The kappa statistic provides a measure of agreement for categorical variables in subjective tests. The kappa coefficient, *K*, is the ratio of the proportion of times that the coders/taggers agree (corrected for chance agreement) to the maximum proportion of times that the coders could agree

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{1}$$

where $P(A)$ is the proportion of times that the k coders agree and $P(E)$ is the proportion of times we would expect the k coders to agree by chance. If there is complete agreement among the coders, then $K = 1$; whereas, if there is no agreement (other than the agreement which would be expected to occur by change) among the coders, then $K = 0$.

The results of the values of kappa statistic, K, for both female and male are shown in Table I. From these results, we can conclude that there is moderate agreement among the taggers (Ang *et al.* obtained similar kappa statistic values for the DARPA Communicator data [17]). Such results might stem from the fact that it is also difficult for humans to distinguish emotions from speech. To see whether these results represent a significant difference from 0, i.e., the agreements by chance, we did a hypothesis test with the null hypothesis of $H_0 : K = 0$ against $H_1 : K > 0$. It has been known that K is approximately normally distributed with zero mean for a large number of samples and thus we can perform normal test (Z test). The results are shown in the second row in Table I and they exceed the $\alpha = 0.05$ significance level (where $Z = 1.96$). Therefore, we can conclude that the tagging exhibits a significant difference from the agreements by chance. After the database preparation, we obtained 776 utterances for female speakers with 575 non-negative and 201 negative utterances and 591 for male (452 non-negative and 139 negative emotion-tagged utterances).

III. Acoustic Information

In this work, we considered 21 different acoustic correlates related to both segmental and suprasegmental information from speech signals. These features were utterance-level statistics corresponding to fundamental frequency (F0), energy, duration, and the first and second formant frequencies.

1) **Fundamental Frequency (F0)** mean, median, standard deviation, maximum, minimum, range (max–min), and linear regression coefficient.
2) **Energy** mean, median, standard deviation, maximum, minimum, range, and linear regression coefficient.
3) **Duration** speech-rate, ratio of duration of voiced and unvoiced region, and duration of the longest voiced speech
4) **Formants** first and second formant frequencies (F1, F2), and their bandwidths (BW1, BW2). Mean of each feature was calculated

For speech-rate, the average length of voiced portion of speech was calculated. The linear regression coefficients for F0 were also calculated in the voiced portion of speech describing the lines that fit the pitch contour. For the linear regression coefficients of energy, the nonzero valued portion of energy was used. These are referred to as base acoustic features (correlates) since they provide the starting point for the study. Note that all the samples were normalized, i.e., the origin was shifted to the means of the features, and the variance of all features was scaled to 1.

A. *Feature Selection*

Largely motivated by proposals in the published literature, all of the base acoustic correlates summarized above may not be equally useful and important for emotion recognition thereby creating the need for systematic feature selection. The rationale for feature selection is that new or reduced features might perform better than the base features because we can eliminate irrelevant features from the base feature set. This can also reduce the dimensionality, which can otherwise hurt the performance of the pattern classifiers. In this study, we used the forward selection (FS) method. First, FS initializes to contain the single best feature with respect to a chosen criterion from the whole feature set. Here, classification accuracy criterion by *nearest neighborhood* rule is used, and the accuracy rate is estimated by *leave-one-out* method. The subsequent features are added from the remaining features which maximize the classification accuracy. Stopping rule is that when the number of features added reaches the pre-set number, the selection stops.

In this study, we experimented with two sets of rank-ordered selected features. The first one had 10 best features (f10), and the other, 15 best features (f15). The best 15 chosen features for the male and female data were

Male: Ratio of duration of voiced and unvoiced region, energy standard deviation (STD), energy median, energy STD, F0 regression coefficient, F0 median, energy regression coefficient, energy max, energy min, energy range, duration of the longest voiced speech, F0 mean, first formant bandwidth, F0 max, and second formant bandwidth.

Female: Ratio of duration of voiced and unvoiced region, energy median, F0 regression coefficient, speech rate, energy min, duration of the longest voiced speech, energy regression coefficient, F0 median, F0 mean, F1, energy mean, energy max, F0 max, energy range, energy STD.

First, note that both male and female data have similar features in their best feature sets. Ratio of duration of voiced and unvoiced region, energy median, and F0 regression coefficient are included in the five-best features for both genders.

Along with the feature sets obtained by FS, a feature set calculated by principal component analysis (PCA) was obtained [21]. PCA involves computation of the sample $d \times d$ covariance matrix Σ of the full feature set with d-dimensions, calculation of the eigenvalues and eigenvectors of Σ, and finally sorting it according to decreasing order of eigenvalues. Then, the k eigenvectors corresponding to the largest eigenvalues are chosen to form a $k \times k$ matrix A whose columns consist of the k eigenvectors. We can obtain a new feature set by preprocessing features according to

$$\boldsymbol{x} = A^T(\boldsymbol{x} - \mu) \tag{2}$$

where μ is the mean vector for $\boldsymbol{x}$. Obviously, the feature set after PCA is different from the base feature set since it is located in the projected feature space, and the dimension of the features can usually be reduced.

IV. Language and Discourse Information

Other sources of information considered in this study are *language* and *discourse* related information. From the data corpus we have, it is apparent that people tend to use specific word choices for expressing their emotions. In fact, while listening to the data which was use to tag the emotion classes, the listeners reported that they tended to feel negative emotions if they heard certain words in the utterances, e.g., "no" or swear words. People tend to use certain words more frequently in expressing their emotions because they have learned the connection between the certain words and the related emotions. This is a topic well-studied in psychology [6]. The question then becomes how to automatically learn and associate emotional information with words (lexical items) that come from data. To make that possible, we propose the notion of emotional salience.

A. *Emotional Salience*

The idea here is to automatically learn and spot "emotionally salient words" so as to improve the recognition of emotions. To identify emotionally salient words in the utterances, we adopted the information-theoretic concept of "emotional salience." An emotionally salient word with respect to a category is one which appears more often in that category than in other categories. We used the salience measure to find and associate words that are related to emotions in the speech data.

In calculating emotional salience, first we denote the words in the utterances by $W = \{v_1, v_2, \ldots, v_n\}$ and the set of emotion classes by $E = \{e_1, e_2, \ldots, e_k\}$ (here $k = 2$, negative and non-negative), and then the self mutual information is given by [22]

$$i(v_n, e_k) = \log \frac{P(e_k \mid v_n)}{P(e_k)} \tag{3}$$

where $P(e_k \mid v_n)$ is the posterior probability that an utterance containing word v_n implies emotion class e_k, and $P(e_k)$ denotes the prior probability of that emotion. We can see that if the word v_n in an utterance highly correlates to an emotion class,

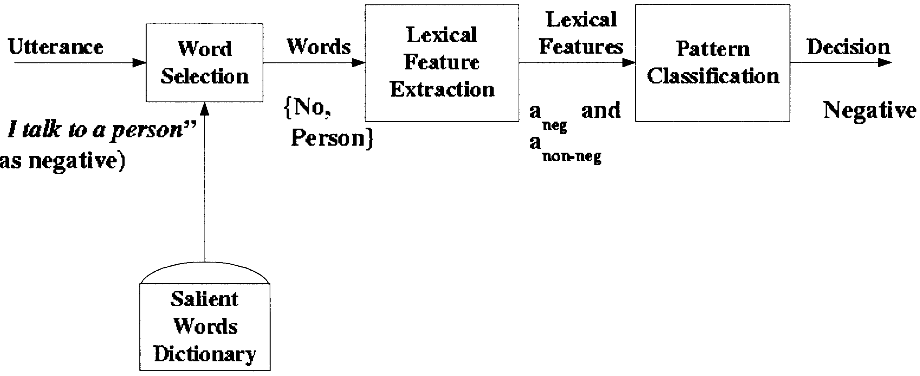


Fig. 1. Block diagram for the classification using lexical information. An example of classification using an utterance from the data corpus is also shown.

TABLE II
PARTIAL LIST OF SALIENT (ISOLATED) WORDS IN THE DATA."EMOTION" REPRESENTS MAXIMALLY CORRELATED EMOTION CLASS GIVEN WORDS, i.e., THE EMOTION CLASS THAT MAXIMIZES THE POSTERIOR PROBABILITY OF EMOTION GIVEN A WORD

Word	Salience	Emotion
Wrong	0.72	negative
Computer	0.72	negative
Damn	0.72	negative
No	0.45	negative
Arrival	0.33	non-negative
Phoenix	0.33	non-negative
Delayed	0.21	non-negative
Baggage	0.20	non-negative

then $P(e_k \,|\, v_n) > P(e_k)$, and $i(v_n, e_k)$ is positive. Whereas, if the word v_n makes a class e_k less likely, $i(v_n, e_k)$ is negative. If there is no effect by the word, $i(v_n, e_k)$ will be zero because $P(e_k \,|\, v_n) = P(e_k)$. The emotional salience of a word for an emotion category is defined as mutual information between a specific word and emotion class

$$\mathrm{sal}(v_n) = I(E; W = v_n) = \sum_{j=1}^{k} P(e_k \,|\, v_n) i(v_n, e_k). \quad (4)$$

In summary, emotional salience is a measure of the amount of information that a specific word contains about a given emotion category. Illustrative examples of salient words in the data corpus are given in Table II. Emotion here represents the one maximally associated with the given word. After identifying the salient words, we retained the words that had salience values greater than a prechosen threshold.

In order to estimate the salience of words in our data corpus, unlike in the acoustic information case, we combined both male and female speech utterances because language information was deemed to be gender-independent.

After computing the emotionally salient words in the (training) data corpus, we calculated the language-related emotion recognition decision at the utterance level. Let E_1 and E_2 represent negative and non-negative emotion classes, respectively, and v_n denote the emotionally salient words obtained from the training data corpus [18], [23]. Each word in an utterance is also assumed to be independent of each other; the effect of automatic speech recognition errors is also ignored for simplicity. An input utterance, $W = \langle v_{n1}, v_{n2}, \ldots, v_{nL} \rangle$, where L is the length of the utterance, is mapped to an emotion category, E_1 or E_2. If the words in an utterance match the emotionally salient words, we can output 1 from those words; otherwise 0, i.e., it is binary output either 0 or 1. And then these binary outputs from an utterance are combined with the emotion class nodes to produce *activations*, a_k, in which k is either 0 (non-negative) or 1(negative). The formula for a_k is as follows:

$$a_k = \sum_{m=1}^{n} I_m w_{mk} + w_k \quad (5)$$

where I_m denotes indicator, which has either 0 or 1 representing either a word matched to a salient word or not, w_{mk} denotes connection weight, and w_k is bias. We can define the connection weights w_{mk}, and bias w_k as follows:

$$\begin{aligned} w_{mk} &= i(v_n, e_k) \\ &= \log \frac{P(e_k \,|\, v_n)}{P(e_k)} \quad (6) \\ w_k &= \log P(E_k) \quad (7) \end{aligned}$$

Finally, the feature related to the language information we chose is the difference, i.e., $a_0 - a_1$ where 0 represents non-negative and 1 represents negative, in activations since we are interested in making a decision between two emotion categories. Fig. 1 shows the block diagram for classification using lexical information. We also show an example in the figure, which comes from the data corpus.

An important and interesting point of our activation is that under the independent assumption of each word in an utterance, the activation provides a maximum a posteriori (MAP) decision [23] since

$$\begin{aligned} P(E_k \,|\, W) &= \frac{P(W \,|\, E_k) P(E_k)}{P(W)} \quad (8) \\ &= \frac{[\prod_{i=1}^{L} P(v_{ni} \,|\, E_k)] P(E_k)}{P(W)} \quad (9) \\ &= \prod_{i=1}^{L} \left[\frac{P(E_k \,|\, v_{ni}) P(v_{ni})}{P(E_k)} \right] \frac{P(E_k)}{P(W)} \quad (10) \\ &= \prod_{i=1}^{L} \left[\frac{P(E_k \,|\, v_{ni})}{P(E_k)} \right] P(E_k) \frac{\prod_{i=1}^{L} P(v_{ni})}{P(W)}. \quad (11) \end{aligned}$$

TABLE III
NUMBER OF EACH DISCOURSE LABEL FROM THE DATA CORPUS

Tag	Male		Female		Total	
	Negative	Non-negative	Negative	Non-negative	Negative	Non-negative
rejection	37	7	72	10	109	17
repeat	4	35	23	38	27	73
rephrase	15	34	10	39	25	73
ask-startover	29	33	33	44	62	77
non	57	350	71	448	128	798
total	142	454	209	579	351	1038

The last term in (11) is equal to (1) because the words in an utterance are independent of each other. Taking log on both sides

$$\log P(E_k \mid W) = \sum_{i=1}^{L} i(v_n, e_k) + \log P(E_k) \qquad (12)$$

which is the same as the activation a_k.

The aforementioned development focuses on emotional salience based on isolated words. Salience of a word can, however, be extended to include word sequences. For example, the word "damn" would be followed by "it" rather than "damn" itself, and thus, we may build salient word pairs. While such an extension appears to be promising, the sparsity of longer word sequences in our data precluded this from the present work. Similarly, we assumed perfect speech recognition (i.e., operated from true transcriptions); one reason was that the underlying automatic speech recognition (ASR) accuracy was relatively high for this constrained grammar task and the effect of ASR errors was not predominant. A possible extension in this direction would be to use word level confidence scores to weight the lexical-based emotion decision. Such extensions will be explored in future work.

B. Discourse Information

Discourse information in human-computer interaction has been suggested as being potentially useful for emotion recognition [15]–[17] and has been combined with acoustic information to improve the performance of emotion recognizers [15], [17]. In our study, the discourse labels are based on the categorization of users' responses. We defined five speech-act like labels for categorizing the user's response to the system: *rejection, repeat, rephrase, ask-start over*, and *none of the above*. This labeling was performed by one person. Since we did not have the system prompts, the labels were based just on the utterance transcriptions. Rejection occurred when a user rejected the system's response. Repeat was defined as the repetition of the contents that a user said. Rephrase had the same definition as repeat, but with the difference that an utterance was repeated in a different way. The utterances in which a user asks for help or tries to go back to the beginning are defined as ask-startover. Note that the labeling of discourse information was done separately from labeling of emotions. Utterances were randomly mingled to avoid contextual influence on the labelers.

The number of utterances corresponding to each discourse label is shown in Table III. Most utterances were labeled as *none of the above* and corresponded to user responses to specific information requests. As expected, a large portion of utterances in the negative emotion category were labeled as rejection (26% for male data and 34% for female utterances). In contrast, only about 2% of the non-negative utterances were labeled as *rejection*. As features to classifiers, we combined *repeat* and *rephrase* labels into a single category because they were user responses under similar situations and this helped reduce the dimension of the feature set.

V. COMBINING INFORMATION SOURCES

This section focuses on combining the three sources of information—acoustic, lexical, and discourse information—for classifying emotions from spoken language. One possible strategy is to combine these three pieces of information at the feature level by constructing a large feature vector [17]. The problem with feature level fusion/combination is the potential of having to face the curse of dimensionality due to the increase in the input feature dimension [21]. Another scheme for combining several pieces of information is decision level fusion. We can generate multiple classifiers to manipulate the set of multiple information available to the learning algorithm. In our case, we have three independent classifiers, one each for acoustic, language, and discourse information streams, and a final decision is made by combining the output results from these classifiers. In this work, we used a simple average of the outputs from each source of information (see Fig. 2) because it achieves good performance in spite of its simplicity and less support by probabilistic interpretation [24], [25].

Let us consider the problem of combining acoustic, lexical, and discourse information at the decision level [26] (see Fig. 2.) Let x_n and y_n denote input and output for a classifier that processes different source of information, where $n = 1, \ldots, N$, total number of classifiers. Probabilistically, y_n is the estimated posterior probability of class label given data, $\hat{P}(E_k \mid x_n)$. And denote y as the final output averaged over y_n and $\mathbf{x}$ as feature vector representing the whole information, i.e., $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$ where T denotes transpose. Then the final output of the overall classifiers is given by

$$y(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} y_n(x_n) \qquad (13)$$

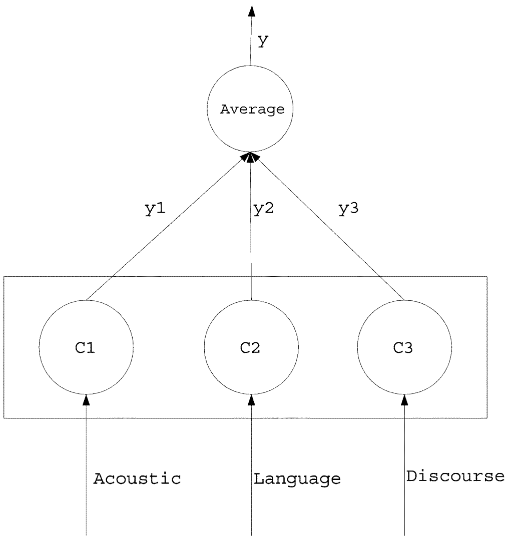


Fig. 2. Combination of acoustic, language, and discourse information by averaging the outputs from each information stream.

where N is the number of classifiers to be combined. The output from each classifier is the confidence level of each information stream. At the confidence level, a classifier outputs a numerical value between [0,1] for each class indicating the probability that the given input belongs to that class. Therefore, in probabilistic notation, it can be written as

$$P(E_k \,|\, \mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} P(E_k \,|\, x_n). \tag{14}$$

Since the estimation of posterior probability, $P(E_k \,|\, \mathbf{x})$, requires large number of training data converging to Bayes error rate, the estimated posterior probability would have significant error. Averaging the outputs from each classifier can provide less error-sensitive estimation of the posterior probability. The final decision is made by

$$P(E_j \,|\, \mathbf{x}) = \arg\max_k P(E_k \,|\, \mathbf{x}). \tag{15}$$

VI. Experimental Results

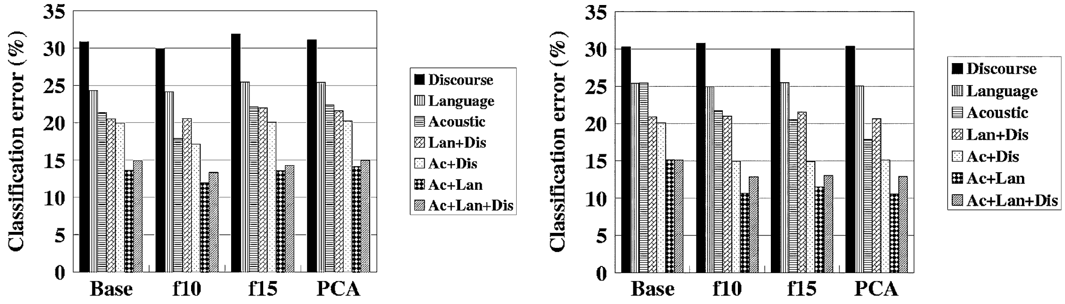
Two different classification methods were investigated using acoustic information: linear discriminant classifiers (LDC) which assume each class has Gaussian probability density with a common covariance matrix, and also k-nearest neighborhood classifiers (k-NN). LDC's were used for classification with both language and discourse information. Since F0-based measures were an integral part of the acoustic feature set, male and female data were considered separately. This allowed us to investigate gender-dependencies in classification using acoustic measures. In real applications, the differentiation of gender can be performed, for example, by building models for both male and female or setting a threshold for mean pitch, which can be obtained from the first 5 or 10 frames of speech data. The training data set was selected 10 times in a random manner from the whole data set for each gender with the same number of data for each class (200 for male data and 240 for female). In the following experiments, all the classification errors were calculated by a tenfold cross-validation. The training data was divided into ten disjoint sets of equal size, and classifiers were trained ten times, each time with a different set held out as a validation set [21]. The estimated classification error is the mean of these ten errors for the validation data. The final errors were calculated averaging over ten training data set randomly chosen from the data pool.

TABLE IV
Classification Error Using Acoustic Correlates For Different Feature Set Choices: Base—All the Features of 21 Dimensions, f10–10 Best Feature Set, f15–15 Best Feature Set, PCA—Feature Set by Principal Component Analysis (for Male, $k = 8$ in k-NN, and $k = 4$ for Female Data)

	LDC		**k-NN**	
Feature set	Male(%)	Female(%)	Male(%)	Female(%)
Base	25.45 ± 6.26	21.47 ± 3.92	21.35 ± 2.40	24.25 ± 2.11
f10	21.70 ± 4.87	**12.04** ± 3.99	**17.85** ± 2.48	**20.87** ± 1.81
f15	20.55 ± 6.90	19.21 ± 6.34	22.10 ± 2.18	21.37 ± 2.13
PCA	**17.85** ± 3.03	19.29 ± 2.38	22.35 ± 2.02	24.79 ± 2.25

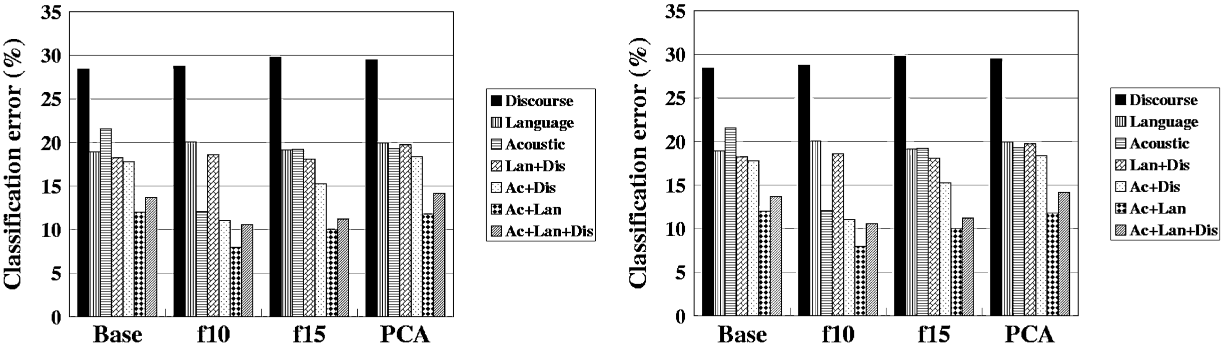
Table IV shows the performance comparison between different feature sets, and also compares the performance of LDC and k-NN classification methods in the context of using acoustic correlates only. The numbers of neighborhood in k-NN classifiers were set to eight for male, and four for female data. Those numbers were estimated by leave-one-out cross-validation on the whole database for each gender. For PCA, we took the dimension explaining 90% of the sum of eigenvalues. In all the cases except the male data with LDC case, the f10 feature set (ten best features) had the best performance in terms of classification error. The performance of a PCA feature set did not show improvements compared with the base feature set, but we should note that dimensionality has been reduced by PCA with comparable performance to both the base feature set and the f15 feature set. From the graph, we can see that the PCA feature set has less variation compared with other feature sets, and thus we can say that the PCA feature set is more reliable.

Next, we investigated the combination of the other sources of information with acoustic information in order to improve the performance with respect to overall classification error. The results are shown in Figs. 3 and 4, and Tables V and VI. For comparison, a baseline classification error results were included in Tables V, and VI. The baseline errors were the case when all the data were classified as non-negative, which is the most frequent class in our database. Overall, combining other sources of information with acoustic information leads to performance improvements in emotion recognition. Note that the case of the combination of acoustic and language information showed the best performance in almost all the settings. The inclusion of discourse information for this data does not seem to provide any significant improvements when used in conjunction with acoustic and lexical information. This may be due to the fact that lexical information is highly correlated with the measured discourse information. It is difficult to investigate the dependency or correlation between information in the feature level but we can explore the similarity between classifiers using the classification results that stem from them. Since each classifier has been trained using different information, if each information is



(a) (b)

Fig. 3. Comparison of combination of input features. Ac = acoustic correlates, Lan = language information, Dis = discourse information. Four sets of features are explored. (a) Male data using LDC, and (b) Male data using k-NN with k = 8 in the classification of acoustic correlates; LDCs were used for classification with lexical and discourse information.



(a) (b)

Fig. 4. Comparison of combination of input features. Ac = acoustic correlates, Lan = language information, Dis = discourse information. Four sets of features are explored. (a) Female data using LDC, and (b) Female data using k-NN with k = in the classification of acoustic correlates, LDCs were used for other information.

TABLE V
CLASSIFICATION ERROR AND ITS STANDARD DEVIATION FOR MALE DATA IN PERCENTILE WITH DIFFERENT FEATURE SETS (BASE = 21 dim, f10 = 10 BEST FEATURE, f15 = 15 BEST FEATURE, PCA = FEATURE SET BY PCA.) Ac = ACOUSTIC CORRELATES, LAN = LANGUAGE INFORMATION, DIS = DISCOURSE INFORMATION. THE BASELINE WAS THE CASE WHEN ALL THE DATA HAS BEEN CLASSIFIED AS NON-NEGATIVE. THE BEST PERFORMING RESULT WITH RESPECT TO CLASSIFICATION ERROR IS BOLD-FACED IN EACH CASE

	Information	**Base**	**f10**	**f15**	**PCA**
LDC	Acoustic only	25.45 ± 6.26	21.70 ± 4.87	20.55 ± 6.90	17.85 ± 3.03
	Language only	25.40 ± 1.85	24.95 ± 1.93	25.50 ± 1.88	25.05 ± 2.68
	Discourse only	30.25 ± 2.11	30.75 ± 1.39	30.00 ± 2.12	30.35 ± 1.76
	Ac+Lan	15.15 ± 2.21	**10.65** ± 3.53	**11.50** ± 4.45	**10.55** ± 2.12
	Ac+Dis	20.10 ± 2.51	14.95 ± 3.86	14.90 ± 5.22	15.10 ± 2.06
	Lan+Dis	20.90 ± 1.55	21.00 ± 1.77	21.55 ± 1.16	20.65 ± 2.21
	Ac+Lan+Dis	**15.10** ± 2.44	12.85 ± 2.85	13.05 ± 2.74	12.90 ± 1.52
k-NN (k = 8)	Acoustic only	21.35 ± 2.40	17.85 ± 2.48	22.10 ± 2.18	22.35 ± 2.02
	Language only	24.30 ± 2.05	24.15 ± 2.49	25.45 ± 1.67	25.40 ± 1.96
	Discourse only	30.80 ± 2.84	29.90 ± 1.71	31.85 ± 2.18	31.10 ± 1.85
	Ac+Lan	**13.60** ± 1.82	**11.95** ± 2.48	**13.55** ± 1.27	**14.10** ± 1.77
	Ac+Dis	19.95 ± 1.89	17.15 ± 2.10	20.05 ± 1.46	20.25 ± 1.25
	Lan+Dis	20.50 ± 2.14	20.55 ± 1.51	21.95 ± 1.99	21.60 ± 1.96
	Ac+Lan+Dis	14.90 ± 1.86	13.30 ± 1.45	14.25 ± 2.04	14.95 ± 1.92
Baseline	23.5				

related to each other, it is highly likely to give a similar decision on a given input data. To see the dependency between the classifiers of lexical and discourse information, we calculated the *Q-statistic*, which measures the similarity between classifiers [27]. *Q-statistic* provides a pairwise symmetrical measure of similarity. For two classifiers y_i and y_j, the *Q-statistic* is defined as

$$Q_{ij} = \frac{N_{11}N_{00} - N_{01}N_{10}}{N_{11}N_{00} + N_{01}N_{10}} \tag{16}$$

TABLE VI
CLASSIFICATION ERROR AND ITS STANDARD DEVIATION FOR FEMALE DATA IN PERCENTILE WITH DIFFERENT FEATURE SETS (BASE = 21 dim, f10 = 10 BEST FEATURE, f15 = 15 BEST FEATURE, PCA = FEATURE SET BY PCA.) AC = ACOUSTIC CORRELATES, LAN = LANGUAGE INFORMATION, DIS = DISCOURSE INFORMATION. THE BASELINE WAS THE CASE WHEN ALL THE DATA HAS BEEN CLASSIFIED AS NON-NEGATIVE. THE BEST PERFORMING RESULT WITH RESPECT TO CLASSIFICATION ERROR IS BOLD-FACED IN EACH CASE

	Information	Base	f10	f15	PCA
LDC	Acoustic only	21.47 ± 3.92	12.04 ± 3.99	19.21 ± 6.34	19.29 ± 2.38
	Language only	18.91 ± 1.59	20.04 ± 1.89	19.12 ± 1.21	19.95 ± 1.43
	Discourse only	28.41 ± 1.77	28.75 ± 1.59	29.75 ± 2.74	29.45 ± 1.34
	Ac+Lan	**12.00** ± 2.76	**7.95** ± 2.51	**10.04** ± 1.83	**11.79** ± 1.64
	Ac+Dis	17.79 ± 3.13	11.04 ± 3.16	15.25 ± 3.73	18.37 ± 2.22
	Lan+Dis	18.25 ± 1.61	18.58 ± 1.40	18.08 ± 2.08	19.75 ± 1.04
	Ac+Lan+Dis	13.66 ± 1.95	10.58 ± 2.47	11.20 ± 1.01	14.16 ± 2.03
k-NN (k = 4)	Acoustic only	24.25 ± 2.11	20.87 ± 1.81	21.37 ± 2.13	24.79 ± 2.25
	Language only	18.91 ± 1.33	18.66 ± 2.15	20.29 ± 1.70	19.29 ± 1.37
	Discourse only	27.50 ± 2.22	28.12 ± 4.16	29.45 ± 2.15	27.75 ± 2.87
	Ac+Lan	**13.04** ± 1.64	**13.12** ± 2.22	**12.37** ± 1.72	**13.75** ± 1.76
	Ac+Dis	22.62 ± 2.20	19.91 ± 3.51	20.62 ± 2.57	22.70 ± 2.82
	Lan+Dis	17.62 ± 1.38	17.75 ± 2.22	19.62 ± 2.44	18.00 ± 2.09
	Ac+Lan+Dis	16.50 ± 1.27	14.83 ± 2.16	16.20 ± 1.71	15.91 ± 0.99
Baseline	25.9				

TABLE VII
Q STATISTICS FOR PAIRWISE INFORMATION IN BOTH MALE AND FEMALE. ACOUS = ACOUSTIC CORRELATES, LANG = LEXICAL INFORMATION, DISC = DISCOURSE INFORMATION

	Male	Female
Q(acous, lang)	0.4403	0.0258
Q(acous, disc)	0.2761	0.1755
Q(lang, disc)	0.9276	0.9154

where

- N_{11} is the number of both classifiers making the correct classification;
- N_{10} is the number of y_i being correct and y_j being incorrect;
- N_{01} is the number of y_i being incorrect and y_j being correct;
- N_{00} is the number of both classifiers making incorrect decision.

$N = N_{11} + N_{01} + N_{10} + N_{00}$ is the total number of data. Q_{ij} has a value between -1 and 1. For a statistically independent classifier, Q_{ij} is 0, and the higher the absolute value of Q_{ij}, the more dependent the classifiers are.

The pairwise *Q-statistic* is shown in Table VII for classifiers with acoustic, lexical, and discourse information sources. For the results, we divided the data into training and testing, and then Q_{ij}'s were obtained only from the testing data. As expected, *Q-statistic* between the classifiers from language and discourse information is almost 1. This is the reason why the addition of discourse information to acoustic and lexical information was not be an improvement. It has been known that the use of classifiers using the information sources dependent to each other cannot improve the performance; it can even worsen performance [27].

VII. DISCUSSION

Automatic recognition of emotions from human speech by machines is gaining increasing attention from the engineering community. The performance by a computer, and the emotional categories it can cover, are far limited compared with those capable by humans. One main difficulty comes from the fact that there is a lack of complete understanding of emotions in human minds, including a lack of agreement among psychological researchers. Agreement among researchers is a prerequisite to satisfaction in attempting to build an effective machine for the task of automatic emotion recognition. Even human beings have difficulty categorizing emotions, as evident in the low kappa statistic values in our corpus. However, we believe that we can design algorithms that will perform reasonably well in constrained domain-specific applications, such as automated call center applications that we focused on in this paper. The knowledge gained from these efforts can help us understand deeper issues and potentially extend to more general applications.

In this paper, we explored domain specific emotion recognition from speech signals using data obtained from a real-world call center dialog application. Language and discourse information, as well as acoustic features that most studies have focused on, were explored to improve the performance of an emotion recognizer. The results show that significant improvements can be made by combining these information sources in the same framework. An information-theoretic concept of *emotional salience* was used to obtain language-related features, which were termed as *emotionally salient words* in an utterance. For discourse information, we considered a simple set of five speech act categories related to the users' response to the automated system. There are several open issues that need to be further explored in the future.

First of all, data sparsity is a significant problem in the modeling of lexical information. In the test phase of classification using lexical information, many utterances were left undecided due to the fact that the words in certain utterances were not in the list of salient words seen in the training data, even if one or more words were apparently related to emotion classes. To explore this problem, we need to experiment with the dependence of language information on the number of salient words. We also need to increase the amount of data available for estimating the salient words. Finally, we need to study effective smoothing techniques for dealing with data sparsity.

Secondly, in this paper, we estimated the emotional salience calculation at a single word level; however, emotional salience should also be extended to word sequences. This may lead to a more reasonable estimation of emotional salience in the sense that human beings often incorporate word sequences to judge emotion states. This should be possible, again, with a larger data corpus. The role of speech recognition errors could be incorporated as well, say through the use of confidence scores as weights in the decision function.

The third issue is how to best combine several kinds of information. In this paper, we formulated this as a data fusion problem and combined information at the decision level. But there are other possible strategies for combining several sources of information; feature level fusion, or using *winner-take-all* method in making a decision. Also, each classifier for each type of information in this work was trained and locally optimized before the combination, and thus the results are suboptimal. Further improvements may be made if we globally optimize the parameters of the combination of classifiers by jointly training the whole system. This combination/fusion problem involving different information sources is still largely an open question that continues to be tackled by the data fusion, signal processing, and machine learning communities.

Since emotional states do not have clear-cut boundaries (even people are usually confused at figuring out the emotional states of other people), we need to explore and develop classification methods that can deal with this vague boundary problem. Preliminary results using fuzzy inference methods seem to hold promise [28]. Furthermore, we need to investigate an expanded set of emotion categories that may be useful in other applications and contexts.

Acknowledgment

The authors wish to thank R. Pieraccini and S. Lee for their valuable help. Any opinions, findings and conclusionsor recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[1] A. Damasio, *Descartes' Error: Emotion, Reason, and the Hitman Brain*. London, U.K.: Putman, 1994.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[3] A. Ortony, G. Clore, and A. Collins, *The Cognitive Structure of Emotions*. Cambridge, U.K.: Cambridge Univ. Press, 1988.

[4] J. Averill, "In the eyes of the beholder," *The Nature of Emotion: Fundamental Questions*, pp. 7–14, 1994.

[5] K. Scherer, "Toward a concept of 'modal emotions'," *The Nature of Emotion: Fundamental Questions*, pp. 25–31, 1994.

[6] R. Plutchik, *The Psychology and Biology of Emotion*. New York: HarperCollins College, 1994.

[7] K. Scherer, "A cross-cultural inverstigation of emotion inferences from voice and speech: Implications for speech technology," in *Proc. ICSLP*, Beijing, China, Oct. 2000, pp. 379–382.

[8] C. Lee, S. Narayanan, and R. Pieraccini, "Recognition of negative emotions from the speech signal," in *Proc. Automatic Speech Recognition Understanding*, Dec. 2001.

[9] V. Petrushin, "Emotion in speech: Recognition and application to call centers," *Artif. Neu. Net. Engr. (ANNIE)*, 1999.

[10] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proc. ICSLP*, Philadelphia, PA, 1996, pp. 1970–1973.

[11] D. Roy and A. Pentland, "Automatic spoken affect analysis and classification," in *Proc. Int. Conf. Automatic Face Gesture Recognition*, Killington, VT, 1996, pp. 363–367.

[12] R. Banse and K. Scherer, "Acoustic profiles in vocal emotion expression," *J. Personality Social Psych.*, vol. 70, no. 3, pp. 614–636, 1996.

[13] I. Murray and J. Arnott, "Toward a simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *Acoust. Soc. Ante.*, vol. 93, no. 2, pp. 1097–1108, 1993.

[14] S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk, and S. Stroeve, "Approaching automatic recognition of emotion from voice: A rough benchmark," in *Proc. ISCA Workshop Speech Emotion*, 2000, pp. 207–212.

[15] A. Batliner, K. Fischer, R. Huber, J. Spiker, and E. Noth, "Desperately seeking emotions: Actors, wizards, and human beings," in *Proc. ISCA Workshop Speech Emotion*, 2000, pp. 195–200.

[16] S. Arunachalam, D. Gould, E. Anderson, D. Byrd, and S. Narayanan, "Politeness and frustration language in child-machine interactions," in *Proc. Eurospeech*, Aalborg, Denmark, 2001, pp. 2675–2678.

[17] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Proc, ICSLP*, Denver, CO, Sep. 2002, pp. 2037–2040.

[18] A. Gorin, "On automated language acquisition," *J. Acoust. Soc., Amer.*, vol. 97, no. 6, pp. 3441–3461, 1995.

[19] S. Siegel and J. N. J. Castellan, *Nonparametric Statistics for The Behavioral Sciences*. New York: McGraw-Hill, 1988.

[20] J. Carletta, "Assessing agreement on classification tasks: The kappa statistic," *Comput. Linguistics*, vol. 22, no. 2, pp. 249–254, 1996.

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

[22] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[23] A. Gorin, A. Gertner, and E. Goldman, "Adaptive acquisition of language," in *Neural Networks, Theory, and Applications*, R. Mammone and Y. Zeevi, Eds. San Diego, CA: Academic, 1991, pp. 125–167.

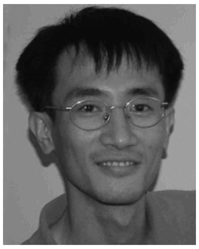
[24] D. Tax, M. van Breukelen, R. Duin, and J. Kittler, "Combining multiple classifiers by averaging or by multiplying," *Pattern Recognit.*, vol. 33, pp. 1475–1485, 2001.

[25] L. Kuncheva, J. Bezdek, and R. Duin, "Decision templates for multiple classifier fusion: An experimental comparison," *Pattern Recognit.*, vol. 34, no. 2, pp. 299–314, 2001.

[26] Z. Chair and P. K. Varshney, "Optimal data fusion in multiple sensor detection systems," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-22, pp. 98–101, 1986.

[27] L. Kuncheva and C. Whitaker, "Measure of diversity in classifier ensembles," *Mach. Learn.*, vol. 51, pp. 181–207, 2003.

[28] C. Lee and S. Narayanan, "Emotion recognition using a data-driven fuzzy inference system," in *Proc. Eurospeech*, 2003, pp. 157–160.



Chul Min Lee (S'01) received the B.S. degree in physics from Yonsei University, Seoul, Korea, the B.S. degree with honors in electrical engineering from University of Wisconsin, Milwaukee, in 1994 and 1998, respectively, the M.S. degree in electrical engineering from University of Southern California (USC), Los Angeles, in 1999, and he is currently pursuing the Ph.D. degree in electrical engineering from USC.

His research interests are in emotion recognition in speech, speech recognition, pattern recognition, and digital signal processing.

Shrikanth S. Narayanan (M'87–SM'02) received the M.S. degree in 1990, the Engineer degree in 1992, and the Ph.D degree in 1995, all from the University of California, Los Angeles (UCLA).

He was with AT&T Research (originally AT&T Bell Labs), first as a Senior Member and later as a Principal Member of its Technical Staff from 1995 to 2000. Currently, he is an Associate Professor of Electrical Engineering, Linguistics, and Computer Science at the University of Southern California (USC), Los Angeles. He is a Research Area Director of the Integrated Media Systems Center and a National Science Foundation (NSF) Engineering Research Center at USC. His research interests include signals and systems with applications to speech, language, multimodal, and biomedical problems. He has published over 125 papers and holds three U.S. patents.

Dr. Narayanan was an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (2000–2003) and serves on the Speech Communication Technical Committee of the Acoustical Society of America, the speech processing and multimedia signal processing technical committees of the IEEE Signal Processing society. He is a member of Tau-Beta-Pi, Phi Kappa Phi, and Eta-Kappa-Nu. He received the NSF CAREER Award, the USC Engineering Junior Research Award, and a Center for Interdisciplinary Research Faculty Fellowship.