# VOICE SIGNATURES

*Izhak Shafran, Michael Riley, and Mehryar Mohri*

AT&T Labs – Research
Florham Park, USA 07932
{zak,riley,mohri}@research.att.com

## ABSTRACT

Most current spoken-dialog systems only extract sequences of words from a speaker's voice. This largely ignores other useful information that can be inferred from speech such as gender, age, dialect, or emotion. These characteristics of a speaker's voice, *voice signatures*, whether static or dynamic, can be useful for speech mining applications or for the design of a natural spoken-dialog system. This paper explores the problem of extracting automatically and accurately voice signatures from a speaker's voice.

We investigate two approaches for extracting speaker traits: the first focuses on general acoustic and prosodic features, the second on the choice of words used by the speaker. In the first approach, we show that standard speech/nonspeech HMMs, conditioned on speaker traits and evaluated on cepstral and pitch features, achieve accuracies well above chance for all examined traits. The second approach, using support vector machines with rational kernels applied to speech recognition lattices, attains an accuracy of about 81% in the task of binary classification of emotion. Our results are based on a corpus of speech data collected from a deployed customer-care application (HMIHY 0300). While still preliminary, our results are significant and show that voice signatures are of practical interest in real-world applications.

## 1. INTRODUCTION

Most current spoken-dialog systems only extract sequences of words from a speaker's voice. This largely ignores other useful information that can be inferred from speech. As humans, we infer a number of important traits about a speaker that may include gender, age, dialect, ethnicity, emotion, level of education, state of intoxication, and even height or weight.

These characteristics of a speaker's voice, *voice signatures*, whether static or dynamic, can be useful for speech mining applications or for the design of a natural spoken-dialog system. When combined, and in some cases accumulated over time, voice signatures can lead to a characterization of a speaker. In speech mining, such traits can be used to rapidly identify relevant speech or multimedia documents out of a large data set. Voice signatures can also be utilized to create more natural human-machine speech interfaces than existing ones. For example, knowing that an elderly person from a Southern state is upset, an automatic customer-care system could comfort the user with an appropriate dialog sequence in a synthesized Southern accent. Similarly, when a young New Yorker enquires, the system could deliver the information at a suitable pace with an accent more familiar to him. Similarly in a marketing dialog, information about age or gender can be used to guide the offers and questions posed.

This paper explores the problem of extracting automatically and accurately voice signatures from a speaker's voice. Our study focuses on the following four signatures: gender, age, dialect, and emotion. Gender is detected in many automatic speech recognition (ASR) systems for the purpose of choosing gender-specific acoustic models. Automatic detection of age and dialect have been studied to a lesser extent.

In a series of experiments, several authors have studied subjective estimation of age on small corpora [4, 3]. Their work shows that humans can estimate age groups reliably even across certain languages. The only study on automatic age detection that we are aware of is that of [17] which shows that, on a small Japanese corpus, a binary classification of elderly versus others can be performed with an accuracy of 95%. The authors found standard Mel-warped cepstral features, speaking rate, and local variation of power (shimmer) to be useful, but not pitch.

While a few researchers have studied the automatic classification of foreign accents in English, studies on dialect are even fewer. In [22, 1], mono-phones conditioned on foreign accents were used to classify 4-6 classes of speakers to obtain an accuracy of about 61-65% on isolated words. Miller and Trischitta from BBN, applied Fisher linear discriminant to classify two distinct dialect groups – Northern (Boston and Midwest) and Southern (Atlanta and Dallas), and obtained a binary classification accuracy of about 92% on digit strings and short phrases [16].

The detection of emotions has been the subject of several recent studies. In most of the work reported, various general classes of emotion are defined and correlated with measurable attributes of speech [8, 20]. Some recent studies are based on artificial speech data: speech collected from actors asked to articulate distinct types of emotions, or data from an artificial laboratory setup [19, 2, 20]. Two recent sets of studies are based on data collected from real applications. Devillers *et al.* [9] investigated automatic classification of utterances into five categories, namely *neutral*, *anger*, *fear*, *satisfaction* and *excuse* using interpolated unigrams conditioned on each class of emotion. Using the highest posterior on the reference transcripts, these authors obtained an accuracy of 68%, when averaged over two balanced test sets of 125 utterances each. Lee *et al.* [11, 12, 13, 14] extracted a set of utterance-level statistics on acoustics such as pitch and energy. In addition, they computed a measure of *salience* over words that occurred in the utterance. It is unclear whether the words were from the reference or from the ASR output. The combination of the two scores, under the assumption of statistical independence, gave a binary classification accuracy of 80%.

Our experiments are based on a speech corpus collected from a deployed customer-care application (HMIHY 0300). In Section 2, we briefly describe the annotation of that corpus for the purpose of detection of voice signatures, introduce the definition of the classes or categories used for each speaker trait, and discuss our test of the consistency of the labeling. Both acoustic (and prosodic) features and word sequences may be used as a clue to determine such speaker traits. Section 3 describes our use of cepstra augmented with pitch and HMM-based classifiers to identify the four speaker traits already mentioned and presents the results of our experiments. Section 4 describes how word lattices output by a speech recognizer can be used to derive these speaker traits. It briefly describes a general discriminant classification technique, support vector machines (SVMs) with rational kernels that are applied to lattices, and presents the results of our experiments, in particular to the binary classification of emotions into two categories.

## 2. CORPUS ANNOTATION AND CATEGORY DEFINITION

To understand how well speaker traits can be extracted in a real-world application, we chose to work with data collected from a deployed customer-care system, AT&T's "How May I Help You" system (HMIHY) [10]. The following sections describe the guidelines used for the annotation of the corpus for our purposes here and discuss how labeling consistency was measured on the corpus.

### 2.1. Annotation

The corpus annotation presented several issues, in particular because the data was from a real application (as opposed to "acted" speech) and because there are no widely accepted definitions for the class values associated with emotion and dialect. We took a pragmatic approach guided mostly by the following three goals: ease of labeling by native speakers, high inter-labeler agreement, and the ability to identify ambiguous labels.

Initially, we adopted the age and dialect labels used in the DARPA Switchboard corpus. However, based on some informal perceptual experiments, we decided to collapse perceptually confusable categories. For example, Mid-West and Western dialects were merged into one category. Similarly, age was grouped into broad categories. Further, to reduce the variability within classes, intermediate groups were defined to isolate speakers who were hard to classify into one category alone. The details of the annotations for each trait are given below.

**Age**. While gender, dialect and emotion are categorical, age is continuous and its estimation could be treated as a regression problem. However, it proved convenient and less confusable for us to group age into discrete bins. Perceived age was grouped into four main categories: *Youth* (below 25 years), *Adult* (between 25 and 50 years), and *Senior* (above 50 years). When the labelers were unsure, they were allowed to assign speakers to intermediate categories: *Youth/Adult*, and *Adult/Senior*, with the motivation that the uncertain assignments can be refined using inter-labeler consistency tests or automatic techniques.

**Dialect**. Based on our perceptual studies and feedback from labelers, dialect was annotated with the following seven categories: *NE* (Boston and similar), *NYC* (Brooklyn and similar), *MW/W* (Mid-West, West and "standard" American dialect), *So* (Southern), *AfAm* (African American), *Hisp/Lat* (Hispanic or natives of Latin America), and *NonNat* (non-native not belonging to above categories). In order to capture the uncertainty in the assignment of categories, the labelers were also requested to qualify their classification with *Strong/Sure* or *Weak/Unsure* labels.

**Emotion**. Emotion was also grouped into seven categories: *PosNeu* (positive, neutral, cheerful, amused or serious), *SWAngry* (somewhat angry), *VAngry* (very angry, as with raised voices), *SWFrstd* (somewhat frustrated), *VFrstd* (very frustrated), *OtherSWNeg* (otherwise category of lower intensity) and *OtherVNeg* (otherwise category of higher intensity). When several negative emotions apply, labelers were requested to choose the dominant one. In case of ambiguity, the category with the earlier rank in the above mentioned list was selected.

The utterances from a customer were presented to the labelers in the order of occurrence, thus they had the advantage of knowing the context beyond the utterance being labeled. Note, this is unlike some of the previous studies [12, 9] where the utterances were presented in a random order. On the average, the utterances were about 15 words long. The total number of utterances in the corpus is 5147 (from 1854 calls) and the observed class distributions are shown in Table 1.

| Trait | Classes (%) |
|---|---|
| Gender | Female (65); Male (35). |
| Age | Youth (2.3); Adult (57.9); Senior (24.0); Youth/Adult (3.0); Adult/Senior (12.9). |
| Dialect | MW/W (51.2); So (17.1); NonNat (5.9); AfAm (4.8); HispLat (4.5); NE (1.4); NYC (15.1). |
| Emotion | PosNeu (70.9); SWFrstd (16.3); SWAngry (7.8); VFrstd (2.7); VAngry (1.7); OtherSWNeg (0.4). |

**Table 1**. Observed distribution of classes in the corpus. Classes with less than 0.1% occurrences are not shown.

| | |
|---|---|
| **Gender** | |
| Male, Female | 1.0 |
| **Age** | |
| Youth, Youth/Adult, Adult, Adult/Senior, Senior | 0.34 |
| {Youth, Youth/Adult}, {Adult, Adult/Senior}, Senior | 0.37 |
| {Youth, Youth/Adult}, Adult, {Adult/Senior, Senior} | 0.50 |
| **Dialect** | |
| AfAm, Hisp/Lat, MW/W, NYC, NE, NonNat, So | 0.58 |
| AfAm, Hisp/Lat, {MW/W, NYC}, NE, NonNat, So | 0.72 |
| AfAm, Hisp/Lat, {MW/W, NYC, NE}, NonNat, So | 0.74 |
| **Emotion** | |
| PosNeu, SWAngry, SWFrstd, OtherSWNeg, VAngry, VFrstd | 0.32 |
| PosNeu, {SWAngry, SWFrstd, OtherSWNeg}, VAngry, VFrstd | 0.34 |
| PosNeu, {SWAngry, SWFrstd, OtherSWNeg, VAngry, VFrstd} | 0.42 |

**Table 2**. Inter-labeler consistency in labeling dynamic and static speaker traits, showing the variations in Cohen's Kappa values when different classes are merged as denoted by parenthesized sets.

### 2.2. Consistency

To test annotation consistency, a subset of 200 calls (627 utterances) were annotated by two human labelers. Subsequently, this was used as a test set and the rest of the data as a training set. Consistency was measured with Cohen's Kappa[1], and this measure was also used to test the impact of merging various classes (e.g., dialect such as Midwest/West and New York City). Table 2 gives a summary of the measured Kappa for grouping various classes.

Among the tags, the highest inter-labeler agreement was observed for gender and dialect. The consistency improves when some of the classes are collapsed. For example, the inter-labeler agreement improves when age is annotated as: {*Youth*, *Youth/Adult*}, *Adult*, {*Adult/Senior*, *Senior*}. Similarly, when emotions are grouped into negative vs. others.

## 3. SPEAKER TRAITS FROM GENERAL ACOUSTIC AND PROSODIC FEATURES

We first investigate the use of general acoustic and prosodic features to distinguish speaker traits, which focuses more on how the speaker spoke and less on what specifically they said. For this, we use speech/non-speech HMM-based classifiers on cepstral features to identify speaker traits. The speech portions in an utterance were tagged with labels as described in the previous section and were modeled with one-state HMMs. The non-speech portions of all the utterances were modeled with a generic one-state noise model. Both speech and non-speech states were modeled with Gaussian Mixture Models, allowing them to take a variable number of mixture components. Standard Mel Frequency Cepstral Coefficient (MFCC), as used in automatic speech recognition (ASR), were used as input, without any normalization. In a second set of experiments, we added pitch information to the MFCC feature vector. The pitch measurements were computed using the Entropics *waves* program and normalized as in [15].

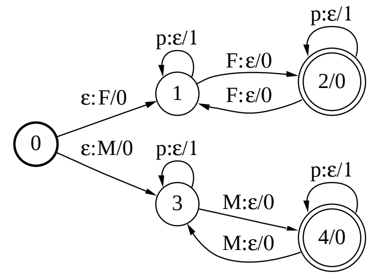


**Fig. 1**. Search space constrained to allow only paths that contain one type of tagged speech state (M or F), optionally, with non-speech (p) states.

For testing, Viterbi decoding was performed over the utterance. The state space was constrained as shown in Figure 1. Each path contained only one type of speech tag, which in some cases was drawn from the cross product of multiple trait labels. The finite-state transducer representing the constrained state space allowed insertion of non-speech states. The time synchronous search was pruned using beam widths as in ASR decoding. For experiments with unbalanced test sets, the priors were incorporated as unigram weights in the weighted finite-state transducers (WFSTs) representing the constrained state space [18].

The HMM-based classifiers were evaluated on the two feature sets mentioned above. The results are reported in Table 3. Accuracies for each trait were evaluated independently by four-fold cross-validation on test set which was balanced by class. The target classes were collapsed according to the groupings that gave maximal Kappa (shown in Table 2). Each trait was evaluated only over those utterances that had the same collapsed class label from both labelers. The balanced test set contained 368, 57, 125 and 276 utterances for gender, age, dialect and emotion respectively. In Table 3, "chance" denotes the accuracy achieved by a trivial classifier that assigns the most probable class label to all test points.

Gender detection performed quite well as expected. Additionally, conditioning the speech states on age and emotion, marginally improved gender detection. Classification of a speaker's age group performed best when modeled alone.

[1] Cohen's Kappa is a statistic often used to assess inter-rater reliability in coding categorical variables, a measure considered to be better than % agreement [5]. It takes values from 0 to 1, where higher value represents better reliability.

| Traits Jointly Modeled | Ceps | Ceps+Pitch |
|---|---|---|
| Gender (chance 50%) | | |
| Gender only | 94.6 | 95.4 |
| (Gender,Age,Emotion) | 95.9 | 95.9 |
| Age (chance 33.3%) | | |
| Age only | 68.4 | 70.2 |
| Dialect (chance 20%) | | |
| Dialect only | 43.2 | 44.8 |
| (Dialect,Gender,Emotion) | 44.8 | 50.4 |
| Emotion (chance 50%) | | |
| Emotion only | 68.5 | 67.8 |
| (Emotion,Gender) | 68.8 | 69.2 |

**Table 3**. Classification accuracy using speech/non-speech HMMs, tagged with speaker traits, and evaluated on (a) cepstral features (Ceps), (b) cepstral features augmented with processed pitch (Ceps+Pitch).

Dialect classification produced the best results when it was modeled jointly with gender and emotion. Emotion produced best result when modeled jointly with gender. The classification of age, dialect and emotion improved further when cepstral features were augmented with pitch information. Although there were only two classes of emotions, it was always found beneficial to model the emotions at a finer level and re-assign the classes to the two categories after classification. This was found to be true for age and dialect as well.

In applications where the occurrence of classes is unequal, the priors can be incorporated into the WFST that represents the state-space constraints. This was evaluated on the full test set, containing 627, 445, 465 and 448 utterances for gender, age, dialect and emotion respectively, to obtain the following accuracies - 95.7% for gender, 64.5% for age, 61.4% for dialect, and 76.8% for emotion. However, the gain observed in augmenting the cepstra with processed pitch disappeared.

## 4. SPEAKER TRAITS FROM THE SPEAKER'S SPOKEN WORDS

We next investigate using the speaker's words to distinguish speaker traits. Here we are specifically making use of what the speaker said, not so much how he said it. We use a general discriminant classification technique that applies to the word lattices output by the recognizer. These lattices encode hypothesized words along with their estimated likelihoods of being correct.

In some tasks such as that of determining the class of emotions corresponding to an input speech utterance, it is natural to use the word transcriptions provided by the recognizer since some words or sequences of words may provide a strong indication of the class of emotions of the utterance. It is often preferable to use the full word lattice output by the recognizer since, in most cases, the word lattice contains the correct transcription.

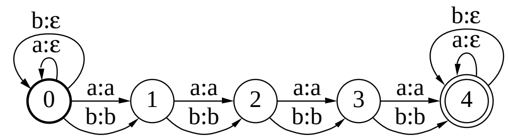


**Fig. 2**. Weighted transducer $T$ defining a 4-gram kernel.

### 4.1. Rational Kernels

Traditional statistical learning techniques have focused on classification of fixed-length input vectors associated to the objects to be classified. The application of discriminant classification algorithms to word lattices, or more generally weighted automata, requires generalizing these algorithms to handle distributions of alternative variable-length sequences.

We recently introduced a general framework, *rational kernels*, to extend statistical learning techniques to the analysis of weighted automata [7, 6]. Our approach is based on *kernels*. Kernel methods are widely used in statistical learning techniques such as Support Vector Machines (SVMs) [23] due to their computational efficiency in high-dimensional feature spaces [21]. Kernels are used for computing inner products or similarity measures between feature vectors in a closed form without explicitly extracting the feature vectors.

The general family of kernels we introduced, rational kernels, is based on weighted transducers or rational relations, which apply to weighted automata and thus word lattices. These kernels are efficient to compute using general weighted automata and graph algorithms and have been successfully used in other applications such as spoken-dialog classification.

### 4.2. $n$-gram Kernels

A kernel can be viewed as a similarity measure between two sequences or weighted automata. One may for example consider two sequences to be similar when they share many common $n$-gram subsequences.

This can be extended to the case of weighted automata or lattices over the alphabet $\Sigma$ in the following way. A word lattice $L$ can be viewed as a probability distribution $P_L$ over all strings $s \in \Sigma^*$. Denote by $|s|_x$ the number of occurrences of a sequence $x$ in the string $s$. The expected count or number of occurrences of an $n$-gram sequence $x$ in $s$ for the probability distribution $P_L$ is:

$$C_L(x) = \sum_s P_L(s)|s|_x$$

Two lattices output by a speech recognizer can then be viewed as similar when the sum of the product of the expected counts they assign to their common $n$-gram sequences is sufficiently high. Thus, we define an $n$-gram kernel $K$ for two lattices $L_1$ and $L_2$ by:

$$K(L_1, L_2) = \sum_{|x|=n} C_{L_1}(x)\, C_{L_2}(x) \quad (1)$$

We have proved elsewhere that $K$ is a rational kernel and that it can be computed efficiently [7]. There exists a simple weighted transducer $T$ that can be used to computed $C_{L_1}(x)$ for all $n$-gram sequences $x \in \Sigma^*$ [7]. Figure 2 shows that transducer in the case of 4-gram sequences ($n = 4$) and for the alphabet $\Sigma = \{a, b\}$. The general definition of $T$ is:

$$T = (\Sigma \times \{\epsilon\})^* \, (\sum_{a \in \Sigma} \{a\} \times \{a\})^n \, (\Sigma \times \{\epsilon\})^* \quad (2)$$

The kernel $K$ can thus be written in terms of the weighted transducer $T$ as:

$$K(L_1, L_2) \quad = \quad w[(L_1 \circ (T \circ T^{-1}) \circ L_2)] \qquad (3)$$

where $w[T]$ represents the sum of the weights of all paths of $T$. This shows that $K$ is a rational kernel whose associated weighted transducer is $T \circ T^{-1}$ and thus that it is *positive definite symmetric* (PDS), or equivalently that it verifies Mercer's condition [6]. This condition guarantees the convergence to a global optimum of discriminant classification algorithms such as SVMs. $K$ can be further used to construct other families of PDS kernels, e.g., *polynomial kernels* of degree $p$ defined by $(K + a)^p$.

### 4.3. Experiments

For our experiments in the binary emotion classification task, we used SVMs with rational kernels based on the $n$-gram kernels described in the previous section. For a given $n$, our *$n$-gram kernel* was defined as the sum of the $k$-gram kernels, $k = 1, \ldots, n$. Since the sum of two PDS rational kernels is a PDS rational kernel, our $n$-gram kernels are PDS rational kernels [6].

We first did experiments with the reference transcriptions of the utterances and $n$-gram kernels of different order combined with a second-degree polynomial. We then applied $n$-gram kernels to the word lattices output by the recognizer, our best results were obtained with $n = 4$.

Table 4 gives the classification accuracy obtained in each case over a test set of 448 utterances, mentioned in Section 3. The results show that rational kernels applied on word lattices ($80.6\%$) can achieve the accuracies close to that obtained from using the reference transcriptions ($81.7\%$).

Some authors have defined and used task-dependent sets of words, *salient words*, as features for classification [14]. We are not relying on the definition of a specific set of words

| | |
|---|---|
| HMM-based classifiers on Cepstra | 76.8 |
| SVM w/ rational kernels on ref. word transcripts | 81.7 |
| SVM w/ rational kernels on ASR word lattices | 80.6 |

**Table 4**. Comparison of two classifiers on the task of binary emotion classification. The baseline HMM-based classifiers use acoustic features. SVMs with rational kernels are applied to the word lattices output by the recognizer.

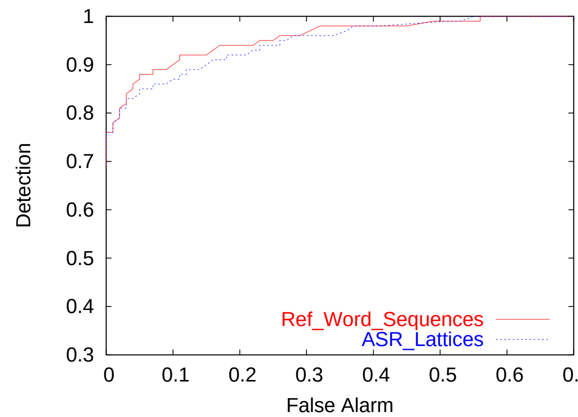


**Fig. 3**. ROC curves for binary emotion classification (negative emotion vs. other) using SVMs with rational kernels applied to reference word sequences and ASR word lattices.

and use as features all $n$-gram sequences. Using a subset of the words may introduce a bias. Furthermore, the generalization ability of SVMs does not suffer from the use of higher-dimensional feature space. Our use of kernels also helps preserve the computational complexity of our technique. Similar observations have been made in the case of spoken-dialog classification where the large-margin classification techniques currently used are not anymore based on salient words.

Figure 3 shows the Receiver Operating Characteristic (ROC) curves, comparing rational kernels applied to reference transcriptions versus to word lattices. The curves plot detection (true positive) versus false alarm (false positive).

Using the same framework, we observed that the spoken words did not provide a significant gain in accuracy over chance in classifying gender (72.6% vs.70.7%) into two classes and age (53.1% vs. 48.1%) into three classes. Dialect performed about as well as chance, indicating that the words alone are not sufficient to identify the dialect of the speaker. Perhaps, in a task-oriented dialog, there is limited choice of gender-, age- and dialect-specific words or phrases. Pronunciations and prosodic variations may hold more cues to identify a speaker's dialect.

## 5. CONCLUSION

We examined two sources of information for identifying speaker traits. The first corresponds more to *how* the speaker spoke. For this, we presented the results of several experiments using HMM-based classifiers to detect static or dynamic traits such as the speaker's gender, age, dialect, or emotion. The second represents more *what* the speaker specifically said. For this, we presented results using SVMs with rational kernels applied to word lattices. This approach performed $3.8\%$ better in the binary emotion classification task over the HMM-based approach.

An interesting avenue for future work is to investigate using both sources of information for classification. For example, rational kernels, which provide a general framework for discriminant classification, could be used with feature sets augmented to contain both kinds of information. Other features such as speaking rate and phone or word duration, which have been found to be useful in previous studies and are available only after a first pass decoding, could also be incorporated into this framework.

## 6. ACKNOWLEDGMENTS

The authors would like to thank several colleagues for their help and contributions: Richard Sproat for help in developing annotations for dialects, Patrick Haffner for making the Library for Large-Margin classification (LLAMA) available, Barbara Hollister and Ilana Bromberg for labeling the corpus, and Corinna Cortes and Daryl Pregibon for useful discussions.

## 7. REFERENCES

[1] L. Arslan and J. Hansen. Language accent classification in American English. *Speech Communication*, 188:353–367, 1996.

[2] A. Batliner, K. Fisher, R. Huber, J. Spilker, and E. Noth. Desperately seeking emotions: actors, wizards, and human beings. In *Proc. of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, pages 195–200, 2000.

[3] A. Braun and L. Cerrato. Estimating speaker age across languages. In *Proc. 14th Int'l Conference on Phonetic Sciences*, pages 1369–1372, 1999.

[4] L. Cerrato, M. Falcone, and A. Paoloni. Subjective age estimation of telephonic voice. *Speech Communication*, 31(2-3):107–112, 2000.

[5] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

[6] C. Cortes, P. Haffner, and M. Mohri. Positive definite rational kernels. In *Proceedings of The Sixteenth Annual Conference on Computational Learning Theory (COLT)*, Washington D.C., August 2003.

[7] C. Cortes, P. Haffner, and M. Mohri. Rational kernels. In *Advances in Neural Information Processing Systems (NIPS)*, volume 15, Vancouver, Canada, March 2003. MIT Press.

[8] R. Cowie. Emotional states expressed in speech. In *Proc. of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, pages 224–231, 2000.

[9] L. Devillers, I. Vasilescu, and L. Lamel. Emotion detection in task-oriented dialog corpus. In *Proc. of the IEEE Int'l Conference on Multimedia*, 2003.

[10] A. L. Gorin, A. Abella, T. Alonso, G. Riccardi, and J. H. Wright. Automated natural spoken dialog. *IEEE Computer Magazine*, 35(4):51–56, 2002.

[11] C. M. Lee and S. S. Narayanan. Emotion recognition using a data-driven fuzzy inference system. In *Proc. of European Conference on Speech Communication and Technology*, 2003.

[12] C. M. Lee, S. S. Narayanan, and R. Pieraccini. Recognition of negative emotions from the speech signal. In *Proc. Automatic Speech Recognition and Understanding*, 2001.

[13] C. M. Lee, S. S. Narayanan, and R. Pieraccini. Classifying emotions in human-machine spoken dialogs. In *Proc. of the IEEE Int'l Conference on Multimedia*, 2002.

[14] C. M. Lee, S. S. Narayanan, and R. Pieraccini. Combining acoustic and language information for emotion recognition. In *Proc. International Conference on Spoken Language Processing*, 2002.

[15] A. Ljolje. Speech recognition using fundamental frequency and voicing in acoustic modeling. In *Proc. IEEE Int'l Conference on Acoustic Signal and Speech Processing*, volume 3, pages 2137–2140, 2002.

[16] D. R. Miller and J. Trischitta. Statistical dialect classification based on mean phonetic features. In *Proc. International Conference on Spoken Language Processing*, volume 4, pages 2025–2027, 1996.

[17] N. Minematsu, M. Sekiguchi, and K. Hirose. Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers. In *Proc. IEEE Int'l Conference on Acoustic Signal and Speech Processing*, pages 137–140, 2002.

[18] M. Mohri, F. C. N. Pereira, and M. Riley. Weighted finite-state transducers in speech recognition. *Computer Speech and Language*, 16(1):69–88, 2002.

[19] V. A. Petrushin. Emotion recognition in speech signal: Experimental study, development, and application. In *Proc. International Conference on Spoken Language Processing*, 2000.

[20] K. R. Scherer. Emotional effects on voice and speech: Paradigms and approaches to evaluation. In *Proc. of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, 2000.

[21] B. Schölkopf and A. Smola. *Learning with kernels*. MIT Press: Cambridge, MA, 2002.

[22] C. Teixeira, I. Trancoso, and A. Serralheiro. Accent identification. In *Proc. International Conference on Spoken Language Processing*, volume 3, pages 1784–1787, 1996.

[23] V. N. Vapnik. *Statistical learning theory*. John Wiley & Sons, New-York, 1998.