EXPRESSIVE SPEECH IDENTIFICATIONS BASED ON HIDDEN MARKOV MODEL

Syaheerah L. Lutfi, J. M. Montero, R. Barra-Chicote, J. M. Lucas-Cuesta
Speech Technology Group, Technical University of Madrid, Spain

A. Gallardo-Antolin
Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Spain

Keywords: Affective computing, Biometrics, Speech processing, Emotion identification.

Abstract: This paper concerns a sub-area of a larger research field of Affective Computing, focusing on the employment of affect-recognition systems using speech modality. It is proposed that speech-based affect identification systems could play an important role as next generation biometric identification systems that are aimed at determining a person's 'state of mind', or psycho-physiological state. The possible areas for the deployment of voice-affect recognition technology are discussed. Additionally, the experiments and results for emotion identification in speech based on a Hidden Markov Models (HMMs) classifier are also presented. The result from experiment suggests that certain speech feature is more precise to identify certain emotional state, and that happiness is the most difficult emotion to detect.

1 INTRODUCTION

Biometrics has been a long used method in the field of information security and wide interest to biometrics is observed to arise in recent years. By definition, biometrics identification is a process to identify (or verify (Vaclave Jr. and Riha, 2000)) a person based on some feature of their biological makeup, with the help of automated systems that apply pattern recognition techniques. There are two types of biometrics identification techniques:

- *Behavioural:* the identification tasks are based on an individual's psychology tendency, style or preference such that through handwritten signature (Bullington, 2005, Gamboa and Fred, 2003), keystroke dynamics or speaker identification via his/her voice print (BenZeghiba et al., 2001). Behavioural biometrics is associated with less intrusive systems, minimizing ethical issues and leading to better acceptability by the users. On the other hand, behavioural identification systems need to be designed dynamically and receptive to variability (Gamboa and Fred, 2004, Gamboa and Fred, 2003, Ronzhin et al., 2004), considering the variability of most behavioural characteristics over time.
- *Physiological:* the identification tasks are based on invariant physical characteristics, such as finger print, hand silhouette, blood vessel pattern in the hand or back of the eye, DNA samples, patterning of the iris or the most popular technology , face recognition (Gamboa and Fred, 2003, Bullington, 2005). Though physiological identification technologies are based on more stable characteristics, usually data cannot be gathered passively, without the knowledge or permission of the person who is being observed.

Hence, these technologies often constitute an invasion of privacy and remains controversial as there were attempts of physiological biometric data being sold to interested third parties (Steinhardt, 2000).

While important in the area of biometrics, the focus of this paper is not the trade-offs between the security and privacy, but to highlight a possible foundation for the next generation of biometrics devices based on the process of the recognition of human emotion or affect via voice. The next sections discuss the importance of integrating the affective computing field to biometrics, followed by the

L. Lutfi S., M. Montero J., Barra Chicote R., Lucas-Cuesta J. and Gallardo-Antolín A. (2009).
EXPRESSIVE SPEECH IDENTIFICATIONS BASED ON HIDDEN MARKOV MODEL.
In *Proceedings of the International Conference on Health Informatics*, pages 488-494
DOI: 10.5220/0001556704880494
Copyright © SciTePress

possible use of vocal-affect biometrics, and the areas where this technology could be deployed. Please note that the words 'affect' and 'emotion' will be used throughout this paper interchangeably.

2 AFFECT RECOGNITION BIOMETRICS

Though 'affect recognition' systems such as the one used for surveillance, the 'HAL 9000' computer from the movie "2001: A Space Odyssey" ("I'm afraid, Dave…") or personal domestic- assistant robots in " I, Robot" that have good senses of their masters' emotional states, sound like science fiction, there is an interest in making them a reality (Bullington, 2005). For example, a recent proposal invitation from DARPA's Small Business Innovation Research Center calls for the development of a "non invasive emotion recognition system…suitable for deployment in military/operational environments or in environments in which discrete observation of potential enemy threats is desired" (DARPA, 2003). In addition, SRI lists 'Affective Computing' as one of their 'Next Generation Technologies:' "Affective-computing technology will reduce the intrusiveness of human-machine interface technology and perhaps make the technology more acceptable to people because of its more natural interactions and its seamless presence in the environment." (SRI-BI, N/A). Machine learning in recognition and adaptation to a human's affective state is important for natural human-computer interaction, but in biometrics, identifying a man's affective state could be aimed at determining a person's psycho-physiological state. So far, very little studies are found in emotion recognition for the purpose of biometric identifications. And the existing ones focus on *facial* emotion recognition biometrics. For example, Gray (Gray, 2003) has proposed a surveillance system that attempts to read a person's involuntary facial muscle-changes or what is termed as "microexpressions", that corresponds to his or her emotional state. Nevertheless, building an affect recognition system particularly based on voice for identification and possible intervention has almost never been attempted (Bullington, 2005).

2.1 Voice Affect Biometrics

In general, biometric analysis of speech aims at identification of a person. However, analysis on the emotion conveyed in the speech could reveal the presence of a particular *psychophysiological* state of the person via extralinguistic information (Ronzhin et al., 2004, Huang, 2001). Simply put, voice-affect identification could seek to predict an individual's 'state of mind' and reach judgments about his or her emotional states, impairments or behavioural intentions (i.e.: criminal intent) (SRI-BI, N/A)), despite the impostor deliberately trying to deceive the system, in some cases. Naturally, such a system will also recommend a possible course of prevention (Bullington, 2005). Physiological-based biometrics is constrained in terms of time, cost and detection of certain emotional 'colouring' such as boredom or fatigue. Additionally, these systems are dependent on human observers, invasive and impose ethical issues to some degree which make them unpopular (Bullington, 2005, Ronzhin et al., 2004). In contrary, voice-affect recognition systems are ideal and easy to deploy as they are individual oriented system, that does need human observers such that in facial-based recognition system. It uses machine learning algorithms to analyze and learn about the distinctive patterns of emotional states responding of individual users. Therefore, speech-based affect recognition is more natural, contact-free and offers high processing speed (Ronzhin et al., 2004). Though this kind of technology is unsuitable to be used in a crowd setting (i.e.: surveillance) on its own, it is ideal to be deployed in smaller scoped, specific high risk environment, where operator error could lead to serious problems such as injuries or fatalities. An example is the transportation industry or nuclear power plant. Apart from that, employment agencies or offices that recruit people for high-level security jobs such that in financial industry, military etc. can take advantage of this technology. In what follows, three possible areas for the implementation for voice-affect recognition technology are discussed:

1. Recruitment: Emotion detection in voice can estimate the psychophysiological state that leads to the determination of the psychological compatibility and the readiness of a candidate to accept a high security or stressful jobs. Industries such as nuclear, aerospace, transportation, financial etc. require workers that are fit for duty in these kinds of work nature. They may be able to determine when employees are not in the right state of mind to complete their tasks or determine the optimal conditions and most productive situations for each individual. Rohnzin et al (Ronzhin et al., 2004) proposed that the system takes into consideration the lexical and grammatical accuracy, apart from phrase understandability of a conversation during

testing. Then, the output is matched against a speech understanding model, to determine their psychological compatibility. Bullington (Bullington, 2005) pointed out that facial affect recognition system can be deployed to detect "possibly sleepy, intoxicated or distressed" worker to alert in-charged security personnel or deliver a warning to the worker himself. Such an application can be integrated with speech modality to enhance its functions.

2. Pathological detection and treatment: As mentioned earlier, analysis of voice can reveal hidden information, such as the true state of emotion, even if a person is faking it. Psychiatrists could benefit from this with regards to diagnosing and treatment of psychological disorders such as Post-Traumatic Stress Disorders (PTSD) (Castellanos et al., 2006, Morales-Perez et al., 2008). An example is the study from Morales-Perez et al (Morales-Perez et al., 2008) that utilizes a speech recognition application that employ time-frequency transformation techniques to extract relevant speech features correlating to both time and frequency domains. This information gives pathological information for detecting the anxiety level of a patient, thus providing the suggestion of the suitable treatment required.
3. Group Decision Support: Affective data obtained from speech analysis provides a more reliable indicator about the level people's feelings (strength of support or dissent) about an issue compared to a typical survey-based model (Bullington, 2005).

The next sections present the speech emotion identification experiments and the results.

3 EMOTION IDENTIFICATION IN SPEECH

This section presents the speech emotion identification experiments and the results obtained

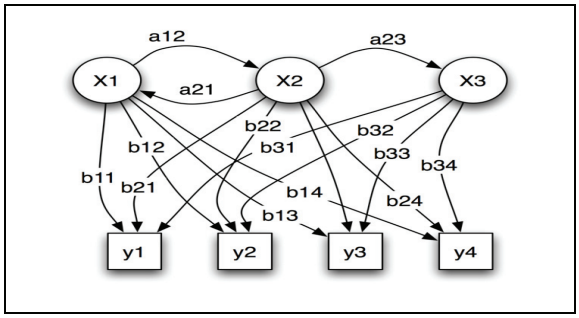
3.1 Experiments

The experiments are run on a classifier based on Hidden Markov Models (HMMs), which is implemented on the Hidden Markov Model Toolkit (HTK) (Young et al., 1995), originally developed for speech recognition. The output of the HMM classifier are 'fake' phonemes that are correlated to emotion with the highest probability among all emotion models together with its score value (accuracy). This will be assumed as the emotion that is identified from the speech input.

3.2 How HMM Works

HMMs are widely used in the field of pattern recognition. Their original application was in speech recognition (Rabiner, 1989). Given an HMM with a predefined architecture, there exist some well-established training algorithms to automatically optimize the parameters of that architecture. An example is the Baum-Welch training procedure (Rabiner, 1989) which uses the Maximum Likelihood Estimation (MLE) criterion. However, the architecture of HMM as well as the number of Gaussian mixtures per state and the number of training iterations are usually empirically determined. This section's summarizes how HMMs work, based on (Gunter and Bunke, 2004). When using HMMs for a classification problem, an individual HMM is constructed for each pattern class. For each observation sequence, i.e. for each sequence of feature vectors, the likelihood that this sequence was produced by an HMM of a class can be calculated. The class whose HMM achieves the highest likelihood is considered as the class that produced the actual sequence of observations. An HMM consists of a set of *states* and *transitions* probabilities between those states. One or several of the states are defined as final states. For each state a likelihood value for each possible observation is defined. If there is a finite number of observations then a probability for each observation, i.e. feature vector, is defined, but if we have continuous observation vectors a probability distribution is used. Each instance of a state in the training set has an impact on the training and leads to better parameter estimation. This software employs the Baum-Welch algorithm for training and the Viterbi algorithm for recognition. These algorithm are used to define the likelihood of an observation sequence for a given HMM. Viterbi recognition uses the highest likelihood of all possible state sequences, and the Baum-Welch recognition considers the *sum* of the likelihoods of all possible state sequences (Huang et al., 2001). The output of the HMM classifier are phonemes that are correlated to an emotion with the highest probability among all emotion models together with its accuracy value.

The figure below depicts the state, iteration and transition probabilities between states in an HMM:



Legend:
x — states
y — possible observations
a — state transition probabilities
b — output probabilities

Figure 1: Probabilistic parameters of a hidden Markov model (Rabiner, 1989).

3.3 Data Set

The use of acted speech recordings, though widely criticized (Douglas-Cowie et al., 2000, Campbell, 2000), is chosen over natural speech for the experiments reported in this paper. Though we are well aware that natural speech corpus such as Speech under Simulated and Actual Stress Database (SUSAS) (Hansen et al., 1998) is more suitable for this kind of study, we chose to use a readily available corpus as a pilot study. A Spanish corpus, The Spanish Emotional Speech corpus (SES), which is described in detail in (Montero et al., 1998) is used. It is a speaker-dependent database that contains three emotional speech recording sessions played by a professional male actor in an acoustically-treated studio. The data taken are from three sessions of recordings of paragraphs and each session contains 4 paragraphs (12 in total). The recordings simulate three primary emotions (sadness, happiness and cold anger), a secondary emotion (surprise) and a neutral speaking style. Each of these paragraphs are recorded in three sessions of each emotion, with exception to neutral, which was recorded only in two sessions, as it is assumed that there would not be a significant difference in any two different neutral speaking style (producing 168 paragraph recordings in all of the emotional states). From this, training dataset contains 165 paragraphs while recognition set consists of the *sentences* from 3 remaining paragraphs for all the emotional states, excluding cold anger (151 sentences). Therefore it should be noted that this is not a text-independent experiment and the result could be influenced by the fact that data have been segmented differently (multiple silence points in the paragraphs of training set; silence-speech-silence-speech-….-silence, and only two silence points in the recognition set; silence-speech-silence. The texts used carry non-emotionally inherent contents; they do not convey any explicit emotional content.

3.3.1 Feature Extraction

All utterances were processed in frames of 25ms window with 10ms frame shift. The two common signal representation coding techniques employed are the Mel-frequency Ceptral Coefficient (MFCC) and Linear Prediction Coefficient (LPC) and are improved using common normalization techniques; Cepstral Mean Normalization (CMN).

Feature extraction in speech recognition system consists of two steps. Kumar and Andreou (Kumar and Andreou, 1998) points out that the dimensionality of the speech waveform is reduced by using procedures such as cepstral analysis, or other form of analysis motivated by human speech perception. In the latter, the dimensionality of the obtained features vector is increased by an extended form of features vector that includes derivative and acceleration information. Therefore, in addition, experiments with speech identification that exclude accelerations and derivatives information are also carried out.

3.3.2 States

As the speech representing the feature combination will have dynamic properties, it is also natural to consider using more than one state in the HMM; common choices are to use three states.(Donovan and Eide, 1998, Donovan and Woodland, 1995). This can be formally thought of as modeling the beginning, middle and end of the sentence file. In the experiments reported here, the training set contains 'forced' silences in the beginning and at the end of the sentence. For this reason, the actual minimum number of state is one. Additionally, the experiment is run using two and three states per HMM, to determine any improvements. This is because it is assumed that each instance of a state in the training set has an impact on the training and leads to better parameter estimation.

3.3.3 Training Iterations and Gaussians

The numbers of training iterations used in all the experiments are from six (6) to thirteen (13), and up to thirty (30) Gaussians per state**.**

4 RESULTS

From the base experiments, a number of iterations with highest accuracy average are chosen. From this, experiments employing training iterations 6, 7 and 9 have the optimum average accuracy. Therefore, the results presented below uses the information from one of those experiments, specifically with *6 training iterations*, for both base and normalized features (with and without derivations and accelerations). The results of emotion identifications are measured in terms of error rates.

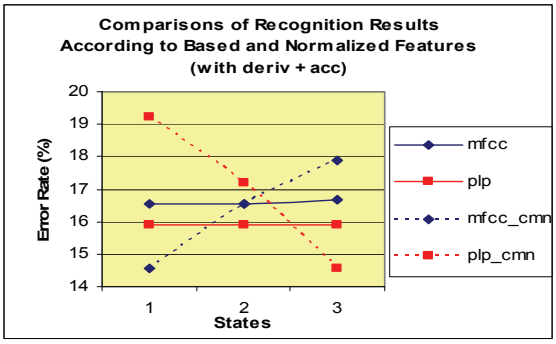


Figure 2: Recognition Results according to Features with 30 Gaussians per state, and 6 training iterations for all states, before and after normalizations (with derivations and accelerations).

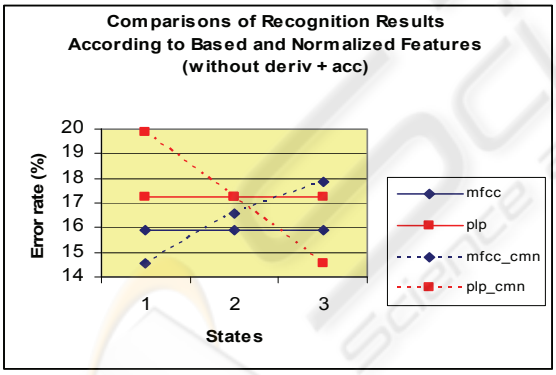


Figure 3: Recognition Results according to Features with 30 Gaussians per state, and 6 training iterations for all states, before and after normalizations (without derivations and accelerations).

Further, Table 1 and 2 below show the comparisons of confusion matrix for emotion identifications for base and normalized features respectively. The last row of the tables show the *precision* of the identification task, in other words, how precise the identification is. This is defined by the number of times the intended emotions are correctly identified in proportion of the total number of identifications (regardless of the accuracy).

5 DISCUSSION

Comparing the base features with derivatives and accelerations in Figure 2, PLP is shown as a slightly better feature, with error rate of 15.9 percent. The horizontal lines representing both the base features show that the number of states does not make much difference. However, after normalization is applied, number of states seems to influence the parameter estimation. After normalization, the optimum number of states for MFCC is one (1) while the worst is three (3). It is however, reversed with the PLP feature. The fact that in certain number of states (i.e.:MFCC with 3 states) worst parametric estimation is produced compared to the base-feature results implies that normalization may cause some important speech information to be eliminated. As for testing the features with and without derivatives and accelerations, the comparison of Figure 2 and Figure 3 clearly depicts that for test-set of 6 training iterations, results for base-PLP feature without derivatives and accelerations declined while it was the opposite for MFCC, though the Level of Confidence (LOC) for the differences of both features were not significant. In contrast, the normalized features without derivatives and accelerations are almost equal to that of features with derivatives and accelerations. Finally, the confusion matrix for emotion identifications of both base and normalized features in Table 1 and 2 shows that *happiness* is the most difficult emotion to detect. Additionally, Table 1 shows that MFCC is a more precise feature at identifying *happiness*. These results for *happiness* identification is also similar to the identification by human listeners using the same corpus in (Barra et al., 2006). As for the identification of the rest of the emotions, both features are almost equal, before and after normalization. Perhaps the use of different features is essential when it comes to the target emotions to be identified, as suggested by the results shown in the Tables.

ACKNOWLEDGEMENTS

This work has been partly funded by the Spanish Ministry of Science and Innovation with the contract: ROBONAUTA (DPI2007-66846-c02-02).

Table 1: Comparisons of Confusion Matrix for Emotion Identification between Base MFCC and PLP Features.

	IDENTIFIED EMOTIONS (%)									
INTENDED EMOTIONS	**Hap**		**Sur**		**Sa**		**Neu**		**Ang**	
Hap	**46,3**	**75,6**	22.0	19,5	0	0	4,87	0	26,8	4,9
Sur	2,22	24,4	**97,8**	**75,6**	0	0	0	0	0	0
Sa	0	0	0	0	**96,8**	**93,5**	3,2	6,5	0	0
Neu	0	0	0	0	0	0	**97**	**97**	2,94	2,94
Precisions	95	73,8	83	80,9	100	100	91,7	94,3		

mfcc	
plp	

Table 2: Comparisons of the Confusion Matrix for motion Identification between Normalized MFCC and PLP Features.

	IDENTIFIED EMOTIONS (%)									
INTENDED EMOTIONS	**Hap**		**Sur**		**Sa**		**Neu**		**Ang**	
Hap	**56,1**	**46,3**	29,3	46,3	2,4	0	0	2,4	12,2	4,9
Sur	6,7	2,2	**93,3**	**95,6**	0	2,2	0	0	0	0
Sa	0	0	0	0	**100**	**100**	0	0	0	0
Neu	8,8	2,9	0	0	0	0	**88,2**	**94**	2,94	2,94
Precisions	79,3	90,5	77,8	69,3	96,9	96,9	100	97		

mfcc	
plp	

REFERENCES

Barra, R., Montero, J. M., Macias, J., D'Haro, L. F., Segundo, R. S. & Cordoba, R. D. (2006) Prosodic and segmental rubrics in emotion identification. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toulouse.

Benzeghiba, M. F., Bourlard, H. & Mariethoz, J. (2001) Speaker verification based on user-customized password. . Institut Dalle Molle d'Intelligence Artificial Perceprive.

Bullington, J. (2005) 'Affective' computing and emotion recognition systems: The future of biometric surveillance? *Information Security Curriculum Development (InfoSecCD)*. Kennesaw, GA, USA.

Campbell, N. (2000) Database of Emotional Speech. *ISCA Workshop on Speech and Emotion.* Belfast.

Castellanos, G., Delgado, E., Daza, G., Sanchez, L. G. & Suarez, J. F. (2006) Feature Selection in Pathology Detection using Hybrid Multidimensional Analysis. *IEEE 2006 International Conference of the Engineering in Medicine and Biology Society (EMBS '06)*. NY, USA.

Darpa (2003) Integrated system for emotion recognition for the enhancement of human performance detection of criminal intent. *DARPA SB032-038*. USA, DARPA.

Donovan, R. E. & Eide, E. (1998) The IBM trainable speech synthesis system. *ICSLP 98.*

Donovan, R. E. & Woodland, P. C. (1995) Automatic speech synthesiser parameter estimation using HMMS. *International Conference on Acoustic Speech Signal Processing.*

Douglas-Cowie, E., Cowie, R. & Schroder, M. I. (2000) A new emotion database: Considerations, sources and scope. *ISCA Workshop on Speech and Emotion.* New Castle, UK.

Gamboa, H. & Fred, A. (Eds.) (2003) *An identity authentication system based on human computer interaction behaviour*, ICEISS Press.

Gamboa, H. & Fred, A. (2004) A Behavioural Biometric System Based on Human Computer Interaction. *Proc. of SPIE,* 5404.

Gray, M. (2003) Urban Surveillance and Panopticism: will we recognize the facial recognition society? *Surveillance and Society,* 1**,** 314-330.

Gunter, S. & Bunke, H. (2004) HMM-based handwritten word recognition: on the optimization of the number

of states, training iterations and Gaussian components. *Pattern Recognition,* 37**,** 2069-2079.

Hansen, J. H. L., Bou-Ghazale, S. E., Sarikaya, R. & Pellom, B. (1998) Getting started with the SUSAS: Speech under simulated and actual stress database.

Huang, X. (2001) *Spoken Language Processing*, Prentice Hall.

Huang, X., Acero, A. & Hon, H.-W. (2001) *Spoken Language Processing: A guide to theory, algortihm and system development,* New Jersey, Prentice Hall.

Kumar, N. & Andreou, A. G. (1998) Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Communication,* 26**,** 283-297.

Montero, J. M., Gutirrez-Arriola, J., Palazuelos, S., Enriquez, E. & Pardo, J. M. (1998) Spanish emotional speech from database to TTS. *ICSLP.* Sydney.

Morales-Perez, M., Echeverry-Correa, J., Orozco-Gutierrez, A. & Castellanos-Dominguez, G. (2008) Feature Extraction of speech signals in emotion identification. *IEEE 2008 International Conference of the Engineering in Medicine and Biology Society (EMBS '08).* Vancouver, Canada.

Rabiner, L. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE,* 77**,** 257-285.

Ronzhin, A. L., Lee, I. V., Karpov, A. A. & Skormin, V. A. (2004) Automatic estimation of human's psychophysiological state by speech. *9th Conference of Speech and Computer (SPECOM '04).* St. Petersburg, Russia.

Sri-Bi (N/A) Next-Generation Technologies: Affective Computing. *SRI Consulting Business Intelligence.*

Steinhardt, B. (2000) Face-off: Is the use of biometrics an invasion of privacy? *Network World.* 05/08/00 ed., Network World Inc.

Vaclave Jr., M. & Riha, Z. (2000) Biometric authentication systems., ECOM-MONITOR.

Young, S. J., Jansen, J., Ordell, J. J., Ollason, D. & Woodland, P. C. (1995) The HTK Hidden Markov Model Toolkit Book. Entropic Cambridge Research Laboratory.