EMOTION DETECTION WITH MULTIMODAL FUSION USING SPEECH - A REVIEW

R. V. Darekar[1], A. P. Dhande[2]
[1]Research Scholar, [2]Professor
[1]Pardmashree Dr. D. Y. Patil Institute of Technology, Pimpri, Pune (Maharashtra) INDIA
[2]Department of Electronics & Telecommunication Engineering, Pune Institute of Computer Technology, Pune (Maharashtra) INDIA

***Abstract*: - Speech is the most natural form of communication. Very less work has been carried out by fusion of speech parameters. This paper attempts to review the fusion of speech parameters to recognize emotions in human speech. Emotion recognition and verification is the process of determination of the psychological state of the speaker. Every emotion comprises different vocal parameters exhibiting diverse characteristics of speech. These features can be extracted using efficient parameters like MFCC coefficients, energy, pitch etc., If the fusion results of these parameters is the input to the trained neural net, it will be possible to analyze correct emotion of speaker. Harnessing the approaches of signal processing and pattern recognition algorithms, a smart and emotions specific man-machine interaction can be achieved. Neural networks can be explored to model the prosodic parameters of the syllables from their positional, contextual and phonological features. It may also be possible to search effective emotion parameters and recognize emotions accurately. In future, it will lead to creation machine capable to work considering human emotions.**

***Keywords*- Emotion detection using speech, multimodal fusion, MFCC, Energy, Pitch.**

I. INTRODUCTION

Emotion can be expressed as a complex set of interactions among subjective and objective factors, mediated by neural/hormonal systems. Speech samples and some efficient parameters can be used as database for emotion recognition. From speech signal, parameters that can be extracted are prosodic and spectral features such as pitch, energy, form ants, speech rate, Mel frequency Cepstrum coefficient and linear prediction Cepstrum coefficient. For database generation, there should be some criteria that can be used to judge how well a certain emotional database simulates a real-world environment. There are very few benchmark databases that can be shared among researchers. M ost of the databases share the emotions like anger, joy, sadness, surprise, boredom, disgust, and neutral. Psychological studies have shown that changes in human emotions reflect through prosodic Parameters of speech. Prosodic features can be selected like phonation time, speech rates, basic frequency averages, basic frequency ranges, basic frequency change rates, amplitude averages, amplitude change ranges, formant change averages, formant change ranges to analyze and find the structural characteristics and distribution patterns of different emotional signal features. There are many approaches towards automatic recognition of emotion in speech by using different feature vectors. Feature vectors can be classified as long-time and short-time feature vectors. The long-time ones are estimated, over the entire length of the utterance, while the short-time ones are determined over window of usually less than 100 ms. The long-time approach identifies emotions more efficiently. Short time features uses interrogative phrases which has wider pitch contour and a larger pitch standard deviation. Speech signal is composed of large number of parameters which indicates emotion content of it. Changes in these parameters indicate changes in the emotions. Therefore proper choice of feature vectors is one of the most important tasks. The specific parameter will effectively represent detected emotion but the fusion of number of detections with appropriate weight will increase emotion recognition accuracy.

II. FEATURE EXTRACTION

Most common features for feature extraction can be summarized as –

Pitch: Bäzinger et al. argued that statistics related to pitch conveys considerable information about emotional status. Yu et al. have shown that some statistics of the pitch carries information about emotion in Mandarin speech. Pitch can be extracted from the speech waveform using a modified version of the RAPT algorithm for pitch tracking implemented in the VOICEBOX toolbox. Using a frame length of 50ms, the pitch for each frame was calculated and placed in a vector to correspond to that frame. If the speech is unvoiced, the corresponding marker in the pitch vector was set to zero. The following statistics are calculated from the pitch and used in the pitch feature vector:

- Mean, Median, Variance, Maximum, and Minimum (for the pitch vector and its derivative)
- Average energies of voiced and unvoiced speech
- Speaking rate (inverse of the average length of the voiced part of the utterance) Hence, the pitch feature vector is 13-dimensional. The value of pitch frequency can be calculated in each speech frame and the statistics of pitch can be obtained in the whole speech sample. These statistical values reflect the global properties of characteristic parameters. For most practical purposes the pitch is just the frequency.

Energy: This is the basic and most important feature in speech signal. We can obtain the statistics of energy in the whole speech sample by calculating the energy, such as mean

value, max value, variance, variation range, contour of energy. Energy Intensity represents loudness of an audio signal, which is correlated to amplitude of signal. Energy Entropy expresses abrupt changes in the energy level of an audio signal. In order to calculate this feature, frames are further divided into K-sub windows of fixed duration.

Qualitative Features: Emotional contents of an utterance are strongly related with its voice quality. The voice quality can be numerically represented by parameters estimated directly from speech signal. The acoustic parameters related to speech quality are voice level, voice pitch, phrase phoneme and temporal structures.

Form ants and related features: Tracking form ants over time is used to model the change in the vocal tract shape. The use of Linear Predictive Coding (LPC) to model formants is widely used in speech synthesis. Prior work done by Petrushin suggests that formants carry information about emotional content. The first three formants and their bandwidths were estimated using LPC on 15ms frames of speech. For each of the three formants, their derivatives and bandwidths, we calculate the mean, variance, maximum and minimum across all frames. We calculate the mean, variance, maximum and minimum of the mean of each formant frequency, its derivative and bandwidth. The formant feature vector is 48-dim ensional.

Linear Prediction Cepstrum Coefficients: L PCC embodies the characteristics of particular channel of speech. Person with different emotional speech will have different channel characteristics, so we can extract these feature coefficients to identify the emotions contained in speech. The computational method of LPCC is usually a recurrence of computing the linear prediction coefficients (LPC), which is according to the all-pole model.

Wavelet Based features: Speech signal is a non-stationary signal, with sharp transitions, drifts and trends which is hard to analyze. Wavelets have energy concentrations in time and are useful for the analysis of transient signals. A time frequency representation of such signals can be performed using wavelets. The Discrete Wavelet Transform (DWT) is computed by successive low-pass and high-pass filtering of the discrete time-domain signals. Speaker emotional state identification applications the Discrete Wavelet Transform offers the best solution. By employing feature extraction technique number of features can be extracted from the emotional speech. To achieve accurate identification of emotion classifier should provided with single best feature. Therefore there is need of systematic feature selection to reduce unuseful features from the base features. To select best features Forward Selection method can be used. The remaining features can be used by classifier to increase classification accuracy.

Mel-Frequency Cepstrum Coefficients: The mel scale, named by Stevens, Volkman and Newman in 1937 is a scale of pitches judged by listeners to be equal in distance from one another. The reference point between this scale and normal frequency measurement is defined by assigning a perceptual pitch of 1000 mels to a 1000 Hz tone, 40 dB above the listener's threshold. Above about 500 Hz, larger and larger intervals are judged by listeners to produce equal pitch increments. As a result, four octaves on the hertz scale above 500 Hz are judged to comprise about two octaves on the Mel scale. The name Mel comes from the word melody to indicate that the scale is based on pitch comparisons. All Psychophysical studies have shown that human perception of the frequency contents of sound for speech signals does not follow a linear scale. The actual frequency (f) is measured in Hertz (Hz) and a subjective pitch is measured on a scale called the mel scale. The mel frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point the pitch of a 1 kHz tone, 40 db above the perceptual hearing threshold is defined as 1000 mels. Therefore we can use the following approximate formula to compute the mels for a given frequency f in Hz.

$$\text{mel}(F) = 2595 \times \log 10(1 + f/700)$$

In the next step, we convert the log mel-spectrum back to time. The result is called the mel frequency Cepstrum coefficients (M FFC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis because the mel spectrum coefficients are real number. We can convert them to the time domain using the Discrete Cosine Transform (DCT). Emotional speech recognition systems perform two fundamental operations namely, signal modeling and pattern matching. Signal modeling represents the process of converting speech signal into a set of parameters. Pattern matching is the task of finding parameter sets from memory which closely matches the parameter set obtained from the input speech signal. After performing these two operations, one can perpetually categorize the emotions with respect to the arousal rate of the dialogue spoken with its valence i.e. the behavioral aspect of the person displaying the emotion.

III. SELECTION OF CLASSIFIER

Selection of classifier depends on the geometry of the input feature vector. Some classifiers are more efficient with certain type of class distributions, and some are better at dealing with many irrelevant features or with structured feature sets. Performance comparison of classifiers can be done on the same large and representative database. M ost advanced researches on a speaker independent mode achieve recognition rates from 55% to 95%, whereas humans could hardly reach emotion recognition rates of about 60% from unknown speakers. Different methods can be used for the recognition of required parameters. Some of this are-

Hidden Markov Model: It is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved

(hidden) states. A H M M can be considered the simplest dynamic Bayesian network. The mathematics behind the HM M was developed by L. E. Baum and coworkers. It is closely related to an earlier work on optimal nonlinear filtering problem (stochastic processes) by Ruslan L. Stratonovich, who was the first to describe the forward-backward procedure. Hidden M arkov Model (HMM) is having the long history in the field of speech applications. The HMM consist of the first order markov chain whose states are hidden from the observer therefore the internal behavior of the model remains hidden. The hidden states of the model capture the temporal structure of the data. Hidden M arkov M odels are statistical models that describe the sequences of events. HMM is having the advantage that the temporal dynamics of the speech features can be trapped due to the presence of the state transition matrix. In this stage of our study, extracted audio features are used to train HM M for distinguishing between happy, sad, angry and aggressive em otions.

Gaussian mixture model: Gaussian mixture model allows training the desired data set from the databases. GM M are known to capture distribution of data point from the input feature space, therefore GM M are suitable for developing emotion recognition model when large number of feature vector is available. Given a set of inputs, GM M refines the weights of each distribution through expectation-maximization algorithm. GMMs are suitable for developing emotion recognition models using spectral features, as the decision regarding the emotion category of the feature vector is taken based on its probability of coming from the feature vectors of the specific model. Gaussian Mixture Models (GMMs) are among the most statistically matured methods for clustering and for density estimation. GM Ms are widely used as probability distribution features, such as vocal-tract related spectral features in a speaker recognition or emotion recognition systems. GMMs having advantage that are more appropriate and efficient for speech emotion recognition using spectral feature of speech.GM M is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities They model the probability density function of observed data points using a multivariate Gaussian mixture density. Here we have considered six emotional states namely Happy, Angry, Sad, Neutral and boredom.

Support Vector Machine: The support vector machine is a learning algorithm which addresses the general problem of learning to discriminate between positive & negative members of given n - dimensional vectors. The SVM is used for classification & regression purpose. The main idea of SVM classification is to a transform the original input set to a high dimensional feature space by using kernel function. The standard SVM takes a set of input data and predicts, for each given input, which of two possible classes the input is a member of, which makes the SVM a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of four categories, SVM testing algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. SVM constructs a hyper plane or set of hyper planes in a high- or infinite-dimensional space, which can be used for classification. A good separation is achieved by the hyper plane that has the largest distance to the nearest training data points of any class, since the larger the margin the lower the generalization error of the classifier.

Artificial Neural Network: It's a classifier used for many pattern recognition applications. They are known to be more effective in modeling nonlinear mappings. Also, their classification performance is usually better than HMM and GMM when the number of training examples is relatively low. Almost all ANNs can be categorized into three main basic types: M LP, recurrent neural networks (RNN), and radial basis functions (RBF) network. The classification accuracy of ANN is fairly low compared to other classifiers. The ANN based classifiers may achieve a correct classification rate of 51.19% in speaker dependent recognition, and that of 52.87% for speaker independent recognition. ANN is an efficient pattern recognition mechanism which simulates the neural information processing of human brain. The ANN processes information in parallel with a large number of processing elements called neurons and uses large interconnected networks of simple and nonlinear units. The quantitative modeling and processing of data using neural networks is effectively performed using the Supervised Learning Neural Network Back-Propagation Algorithm. For a given set of training input output pair, this algorithm provides a procedure for changing the weights in a back-propagation network (BPN) to classify the input patterns correctly. The aim of this neural network is to train the network to achieve a balance between the networks ability to respond (memorization) and its ability to give reasonable responses to the input that is similar but not identical to the one that is used in training. A BPNN is a multi-layer, feed-forward neural network consisting of an input layer, a hidden layer and an output layer. The hidden layers are used to classify successfully the patterns into different classes. The inputs are fully connected to the first hidden layer, each hidden layer is fully connected to the next, and the last hidden layer is fully connected to the outputs. The trained Artificial Neural Network (A N N) is required to be tested with features extracted from the test-utterances. The ANN is trained with multiple voice samples taken at different instances uttering the same phrase at all times. The total log-likelihood of these test-vectors of one test utterance with respect to the trained matrix corresponding to each emotion-class is computed. The test utterance is considered belong to that emotion-class with respect to which the total log-likelihood becomes the largest. An ambiguity may arise when surprise may be expressed along with any other emotion such as anger-surprise, happy- surprise, etc. Also some of the emotions like surprise-anger, surprise-happy, anger-happy and

sad-neutral appear to have similar acoustic characteristics. So a Confusion-M atrix can be prepared which would take care of these uncertainties up to a certain level.

IV. DISCUSSION

The use of three different language databases for emotion recognition has resulted in the higher level of accuracy. The higher recognition accuracies can be obtained with the fusion of feature extracted with Pitch, Energy and M FCC and ANN as the classification technique. Speech features such as spectral and prosodic feature can be extracted from emotional speech samples such as pitch, energy, formant frequency, speech rate, M FCC and by fusion of results, performance of system will get increased For accurate emotional speech database system will provide more efficiency An important task in speech emotion recognition system is selection of classifier. After calculation of Speech features, the appropriate features are provided to the classifier. A classifier detects the emotion from speech utterances of different speakers. ANN is having ability to find nonlinear boundaries for separating the emotional states. in speech emotion recognition Multilayer perception layer neural networks are commonly used because it has well defined training algorithm as it is relatively easy to implement. Most frequently used feed forward neural network for purpose of speech emotion recognition. In databases for emotion recognition, it is common to record the same sentence with different emotions, thus reducing the effect of lexical content on perceived emotions. This suggests that, a particular lexical content can be expressed in more than one type of emotion. In such cases, it can be seen from our data mentioned, that suitable emotiphons can be used by speakers, to effectively express the respective emotion. Improvement in speech emotion recognition performance has been attempted by combining other information such as facial expressions or specific words along with acoustic correlates. It has been shown that searching for emotional keywords or phrases in the utterances and integrating linguistic classifier with acoustic classifier have improved emotion classification accuracy (Ayadi et al., 2011). Computational techniques used in these approaches could be varied depending on the sophistication of the system application. The emotiphons discussed in this paper would be an additional source for emotion recognition. The presence of emotiphons heavily affects the prosody and convey emotions effectively. Some of the emotiphons are standalone and hence may be identified through a pre-processing stage, such as keyword spotting, whereas other emotiphons would have to be viewed along with prosody. Stochastic model based recognition would be required in most cases, because of the subjective variability of pronunciation. Cowie and Cornelius (2003) have described issues related to speech and emotion in great details, covering the basic concepts and relevant techniques to study conceptual approaches. It is well recognized that emotion analysis in human communication is multi-faceted and varied. It is also intertwined with the culture of the language users. Speech emotion recognition systems are similar to speech recognition, speech segmentation and isolated word recognition. A new framework for the emotion recognition can be envisaged to demonstrate the feasibility of integrating the new features derived by wavelet decomposition with the spectro-temporal and the baseline features (M FCC, pitch) in the recognition of human emotional states. The accuracy of emotion recognition and the robustness of the system deals with the extracted features set, classifier and the database. On the other hand, in case of emotion detection using image, when the facial expression changes, the shape and the location of the organs on the face, will change accordingly. It lacks off detailed characterization of changes in the eyes, nose, mouth and other facial organs. We need to build a correction model associated with the expressions containing a variety of rules, and modified the image recognition results using the model. In addition, in the term of real-time applications, besides enhancing the robustness of the system and improving the accuracy, the efficiency of the recognition algorithm is also a key factor. The automatic recognition of emotional states from human speech has found a broad range of applications, and as such has drawn considerable attention and interest over the recent decade. Speech emotion recognition can be formulated as a standard pattern recognition problem and solved using machine learning technology. Specifically, feature extraction, processing and dimensionality reduction as well as pattern recognition have been discussed in this chapter. Three short time cepstral features, Linear Prediction-based Cepstral Coefficients (L PCC), Perceptual Linear Prediction (PLP) Cepstral Coefficients, and Mel-Frequency Cepstral Coefficients (M FCC), are used in our work to recognize speech emotions. Feature statistics are extracted based on speech segmentation for capturing longer time characteristics of speech signal. In order to reduce computational cost in classification, Principal Component Analysis (PCA) is employed for reducing feature dimensionality. The Support Vector Machine (SVM) is adopted as a classifier in emotion recognition system. The experiment in the classification of 15 emotional states for the samples extracted from the LDC database has been carried out. The recognition accuracies achieved with different segmentation forms and different feature set sizes are compared for speaker dependent training mode. Research can also be preceded with differentiation between “opposing” emotional states. Six different “opposing” emotion pairs were chosen: despair and elation, happy and sadness, interest and boredom, shame and pride, hot anger and elation, and cold anger and sadness. For each emotion pair, we formed data sets comprising of emotional speech from all speakers, only male speakers, and only female speakers because the features are affected by the gender of the speaker. For example, the pitch of males ranges from 80Hz to 200Hz while the pitch of females ranges from 150Hz to 350Hz. This corresponds to a total of eighteen unique data sets. For each data set, we formed inputs to our classification algorithm comprising of feature vectors from: Pitch only, M FCCs only, Form ants only, Pitch & M FCCs, Pitch & Form

ants, M FCCs & Form ants, and Pitch, M FCCs & Form ants. Hence, for each emotion pair, the classification algorithm was run on twenty one different sets of inputs. The Indian subcontinent is a good example of a sprachbund (Emeneau, 1956), because there are two distinct language families, Indo-Aryan and Dravidian. However, there is a lot of interaction and similarity across the languages belonging to these two families due to centuries of language and culture contact. While grouping Indian languages using machine learning techniques, based on their text, it is observed that, Marathi is the closest language to the Southern zone consisting of the languages from Dravidian family and can be grouped with Hindi, Punjabi and Gujarati. The grouping corresponds well with the geographic proximity also (Ghosh et al., 2011). Emotiphons can be categorized in different groups. Phonetic representation of emotiphons is given using IPA symbols. A dditional characteristics are mentioned wherever they are significant. The emotiphons that are smallest expressions consisting a single vowel or a diphthong. They are "stand-alone" expressions, i.e., can be used in isolation to express the respective emotions and may not need conversation mode. It can be seen that the number of emotions covered by emotiphons are far more than those expressed in the databases mentioned in the literature for emotion recognition. Common emotions covered by the databases are anger, fear, joy, sadness, disgust, surprise and neutral, mainly in the prosody at the acoustic level. However, emotiphons express many more shades and nuances of emotions like affection, pain, disbelief, sympathy, boredom and so on, which are all important for the semantic context. Emotiphons can be grossly classified into three categories; Vowel like, Fricative like and Multi-phons. Emotiphons are exclamatory in nature. "Stand-alone" emotiphons are unaffected by the linguistic parameters such as gender and number. They cover large number of emotions. Although Its believed that many emotions would be common across all humans which is a Dravidian perspective (Hozjan and Kacic, 2003), we feel that expression of emotion is dependent on culture and society. Many emotiphons are common across Marathi and Kannada, suggesting that people using these languages share similar cultural values, although the languages belong to two different families. The common emotiphons across Marathi and Kannada are of "stand-alone" type and are independent of linguistic parameters such as gender and number. Its identified that there exist many emotion markers, referred to as emotiphons in two Indian languages, Marathi and Kannada belonging to Indo-Aryan and Dravidian language family, respectively. We find that emotiphons are short lexical expression used in conversational speech to convey many different specific emotions explicitly and effectively. Although Marathi and Kannada are from two different language families, we notice that there are many common emotiphons across the two languages. Commonality of emotiphons across the languages would lead us to understand cognitive aspects of the emotion communication as well as the linguistic evolution. Emotiphons would play a major role in identification of emotion in speech processing.

V. CONCLUSION

The application area of emotion recognition from speech is expanding as it opens the new means of communication between human and machine. It is needed to model effective method of speech feature extraction so that it can provide emotion recognition of real time speech. Success of emotion recognition is dependent on appropriate feature extraction as well as proper classifier selection from the emotional speech. It can be seen that Integration of various features can give the better recognition rate. Speech signals in comparison with image signal, for emotion detection can be considered to be more efficient as its possible to retrieve some hidden parameters from speech signals which generally cannot be changed by the speaker with ease. Also by fusion of three or more parameters, higher level of recognition accuracy can be achieved. The process can be made more efficient by selecting best classifier for appropriate emotions and hence can improve the detection results. The previous recognition results show that on the basis of prosodic information, we can initially recognize basic emotional categories, and apply it into the emotion recognition system. It limits the amount of storage, computation and doesn't have strict recognition accuracy. The strategies such as codebook pruning, data compression can also improve the recognition rate effectively. Multimodal recognition systems integrating with images, voice and other emotional information is the inevitable trend of future human-computer interaction development. Although there are still many insurmountable technical problems, with the continuous progress of science and unremitting efforts of the researchers, the real-time systems of multi-modal speech recognition will have more potential development.

REFERENCES

[1] Ian McLoughlin, "Applied Speech and Audio Processing with M ATLA B Examples", Cam bridge University Press, 2009.

[2] Beth Logan, "M el Frequency Cepstral Co-efficient for M usic Modeling", Proceedings of the International Symposium on Music Information Retreival (ISIM IR), Plymouth, MA 2000.

[3] Yashpalsing D. Chavhan and M.L. Dhore, "Speech Emotion Recognition using SVM" IEEE International Conference on „Computer Vision and Information Technology, Advances and Applications~, ACVIT-09, December 2009, pp. 799-804.

[4] L. Rabiner, B. H. Juang, "Fundamentals of Speech Recognition", Pearson Education, 2009.

[5] Madhavi S. Pednekar, Kavita Tiware and Sachi n B hagwat, "Continuous Speech Recognition for Marathi Language Using Statistical Method", IEEE International Conference on „Computer Vision and Information Technology, A dvances and Applications~, ACVIT-09, December 2009, pp.

810-816.

[6] Firoz Shah. A, Raji Sukumar. A, and Babu Anto. P, "Discreet Wavelet Transforms and Artificial Neural Networks for Speech Emotion Recognition", International Journal of Computer Theory and Engineering, Vol. 2, No. 3, 1793-8201, June 2010, pp.319-322.

[7] S.N. Sivanandam, S.N. Deepa, "Principles of Soft Computing", WILEY India, 2009.

[8] A.B. Kandali, A.B. Routray, Basu T.K., "Emotion Recognition From A ssamese Speeches Using M FCC And GM M Classifier", IEEE Region Conference TEN CON 2008,India, pp. 1-5.

[9] X.Huang, A.Acero, H.-W. Hon, "Spoken Language Processing", Prentice Hall PTR, 2001

[10]T. Dutoit, "An Introduction to Text-to-Speech Synthesis", Kluwer Academic Publishers

[11] D. Juraf sky and J.H. Martin, "Speech and Language Processing", Pearson Education, 2000 [12]H.Zen, K.Tokuda , &A.W Black " Statistical parametric speech synthesis", speech communication, doi:10.1016/j.specom.2009.04.004 2009

[13]L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition", In proc. of the IEEE, Vol. 71, no.2, pp.227-286, Feb 1989

[14]A.Falaschi, M.Guistianiani, M.Verola, "A hidden markov model approach to speech synthesis", In proc. of Eurospeech, Paris, France, 1989, pp 187-190

[15]S. Martincic- Ipsic and I. Ipsic, "Croatian H M M Based Speech Synthesis," 28th Int. Conf. Information Technology Interfaces ITI 2006, pp.19-22, 2006, Cavtat, Croatia

[16] S.S. Agrawal, " Speech Synthesis for Natural Sounding" 10th M.S. Narayana Memorial Lecture (Keynote address) delivered during NSA-2001, held at VIT, Vell ore(Tam i lN adu) ,2001

[17]Cahn, J. E., "Generating Expression in Synthesized Speech", Master's Thesis, M IT, 1 989 Cahn, J. E., The Generation of Affect in Synthesized Speech, Journal of the American Voice I/O Society, 8, July 1990, p. 1-19.

[18] Murray, I. R., "Simulating emotion in synthetic speech", PhD Thesis, University of Dundee,

[19] Murray, I. R., & Arnott, J. L., "Implementation and testing of a system for producing emotion-by-rule in synthetic speech", Speech Communication, 16, p. 369

[20] Montero, J. M., Gutiérrez-Arriola, J., Palazuelos, S.,Enríquez, E., Aguilera, S., & Pardo, J. M., " Emotional Speech Synthesis: From Speech Database to T-T-S", ICSLP 98, Vol. 3, p. 923-926. Burkhardt, F., "Simulation emotionaler Sprechweise mitSprachsyntheseverfahren" [Simulation of emotional manner of speech using speech synthesis techniques], PhD Thesis, TU Berlin, 2000

[21] Burkhardt, F., & Sendlmeier, W. F., "Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis",

[22] ISCA Workshop on Speech & Emotion, p. 151-156.

[23] S.Lemmetty, "Review of Speech Synthesis Technology", Master's Thesis, H eli nski University of Technology

[24] Heuft, B., Portele, T., & Rauth, M. (1996), "Emotions in Ti me Domain Synthesis" ICSLP

[25] Edgington, M., "Investigating the Limitations of Concatenative Synthesis", Eurospeech 97.

[26] Vroomen, J., Collier, R., & Mozziconacci, S. J. L., "Duration and Intonation in Emotional Speech", Eurospeech 93, Vol. 1, p. 577-580.

[27] Rank, E., & Pirker, H., "Generating Emotional Speech with a Concatenative Synthesizer", ICSLP 98, Vol. 3, p.67 1-674

[28] Montero, J. M., Gutiérrez-Arriola, J., Colás, J., Enríquez,E., & Pardo, J. M., "Analysis and Modeling of Emotional Speech in Spanish", ICPhS 99, p. 957-960.

[29] Iriondo, I., Guaus, et al., "Validation of an Acoustical Modeling of Emotional Expression in Spanish using Speech Synthesis Techniques", ISCA Workshop on Speech & Emotion, Northern Ireland 2000, p. 161-166. [31]Murray, I. R., Edgington, M. D., Campion, D., & Lynn., " Rule-based Emotion Synthesis Using Concatenated Speech", ISCA Workshop on Speech & Emotion, 2000, p. 173-177. [32]Schröder, M., "Can emotions be synthesized without controlling voice quality?" Phonus 4, Research Report of the Institute of Phonetics, University of the Saarland, p.37-55.

[33]Mozziconacci, S. J. L., "Speech Variability and Emotion: Production and Perception", PhD Thesis, Technical University, Eindhoven, 1998.