TASHKENT UNIVERSITY OF
INFORMATION TECHNOLOGIES
NAMED AFTER MUHAMMAD AL-KHWARIZMI

MUHAMMAD AL-XORAZMIY NOMIDAGI
TOSHKENT AXBOROT TEXNOLOGIYALARI
UNIVERSITETI

BULLETIN OF TUIT:

MANAGEMENT AND COMMUNICATION TECHNOLOGIES



ISSN:2181-1083 Date: 2025-09-20 DOI:
Furkat Rakhmatov, Fakhriddin Abdirazakov, Azizbek Temirov, Baxodir Achilov, Gulzira Khaldarova, Ruslan Baydullayev
№ 3(21)2025

A Model for Synthetic Speech Detection Based on Emotional Variations Using Artificial Intelligence

Furkat Rakhmatov
Tashkent university of information technologies named after Muhammad al-Khwarizmi,
Tashkent, Uzbekistan
f.raxmatov@tuit.uz

Fakhriddin Abdirazakov
Tashkent university of information technologies named after Muhammad al-Khwarizmi,
Tashkent, Uzbekistan
faxriddinabdirazzoqov@gmail.com

Azizbek Temirov
Tashkent university of information technologies named after Muhammad al-Khwarizmi,
Tashkent, Uzbekistan
aotemirov@gmail.com

Baxodir Achilov
Tashkent university of information technologies named after Muhammad al-Khwarizmi,
Tashkent, Uzbekistan
borya19861804@gmail.com

Gulzira Khaldarova
Tashkent university of information technologies named after Muhammad al-Khwarizmi,
Tashkent, Uzbekistan
khaldarovagulzira91@gmail.com

Ruslan Baydullayev
Tashkent university of information technologies named after Muhammad al-Khwarizmi,
Tashkent, Uzbekistan
ruslan.baydll88@gmail.com

Abstract— Today, the automatic recognition of emotions in human speech is gaining increased importance due to the growing role of speech interfaces in human-computer interaction applications. In this context, the present study analyzes several methods for detecting synthetic speech and proposes a classification approach based on emotion recognition. Five primary human emotions — anger, boredom, happiness, neutral state, and sadness — are considered, and speech that does not correspond to any of these emotional categories is investigated as potentially synthetic.To achieve more accurate results, statistical methods were employed to combine various feature streams. For emotion detection in speech, a feature vector was constructed by combining the following acoustic features: 16 Linear Predictive Coding (LPC) coefficients, 12 Linear Predictive Cepstral Coefficients (LPCC), 16 Log Frequency Power Coefficients (LFPC), 16 Perceptual Linear Prediction (PLP) coefficients, 20 Mel-Frequency Cepstral Coefficients (MFCC), and jitter.The proposed method is based on three classification techniques: Linear Discriminant Analysis (LDA), k-Nearest Neighbors (K-NN), and Hidden Markov Models (HMM). The experimental results demonstrate that the selected features are stable and effective in recognizing emotions and yield high performance based on two evaluation metrics.

Keywords **—** Speech variations, emotion recognition, LPC, LPCC, LFPC, PLP, MFCC.

1. INTRODUCTION

Research on understanding and modeling human emotions has traditionally been explored within the fields of psychology and linguistics. However, in recent years, it has attracted increasing attention in the domain of engineering as well. This growing interest is largely driven by the need to enhance the naturalness and effectiveness of artificial speech interfaces in human-machine interaction. Nevertheless, studying emotions remains highly complex for several reasons. One of the main challenges lies in the difficulty of providing a clear and

comprehensive definition of what emotion actually is. The diversity of interpretations across various researchers makes it difficult to establish a universal concept. A general overview and analysis of the most common definitions of emotion proposed by different scholars can be found in [1], where the key theoretical perspectives and foundational concepts are discussed in detail.

Research on the cognitive component of emotions focuses on understanding the contexts and situations that trigger emotional responses, whereas studies on the physiological component are directed toward analyzing the immediate bodily reactions that occur simultaneously with emotional experiences. Emotions are often considered as a form of interaction either within the individual or between individuals [2]. Traditionally, emotions are divided into two main categories: primary (basic) and secondary (derived) emotions [3]. Primary emotions are universal to all human speech and are associated with specific patterns of expression (such as vocal/facial expressions, behavioral tendencies, and physiological responses). Secondary emotions, in contrast, are viewed as combinations or derivatives of primary emotional states. The dimensional classification of emotions aims to simplify the representation of their essential characteristics. According to the theoretical model developed by [4] and further applied in subsequent psychological research [5], emotions can be quantitatively assessed along three main connotative dimensions: arousal, valence (positive–negative), and dominance (power). In general, these dimensions — intensity, valence, and dominance — are widely used to distinguish between basic emotional states.

The spatial distribution of emotions such as happiness, neutrality, anger, sadness, and boredom is illustrated in Figure 1, which provides a clear and simplified representation suitable for various practical applications of emotion modeling.

If an emotion falls within a recognized category, its characteristics are considered to reflect natural speech due to their alignment with typical human vocal behavior. However, when vocal characteristics deviate significantly from expected emotional patterns, the speech is likely to be synthetic, i.e., a signal generated by a computer. Rather than modifying the classifier architecture, this study presents a robust and efficient set of vocal features aimed at detecting synthetic speech through emotional expression. These vocal features are extracted with a focus on capturing emotional authenticity. Traditional vocal attributes—such as intonation contour, pitch level, speech rate, and duration—are often insufficient for accurately distinguishing between emotion pairs such as anger vs. happiness and boredom vs. sadness.

To overcome this limitation, the study employs emotion-related features linked to arousal and intensity levels, which facilitate the categorization of emotions into discrete classes.

A series of systematic experiments were conducted, and the extracted features were evaluated based on their effectiveness and distinctiveness. The results indicate that the proposed feature set is not only simple and interpretable, but also possesses strong discriminative power, making it suitable for reliable emotion-based synthetic speech detection.

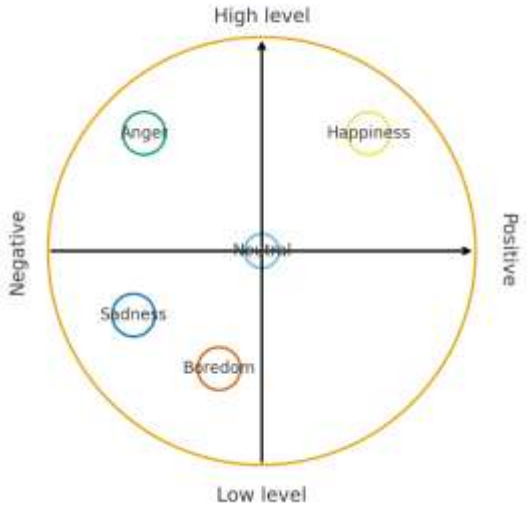


Figure 1. Spatial distribution of emotions

An emotional-speech corpus dedicated to emotion classification was compiled by the authors of [6]. The database comprises recordings of five primary emotions—anger, boredom, happiness, neutral state, and sadness—as well as short utterances of synthetic (artificial) speech, i.e., computer-generated voices designed to resemble human speech. The selection criteria for both natural and synthetic speech utterances are presented in Table 1.

Overall, this corpus provides a controlled benchmark for comparative analysis of natural versus synthetic speech within the same experimental framework, enabling the training and validation of models for fine-grained emotion recognition as well as the detection of artificial (computer-generated) speech.

Table 1. Speech Types and Their Key Characteristics

Speech Type	Key Characteristics
Natural speech	– Absence of exaggerated emotional intonation. – Sentences are suitable for naturally expressing diverse emotions (anger, sadness, happiness, boredom, neutral). – Utterances are simple and fluent. – Phonetically diverse, encompassing a wide range of speech sounds.
Synthetic (artificial) speech	– No bias toward exaggerated emotional coloring. – Capable of rendering all target emotions through text-to-speech (TTS) generation. – Short, uniform utterances that are easy to process with computational algorithms. – Phonetic diversity is considered to produce output that approximates natural speech.

2. CONSTRUCTION OF A SAMPLE SPEECH-SIGNAL DATASET

To avoid artifacts arising from acted delivery, non-professional speakers were recruited. In total, 12 native Uzbek speakers (7 female, 5 male) participated in reading sentences targeted to specific emotions. Recordings were conducted in a quiet environment using a close-talk (head-mounted) microphone at an 8 kHz sampling rate. Each native speaker was asked to produce every sentence in each of the five emotions (anger, boredom, happiness, neutral, sadness), yielding 1,200 utterances overall. After removing utterances contaminated by excessive noise, the emotional-speech corpus was subjected to subjective evaluation by human listeners. The purpose of this evaluation was to eliminate sentences with ambiguous affect; ultimately, 558 utterances that achieved a

listener-rated accuracy above 80% were retained. In addition, synthetic speech samples were created using text-to-speech (TTS) models based on deep neural networks to generate natural-sounding voices from text. The following steps were followed in generating the synthetic speech:

Text selection — sentences provided to speakers (typically emotionally neutral in content, yet suitable for being produced with different emotions).

Generation via a TTS model — the selected text is converted to speech.

Emotional modulation — prosodic parameters of the synthetic speech (intonation, speaking rate, and pitch) are adjusted to render anger, happiness, sadness, boredom, and a neutral state.

Subjective evaluation — human listeners assess the naturalness and the degree of emotional expressiveness of the generated speech.

The resulting speech corpus is used for two complementary research objectives: (i) classifying emotions in natural speech, and (ii) detecting synthetic (TTS-generated) speech produced by generative models and discriminating it from natural speech.

The use of sentence-level utterances, together with the availability of native speakers, facilitates reliable data collection. In general, speakers express emotions more naturally in their native language than in a foreign language. For practical applications—such as call-center analytics [7]—an 8 kHz sampling rate was employed to satisfy computation-time requirements and network-traffic constraints, and this resolution proved sufficient to meet the target performance. In a separate study [8], two professional speakers produced a total of 503 sentences covering five emotions (see Table 2). Those recordings were likewise downsampled to 8 kHz for subsequent processing.

Table 2. Example Emotional and Synthesized Speech Utterances (Set 1)

Emotion	Female	Male	Total	Synthetic speech	Total
Anger	75	76	151	50	201
Boredom	37	46	83	50	133
Happiness	56	40	96	50	146
Neutral	58	58	116	50	166
Sadness	54	58	112	50	162
Total	280	278	558	250	808

Table 3. Example Emotional and Synthesized Speech Utterances (Set 2)

Emotion	Female	Male	Total	Synthetic speech	Total
Anger	36	72	108	50	158
Boredom	72	72	144	50	194
Happiness	36	36	72	50	122
Neutral	36	36	72	50	122
Sadness	72	35	107	50	157
Total	252	251	503	250	753

In the sample speech-signal corpus, sentences are partitioned into two subsets: one for training and the other for testing. Models trained on the training subset are then

compared using the held-out test subset—this constitutes the simplest form of cross-validation. Relying on a single split, however, can introduce variance and occasional bias. A more robust strategy is to generate multiple train–test partitions and report the average performance across them. In the extreme case, with ppp total sentences, one performs leave-one-out cross-validation (LOOCV): for each of the p iterations, train on $p-1$ sentences and test on the single remaining sentence, cycling through all possibilities. The advantage of LOOCV is that it leverages nearly all available data for training in each iteration and avoids the need for a permanently separate test set. This evaluation protocol applies not only to the natural-speech corpus, but equally to synthetic (TTS-generated) and robotic speech recordings, enabling a consistent and reliable assessment of classification effectiveness across all categories.

Synthetic-Speech Detection via Emotional Cues. The proposed model for detecting synthetic speech comprises three stages: feature extraction, feature computation/encoding, and classification. First, core emotion-related descriptors and their statistical measures are extracted from the speech signal. Next, these descriptors are encoded into a dedicated feature vector. Finally, the vector is analyzed by a classifier to discriminate between natural and synthetic speech.

Extraction of Emotional Features. In detecting synthetic speech, emotions are treated as a critical factor, because the emotional expressiveness of artificial speech differs markedly from that of natural speech. For this reason, the careful selection of appropriate emotion-related features is of primary importance. All extracted features are required to contain a sufficiently informative set of cues to distinguish synthetic speech from natural speech. In [9], a wide range of acoustic attributes is recommended for recognizing the principal emotions. The features of the speech signal that are associated with emotions are summarized in Table 3. In natural speech, anger is typically manifested by a faster speaking rate, elevated pitch, and greater intensity, whereas in synthetic speech these cues tend to be simplified or homogenized. Accordingly, parameters such as pitch, intensity (energy), and articulation (e.g., clarity and dynamics of phonetic realization) play a decisive role in differentiating synthetic from natural speech.

Table 4. Emotions and Speech Features

Feature	Anger	Happiness	Sadness	Fear	Boredom	Synthetic speech
Speech rate	Slightly faster	Faster or sometimes slower	Slower	Very fast	Very slow	Stable, uniform rhythm
Mean pitch	Very high	High	Lower	Very high	Very low	Smooth contour, no abrupt changes
Pitch range	Wider	Wider	Narrower	Wider	Narrower	Restricted, near-monotone

Intensity (energy)	High	High	Low	High	Lower	Moderate and even
Voice quality	Harsh/rough	Bright	Resonant	Irregular/uneven	Thick/dull	Artificial, metallic or flat timbre
Articulation	Tense	Normal	Reduced	High clarity	Normal	Very clear, but lacking natural subtleties

Many acoustic features have been used in prior work; however, the authors of [10] argue that these features are often limited in their ability to capture the naturalness of emotional expression. For example, pitch- and energy-based descriptors are highly discriminative in natural speech, yet in synthetic speech their inherent variability and tension are insufficiently represented. To address this limitation, the present approach seeks to detect synthetic speech through a deeper analysis of emotion-related features. In addition to conventional acoustic measures such as MFCC, LPCC, and PLP, we incorporate parameters that reflect micro-perturbations in phonation—namely jitter and shimmer—to better capture subtle deviations characteristic of artificial speech [11]. These parameters are particularly discriminative for separating synthetic from natural speech, because micro-emotional fluctuations are almost absent in synthetic speech. The principal acoustic cues are categorized as follows: speech rate, pitch range, intensity, voice quality, and articulation. Using these cues enables the detection of emotions such as anger, happiness, sadness, fear, and boredom. At the same time, in synthetic speech these cues are insufficiently expressed, which limits the fidelity of emotional conveyance [12]. In natural speech, the pitch range exhibits wide variability, whereas in synthetic speech it tends toward near-monotony. Likewise, voice quality is often either harsh or overly flat, thereby diminishing perceived emotional naturalness. Consequently, differentiating natural from synthetic speech through emotion-related cues constitutes an effective approach [13].

The experimental analyses above indicate that several simple prosodic features—such as duration and pitch level—do not reliably discriminate all primary emotions [14]. Moreover, gender-related differences in prosody (between female and male speakers) are naturally present in speech. In addition, conventional methods for computing energy-based features are not well aligned with human auditory perception. The block diagram of the feature-extraction pipeline is shown in Figure 2.

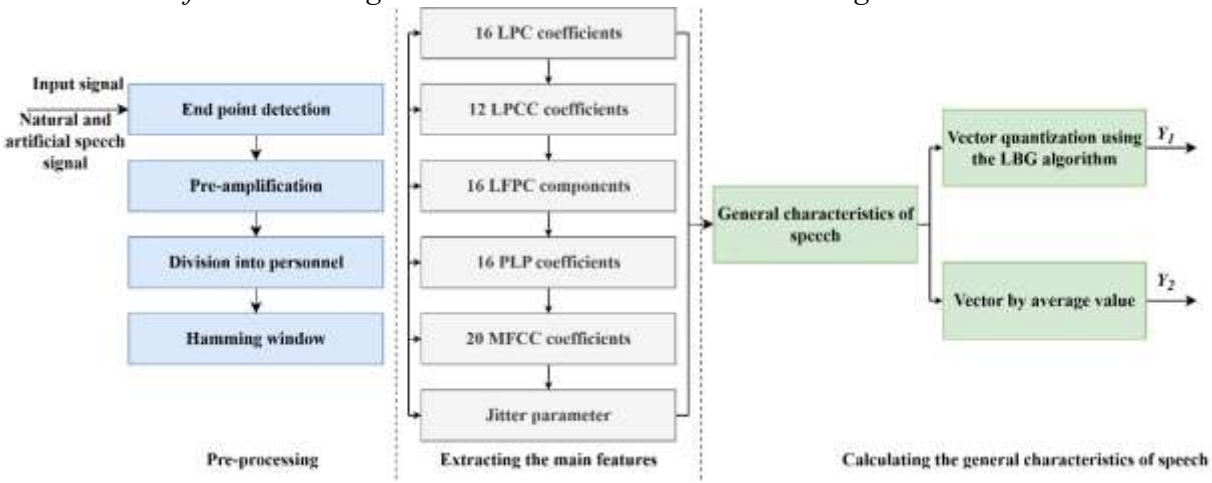


Figure 2. Speech Signal Processing Models

We employed a regression-based feature selection procedure to identify the most effective predictors from a pool of more than 200 speech features. Ten candidates were retained: LPC, LPCC, MFCC, Delta-MFCC, Delta-Delta-MFCC, PLP, RastaPLP, LFPC, jitter, and shimmer. Subsequently, a feature vector was computed for each segment of the sentences in Corpus I. At each stage, classification accuracy was evaluated using cross-validation with a k-NN classifier (k = 3).

Table 5. Accuracy by Individual Feature

Feature	Accuracy (%)
LPCC	68.68
MFCC	68.21
LPC	68.20
PLP	65.59
RastaPLP	65.23
D-MFCC	60.59
LFPC	58.42
Shimmer	53.05
D-D-MFCC	50.18
Jitter	34.77

Table 6. Accuracy for Two-Feature Combinations

Feature	Accuracy (%)
MFCC	68.97
D-MFCC	67.38
LPC	66.52
PLP	66.52
LFPC	66.52
RastaPLP	66.16
D-D-MFCC	60.06
Jitter	54.33
Shimmer	42.86

Table 7. Accuracy for Six-Feature Combinations

Feature	Accuracy (%)
LPCC	83.91
MFCC	83.91
LFPC	83.19
Jitter	83.19
PLP	83.19
Shimmer	79.40

Tables 5–7 report the accuracy levels for the initial set of ten candidate features. The highest accuracy, 83.91%, was obtained during forward selection. In this procedure, the observed accuracy may increase or decrease depending on the effectiveness of specific feature combinations. Tables 5–7 present the forward-selection outcomes when 1, 2, and 6 features are selected, respectively.

Based on these experimental results, the core feature set for emotion-based synthetic-speech detection comprises six parameters—LPCC, MFCC, LFPC, jitter, PLP, and LPC—which together constitute the most effective combination identified.

16 Linear Predictive Coding (LPC) coefficients;

12 Linear Predictive Cepstral Coefficients (LPCC);

16 Log-Frequency Power Coefficients (LFPC);

16 Perceptual Linear Prediction (PLP) coefficients;

20 Mel-Frequency Cepstral Coefficients (MFCC);

Jitter;

The resulting feature vector comprises 81 parameters in total. LPC provides a cost-effective method for modeling the short-term spectral envelope of speech. LPCC and MFCC are among the most effective for representing phonetic content and the distribution of energy across frequency bands. LFPC applies a logarithmic frequency filterbank, enabling the modeling of the human ear's varying sensitivity across frequencies. PLP combines the discrete Fourier transform with LPC to yield a compact and efficient representation. Jitter quantifies natural cycle-to-cycle variations in fundamental period and is often markedly present in synthetic speech.

These findings indicate that combining multiple acoustic and spectral features—MFCC, LPCC, LPC, PLP, and LFPC—together with jitter achieves high effectiveness in distinguishing synthetic speech from natural speech.

Vector Quantization of Feature Vectors. Each feature vector contains 81 parameters, which makes classification computationally intensive. To reduce data dimensionality and accelerate inference, we applied vector quantization (VQ). All frames of the speech signal are assigned to a cluster and encoded by the index of the best-matching codebook entry. We used the K-means algorithm and tested different numbers of cluster centroids. The results showed that performance degraded once the number of clusters exceeded 16; therefore, all experiments employed 16 LBG (Linde–Buzo–Gray) codebook centroids. Consequently, for a speech utterance comprising N frames, the feature vector is defined as follows:

$$Y_1 = [c_1^*, c_2^*, \dots, c_N^*] \quad (1)$$

here, N denotes the number of frames.; c_i *– i- best-matching frame.* In an alternative, simple quantization scheme, the mean of each parameter across all frames was computed, yielding the following vector:

$$Y_2 = [p_1, p_2, \dots, p_{81}] \quad (2)$$

here, p_i - the mean of the i-th parameter across all frames.

Classification process. Three classification methods were employed in the study: Linear Discriminant Analysis (LDA), k-Nearest Neighbors (k-NN), and Hidden Markov Models (HMMs). In the k-NN method, the three nearest neighbors to the test instance were used (k = 3). In the HMM setup, the state-transition and emission probabilities were initialized with equal (uniform) initial values.

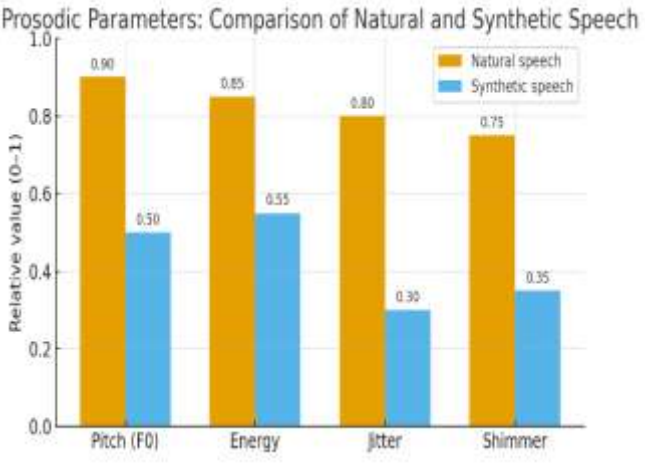


Figure 1.2. Comparison of Prosodic Parameters in Natural and Synthetic Speech

Table 8. Expanded Results for Emotions and Synthetic Speech Based on Conventional Prosody

Actual \ Predicted	**Anger**	**Boredom**	**Happiness**	**Neutral**	**Sadness**	**Synthetic**
Anger	54.81	1.01	29.85	4.05	2.39	7.89
Boredom	0.00	43.11	1.01	18.79	29.20	7.89
Happiness	29.85	2.30	54.07	3.87	2.03	7.89
Neutral	8.66	7.09	8.01	47.99	20.36	7.89
Sadness	1.57	27.08	2.21	16.21	45.04	7.89
Synthetic	1.00	0.60	2.00	4.00	1.20	91.2

The experimental results indicate that a four-state discrete ergodic HMM achieved superior performance for emotion-based synthetic-speech detection compared with alternative architectures. While prosodic features can differentiate emotions to a certain extent, they yield notable confusions for similarity pairs—e.g., anger ↔ happiness and boredom ↔ sadness—with an overall mean accuracy of approximately 53.2%. Accordingly, the prosodic baseline was integrated with a synthetic-speech detection objective: prosody was augmented with spectral–cepstral indicators (MFCC/LPCC/LPC/PLP), phonatory measures (jitter/shimmer), and codebook entropy. In parallel with emotion classification, a binary "Natural vs. Synthetic" detector was constructed to improve discrimination of artificial speech.

3. EXPERIMENTAL RESULTS ON EMOTION AND SYNTHETIC-SPEECH DETECTION

Alongside the pairs anger/happiness and boredom/sadness, the task also includes discriminating synthetic (TTS) speech from natural speech within the same pairs. Accordingly, each pair is treated as a three-class problem: Emotion-1, Emotion-2**,** and Synthetic speech. The selected acoustic features **(**LPC, LPCC, LFPC, PLP, MFCC, and jitter) were converted into feature vectors using two schemes: Y_1 — vector quantization based on LBG (k-means**)**; — a vector formed from the framewise mean of each parameter. For classification, LDA**,** k-NN, and HMM models were applied. The results show that, when synthetic-speech discrimination is added to the emotion-pair tasks, HMM (in particular the Y_2 variant) generally yields the most stable and highest performance**.** Tables 9 and 10 report the training-set and test-set accuracies, respectively.

Table 9. Anger/Happiness + Synthetic Speech (Y_1 and Y_2– Average)

Sinf	***LDA Y_1***	***LDA Y_2***	***K-NN Y_1***	***K-NN Y_2***	***HMMs Y_1***	***HMMs Y_2***
Anger	93.1	93.4	93.7	91.6	93.9	92.6
Happiness	87.7	91.2	90.4	92.8	91.2	93.5
Synthetic Speech	90.8	91.6	90.9	91.3	92.8	93.2
Average	90.5	92.1	91.7	91.9	92.6	93.1

Table 10. Boredom/Sadness + Synthetic Speech (Y_1—LBG; Y_2—Mean)

Sinf	***LDA Y_1***	***LDA Y_2***	***K-NN Y_1***	***K-NN Y_2***	***HMMs Y_1***	***HMMs Y_2***
Boredom	89.5	90.5	89.7	92.1	90.5	94.3
Sadness	92.2	87.6	93.5	90.4	93.2	90.9
Synthetic Speech	90.2	91.0	90.8	91.4	92.0	92.6
Average	90.6	89.7	91.3	91.3	91.9	92.6

Accurately assessing a speaker's emotional state is among the most challenging problems in speech-processing technologies. In this work, we extend the classical emotion-recognition task by additionally aiming to distinguish synthetic speech from natural speech within an emotional context. Using an acoustic feature set comprising 16 LPC, 12 LPCC, 16 LFPC, 16 PLP, 20 MFCC components, and jitter, features were mapped to vectors via two schemes: Y_1 (LBG-based vector quantization) and Y_2 (framewise averaging). For classification, we employed LDA, k-NN, and HMM.

In addition to 558 natural emotional sentences recorded by 12 native speakers and 503 sentences from two professional speakers, we also included synthetic speech samples for each emotion pair. This expanded the overall task to a six-class formulation: the five primary emotions—anger, boredom, happiness, neutral, and sadness—plus synthetic speech.

Table 11. Emotion + Synthetic Speech (Held-Out Set)

Class	**LDA Y_1**	**LDA Y_2**	**K-NN Y_1**	**K-NN Y_2**	**HMMs Y_1**	**HMMs Y_2**
Anger	81.5	80.4	82.3	84.8	86.4	86.7
Boredom	80.3	79.8	84.9	82.3	89.1	88.4
Happiness	76.5	72.3	79.5	82.1	82.3	83.6
Neutral	83.6	78.9	89.1	93.3	90.4	92.3
Sadness	82.5	81.3	91.2	89.1	92.4	92.3
Synthetic	90.4	91.1	91.0	91.8	92.3	93.1
Average (5 emotions)	79.8	78.8	83.6	83.9	86.9	88.3
Average (6 classes, with synthetic added)	81.6	80.9	84.8	85.2	87.8	89.1

Table 12. Results for 5 Emotions + Synthetic Speech

Sinf	**LDA Y_1**	**LDA Y_2**	**K-NN Y_1**	**K-NN Y_2**	**HMMs Y_1**	**HMMs Y_2**
Anger	82.4	76.2	83.2	84.5	90.2	91.4

Boredom	80.9	80.2	85.1	83.4	90.4	86.7
Happiness	81.4	77.8	86.4	82.5	87.5	88.1
Neutral	76.5	79.8	84.1	83.2	90.3	86.0
Sadness	80.3	76.5	86.0	87.5	89.5	91.5
Synthetic	90.1	90.8	90.7	91.5	92.1	93.0
Average (5 emotions)	79.9	78.1	84.2	83.7	88.3	88.7
Average (6 classes)	81.6	80.2	85.3	85.0	88.9	89.4

The selected acoustic feature set, together with the Y_2 –HMM configuration, demonstrated reliable discrimination not only of the five primary emotions but also of synthetic speech within an emotional context. Accordingly, the joint emotion-recognition and anti-spoofing approach constitutes an effective and robust solution suitable for integration into practical systems.

CONCLUSION

In classification over the emotion pairs—anger vs. happiness and boredom vs. sadness—and their association with synthetic speech, the HMM with the Y_2 representation consistently delivered the most stable results. The synthetic-speech class was successfully separated even from natural speech with similar emotional expressivity; HMM (Y_2) scores frequently fell in the ≈92–93% range. The average accuracies across emotion recognition and synthetic-speech detection using HMM (Y_2) were ≈89–89.5%, indicating that the selected feature set can robustly discriminate synthetic speech against an emotional background. The proposed model can operate in a joint setting that simultaneously performs emotion recognition and synthetic-speech (anti-spoofing) detection. Potential application areas include customer-support systems, voice assistants, interactive games, audiobook/TTS monitoring, educational and clinical/therapeutic interfaces, and security systems.

REFERENCES

[1] Ata, B. S., “Effectiveness of Linear Prediction Characteristics of the Speech Ware for Automatic Speaker Identification and Verification,” Journal of the Acoustical Society of America, 55 (1974), pp. 1304–1312.

[2] Cheng, P. Y., “Automated Recognition of Emotion in Mandarin,” MD thesis, National Cheng Kung University, 2002.

[3] Chuang, Z. J. and C. H. Wu, “Multi-Modal Emotion Recognition from Speech and Text,” International Journal of Computational Linguistics and Chinese Language Processing, 9(2), 2004, pp. 1–18.

[4] Cowie, R., E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J. G. Taylor, “Emotion Recognition in Human-Computer Interaction,” IEEE Signal Processing Magazine, 18(1), 2001, pp. 32–80.

[5] Davis, S. and P. Mermelstein, “Comparison of Parametric Representations of Monosyllabic Word Recognition in Continuously Spoken Sentences,” IEEE Transactions on Acoustics, 28(4), 1980, pp. 357–366.

[6] Ekman, P., Handbook of Cognition and Emotion, John Wiley & Sons, New York, 1999.

[7] Hermansky, H., “Perceptual Linear Predictive (PLP) Analysis of Speech,”

Journal of the Acoustical Society of America, 87(4), 1990, pp. 1738–1752.

[8] Holzapfel, H., C. Faion, M. Dennece and A. Waibel, “Integrating Emotional Cues into a Framework for Dialogue Management,” in Proceedings of International Conference on Multimodal Interfaces, 2002, Pennsylvania, USA, pp. 141–148.

[9] Iznanoglu, Z. and R. Canel, “Emotive Alert: HMM-Based Emotion Detection in Voicemail Messages,” in Proceedings of Intelligent User Interfaces, 2005, San Diego, USA, pp. 251–258.

[10] Kaiser, J. F., Discrete-Time Speech Signal Processing, Prentice Hall, New Jersey, 2002.

[11] Kleinginna, P. R. and A. M. Kleinginna, “A Categorized List of Emotion Definitions with Suggestions for a Consensual Definition,” Motivation and Emotion, 5(4), 1981, pp. 345–379.

[12] Kwon, O. W., K. Chan, J. Hao and T. W. Lee, “Emotion Recognition by Speech Signals,” in Proceedings of Eurospeech, 2003, Geneva, Switzerland, pp. 125–128.

[13] Le, X. H., G. Ouenot and P. de Lachel, “Recognizing Emotions from the Audio-Visual Document Using Speech Prosody,” in Proceedings of the Ninth IEEE International Symposium on Computers and Communications, 2004, Alexandria, Egypt, pp. 580–584.

[14] Linde, Y., A. Buzo and R. M. Gray, “An Algorithm for Vector Quantizer Design,” IEEE Transactions on Communications, 28(1), 1980, pp. 84–95.