ANALYZING AUDIO FEATURES TO DISTINGUISH HUMAN VOICE FROM AI

Muxriddin Abduganiev
Artificial intellegence
Tashkent university of information technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan
mr.muhriddin.20@gmail.com
Mamura Uzakova
Artificial intellegence
Tashkent university of information technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan
marynmm262321@gmail.com
Aygul Burxanova
Artificial intellegence
Tashkent university of information technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan

Abstract. Telling a real human voice apart from AI-generated speech is becoming incredibly important for things like security, forensics, and everyday tech. In this study, a more detailed analysis was performed by studying three acoustic features: Mel-Frequency Cepstral Coefficients (MFCC), Zero Crossing Rate (ZCR), and Spectral Centroid. The audio was preprocessed in Python using the librosa library, with the silent parts of the audio removed. The results were quite clear: the spectral centroid of real human speech is higher, and the variation of the MFCCs is much greater. Interestingly the ZCR values did not differ greatly between the two. In conclusion, the results presented here demonstrate that simple audio parameters can be effectively used for automatic detection of synthetic voices.

Keywords: speech synthesis, acoustic features, MFCC, zero crossing rate, spectral centroid, librosa, human-computer interaction, voice forensics

Аннотация. Различение настоящего человеческого голоса и речи, сгенерированной искусственным интеллектом, становится чрезвычайно важным для обеспечения безопасности, судебной экспертизы и повседневных технологий. В данном исследовании был проведён более детальный анализ трёх акустических признаков: мел-частотных кепстральных коэффициентов (MFCC), частоты пересечения нуля (ZCR) и спектрального центроида. Аудиоданные были предварительно обработаны в Python с использованием библиотеки librosa, при этом из записи были удалены участки тишины. Результаты оказались достаточно очевидными: спектральный центроид человеческой речи выше, а вариативность MFCC значительно больше. Интересно, что значения ZCR существенно не различались между двумя типами речи. В заключение

результаты исследования показывают, что простые акустические параметры могут эффективно использоваться для автоматического обнаружения синтетических голосов.

Ключевые слова: синтез речи, акустические признаки, MFCC, частота пересечения нуля, спектральный центроид, librosa, взаимодействие человека и компьютера, судебная экспертиза голоса.

Annotatsiya. Haqiqiy inson ovozini sun’iy intellekt tomonidan yaratilgan nutqdan ajratish xavfsizlik, sud ekspertizasi va kundalik texnologiyalar uchun tobora muhim ahamiyat kasb etmoqda. Ushbu tadqiqotda uchta akustik xususiyat: Mel-Frequency Cepstral Coefficients (MFCC), Zero Crossing Rate (ZCR) va Spectral Centroid chuqurroq tahlil qilindi. Audio ma’lumotlar Python dasturida librosa kutubxonasi yordamida oldindan qayta ishlanib, audio yozuvdagi jim qismlar olib tashlandi. Natijalar aniq bo‘ldi: haqiqiy inson nutqining spektral centroid qiymati yuqoriroq, MFCC koeffitsiyentlarining o‘zgaruvchanligi esa ancha katta ekanligi kuzatildi. Qiziqarli jihati shundaki, ZCR qiymatlarida ikki turdagi nutq o‘rtasida katta farq aniqlanmadi. Xulosa qilib aytganda, ushbu tadqiqot natijalari oddiy akustik parametrlar sintetik ovozlarni avtomatik aniqlashda samarali qo‘llanilishi mumkinligini ko‘rsatadi.

Kalit so‘zlar: nutq sintezi, akustik xususiyatlar, MFCC, nol kesishish chastotasi (ZCR), spektral centroid, librosa, inson-kompyuter o‘zaro aloqasi, ovoz sud ekspertizasi.

INTRODUCTION

The quality of the synthetic voices is now very realistic with the use of text to speech (TTS) technology in recent years. The modern systems such as WaveNet or Tacotron produce sound that is difficult to distinguish from that of human beings only by listening. But despite all these human tricks, the sound signals have some differences between them that we can discover by applying math and code. It's important to be able to see these differences for a number of reasons. First, it can make life easier for security as it helps in identifying “audio deepfakes” and fending off scams. Secondly, it enables researchers to compare the quality of computer-generated speech to natural human speech. Lastly, the analysis of this type can be used in forensics to determine whether a recording is genuine or tampered with. The objective of this research is to investigate whether simple acoustic features, namely MFCC, ZCR and Spectral Centroid, can be used to distinguish human speech from synthetic audio signals with good accuracy. We believe that natural speech will exhibit a lot more variation and fluctuation with time. This is because natural voices are created by physical actions that computers have yet to be able to accurately imitate.

RELATED WORKS

In the beginning, most research on speech classification was about identifying who is speaking or guessing their emotions. The big leap came with the development of Mel-Frequency Cepstral Coefficients (MFCCs) by Davis and Mermelstein [1]. This method is now the standard method of representing speech signals and many studies confirm the usefulness of MFCCs for different language recognition, speech style recognition, even medical condition recognition.

With the increasing prevalence of "audio deepfakes" and voice spoofing, the identification of synthesized speech emerged as a key research area. The competitions, such as ASVspoof challenge (Kinnunen et al. [2]) and Todisco et al. [3] had contributed a great deal with standard datasets and rules to test the ability to detect computer-generated voices. Most of the systems which are being used in the present day employ a combination of various sound features and the MFCCs still give very good performance in these tests.

Some other simple features, such as Zero Crossing Rate (ZCR) and Spectral Centroid are also commonly combined with MFCCs. ZCR can be used to estimate the noisiness of a sound, and the Spectral Centroid can be used to estimate the "brightness" or high frequency energy of the voice. These can be mixed together and give a much more accurate representation of the difference between a natural human voice and a voice synthesized by a machine.

Boyer-Moore [2] and KMP [3] are two classic singlestring matching algorithms. Both of these algorithms also use a window of size n, but they use a skip or shift table to determine where to look next after each mismatch.

METHODS (OPTIONAL).

Two audio samples were analyzed: (1) a natural human voice recording (Human_voice.wav) and (2) a synthetic speech sample generated by a TTS system (synthetic_speech.wav). Both recordings were in WAV format. The natural voice sample had a duration of approximately 110 seconds; the synthetic sample was approximately 70 seconds in length. Both samples were recorded or generated in English and contained continuous speech rather than isolated words. For this study, we analyzed two different audio files in WAV format [4-6]. The first was a 110-second recording of a natural human voice, and the second was a 70-second sample created by a text-to-speech system. Both recordings were in English and consisted of full sentences rather than just random words. We started by loading the audio into Python using the librosa library, making sure to keep the original sound quality. A very important step in our preparation was cleaning up the files by removing any

silence from the beginning and end. We used a 20 dB threshold for this because we wanted to make sure that quiet moments wouldn't change our final data or lead to wrong conclusions.

After the audio was ready, we focused on pulling out three specific types of information. First, we calculated 13 MFCC coefficients to see the overall shape of the sound. Next, we looked at the Zero Crossing Rate to check how often the signal changed, which helps us understand the "noisiness" of the speech. Finally, we measured the Spectral Centroid to see the "brightness" of the voices. For each of these features, we calculated the average values and how much they varied throughout the recordings.

To make the differences easier to see, we put everything together into a single 3x2 grid of charts. Using matplotlib and librosa, we plotted the waveforms, spectrograms, and MFCC maps for both samples. This visual comparison helped us identify exactly where the human voice and the computer-generated voice look different on a technical level[7-9].

RESULTS

Once we ran the tests and processed the audio, we collected the audio's visual and numerical data to compare the two voices. First, we examined the pattern of visual shapes in the sound waves and frequencies to gain an understanding of the results.

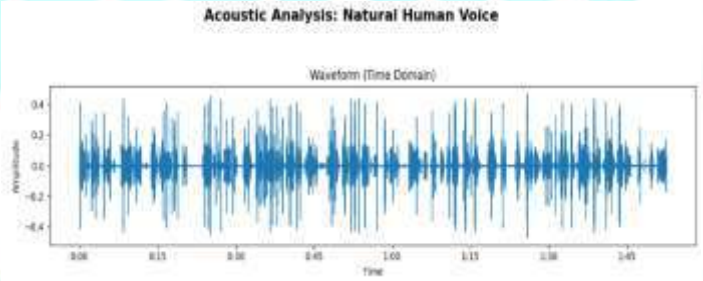

Figure 1. Waveform (Time Domain) of natural human speech

As shown in Figure 1, high amplitude variation is the characteristic of the waveform of the natural human voice. Peaks and valleys represent the natural dynamics of expressing speech in the sense of emphasis on words and pauses in breath. The complicated and non-linear nature of the signal suggests the physical mechanisms underlying the production of sound by the human vocal tract.

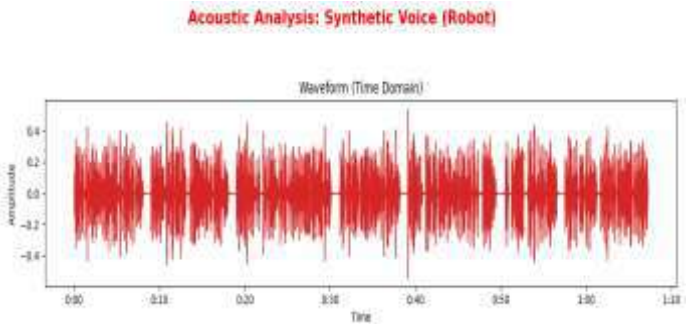

Figure 2. Waveform (Time Domain) of synthetic speech (Robot)

In contrast, Figure 2 shows that the synthetic waveform is much flatter and uniform compared to the human sample. The signal's density is fairly consistent, and it doesn't exhibit sudden dynamic changes, as is typical for computer-generated sounds. This is the simplified articulation pattern used by the TTS vocoder.

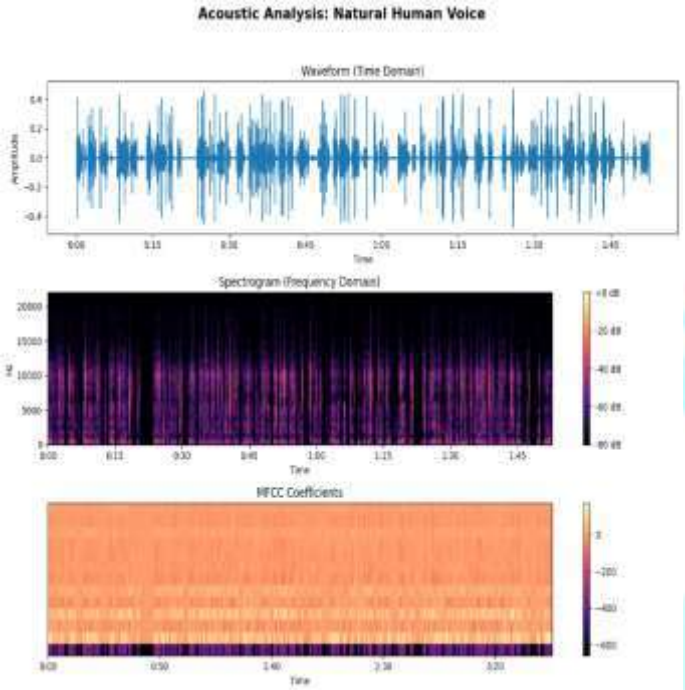

Figure 3. Acoustic Analysis: Natural Human Voice. From top to bottom: Waveform, Spectrogram, and MFCC Coefficients.

Figure 3 provides a comprehensive view of the natural voice. The spectrogram shows a very rich harmonic structure, especially at higher frequencies above 10,000 Hz, where the energy is contributed by the human vocal tract during resonance. Moreover, the MFCC matrix in Figure 3 is characterized by high temporal variability due to the dynamic "shape" of natural articulation.

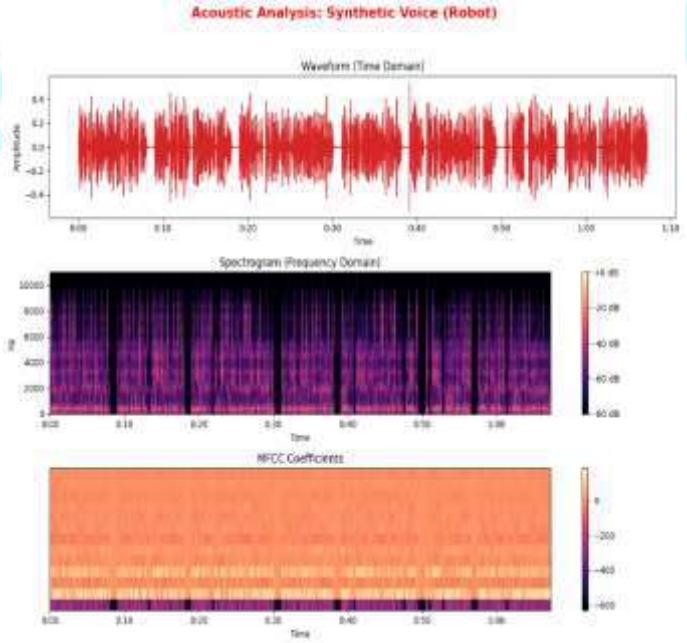

Figure 4. Acoustic Analysis: Synthetic Voice (Robot). From top to bottom: Waveform, Spectrogram, and MFCC Coefficients.

As illustrated in Figure 4, the synthetic waveform is much flatter and smoother compared to the human waveform. The synthetic voice's spectrogram contains a very steep cut-off point for frequency, indicating that the vocoder has ceased producing high-frequency sound. The MFCC matrix of the robot voice is very consistent over time, showing a decrease of complexity in the vocal texture.

To back up what we saw in the pictures, we also calculated specific numbers for each feature. These are summarized in Table 1 below.

The perceptual brightness of the sound is determined by the "center of mass" of the sound spectrum (called the Spectral Centroid) and is calculated by the following equation (1):

$$C = \frac{\sum_{k=1}^{N} f(k)|X(k)|}{\sum_{k=1}^{N} |X(k)|}$$

where:

C - the calculated Spectral Centroid.

N - the total number of frequency bins in the analyzed spectrum.

k - the index of the current frequency bin.

f(k) - the center frequency of bin k in Hertz.

|X(k)| - the magnitude (or amplitude) of the frequency bin k.

Also, the Zero Crossing Rate (ZCR), reflecting the sign-changing speed of the signal and measuring the noisiness of speech (e.g., fricatives) is defined as (2):

$$ZCR = \frac{1}{2T} \sum_{t=1}^{T} |sgn(s_t) - sgn(s_{t-1})| \quad (2)$$

where:

ZCR - the calculated Zero Crossing Rate.

T - the total number of audio samples in the evaluated time frame.

t - is the discrete time index.

s_t - the signal amplitude at the current time sample t.

s_{t-1} - the signal amplitude at the previous time sample t-1.

sgn() - the signum function, which yields 1 for positive values, -1 for negative values, and 0 for zero.

Finally, the complexity of articulation along time and its dynamic shape was assessed by computing standard deviation of Mel-Frequency Cepstral Coefficients (MFCCs) (3):

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \quad (3)$$

where:

σ - is the standard deviation, representing the variance or dynamic complexity of the features.

N- is the total number of MFCC observations (frames) analyzed.

i - is the index of the current observation.

X_i - is the value of the MFCC feature at index i.

μ - is the mean (average) value of all the analyzed MFCCs in the sample.

These extracted metrics are summarized in table 1 below.

table-1

Summary of acoustic feature statistics for natural and synthetic speech

Metric	Natural Voice	Synthetic Voice
MFCC Mean	–28.81	–20.12
MFCC Std Dev (Complexity)	135.07	100.83
ZCR Mean (Noisiness)	0.1345	0.1286
ZCR Std Dev	0.1104	0.1252
Spectral Centroid Mean (Hz)	5252.42	2073.05
Spectral Centroid Std Dev	2276.67	1447.65

The numbers show that the MFCC standard deviation is much higher for the natural voice (135.07 compared to 100.83). This confirms that a real human voice changes its "shape" more often than a machine does. Interestingly, the Zero Crossing Rate (ZCR) was very similar for both, which means the synthetic system is actually quite good at handling noisy sounds like "s" or "f."

But the most significant change was that of the Spectral Centroid. The mean frequency of the human voice was 5252.42 Hz and for the synthetic voice was 2073.05 Hz. This shows that with the natural voice there is a much greater "brightness" and higher energy in the higher frequencies. The team also noticed that the standard deviation of the human speech is higher, indicating that its volume fluctuates more frequently compared to the robot's speech.

Discussion

Our results consistently show that natural speech is much more varied than synthetic speech, especially when we look at the spectral energy and MFCC complexity. This makes sense because human speech is produced by physical movements of the throat and mouth, which are controlled by the nervous system. This creates natural changes in rhythm, emotion, and tiny unique details that current text-to-speech (TTS) systems try to copy but cannot perfectly replicate.

The most striking difference we found was in the Spectral Centroid. The natural voice was about 3000 Hz "brighter" than the robot voice (5252 Hz vs 2073 Hz). Most modern TTS systems use something called a vocoder, which often cuts off the sound at higher frequencies (above 8,000 Hz). In contrast, a real human voice has a lot of energy in those high ranges because of natural breathing and the way we pronounce certain sounds.

Interestingly, we thought the Zero Crossing Rate (ZCR) would be much higher for the human voice, but the numbers were actually very close. This is telling us that the TTS system that we evaluated is pretty good at simulating noisy sounds such as "s" or "f. This is particularly significant because the results indicate that the Spectral Centroid approach could be much more viable for detection of fake voices than ZCR.

Of course, this study has some limits. We only compared one person’s voice with one specific TTS system. In the future, we should test more voices and different types of AI systems to see if these patterns stay the same. However, these results already suggest that even simple math features can be a very powerful tool for building systems that automatically detect "audio deepfakes."

Conclusion

In this paper, we compared natural and synthetic English speech using three main features: MFCC, ZCR, and Spectral Centroid. As we expected, the human voice showed much more variety. The biggest difference was in the Spectral Centroid, where the human voice had a mean value more than double that of the synthetic voice.

These results are of great relevance in areas such as audio forensics and digital security. They show it is not always necessary to have super-complex algorithms to identify "digital fingerprints" in audio. Using basic acoustic characteristics, we can make fast and efficient tools that could be used to verify whether a voice is real or generated by a machine. It's a modest but substantial progress towards a more secure digital communication.

REFERENCES

[1] Maxkamov B.Sh., Zaynidinov X.N., Nurmurodov J.N. “Sun’iy intellekt asoslari” / Toshkent axborot texnologiyalari universiteti. – Toshkent: Javohir ilm

nashr, 2024. – 264 b.

[2] Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), 357–366.

[3] Kinnunen, T., Sahidullah, M., Delgado, H., Todisco, M., Evans, N., Yamagishi, J., & Lee, K. A. (2017). The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. *Proceedings of Interspeech 2017*, 2282–2286.

[4] Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., ... & Evans, N. (2019). ASVspoof 2019: Future horizons in spoofed and fake audio detection. *Proceedings of Interspeech 2019*, 1008–1012.

[5] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in Python. *Proceedings of the 14th Python in Science Conference*, 18–25.

[6] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

[7] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... & Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. *Proceedings of Interspeech 2017*, 4006–4010.

[8] Kim, J., Kong, J., & Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 5530–5540.

[9] Zakariah, M., Khan, S. A., & Al-Khouri, A. M. (2022). Digital audio forensics: Review, methods, and challenges. *IEEE Access*, 10, 11200–11225.