I645 Human Perceptual Systems and its Models

Report 2

Discuss how will you apply (or use) your knowledge with regard to auditory perception to (or to solve) engineering problems.

Name： Cheng Haowei

Student ID: 2110415

Introduction:

Deep-fake speech has become a growing concern in recent years, as it has the potential to manipulate and spread false information. As a result, the need for reliable and accurate deep-fake speech detection methods have become increasingly important. One approach to detecting deep-fake speech is using auditory perception. To do this, it is important to understand the properties of speech signals and how they are perceived by the auditory system [1].

Auditory model:

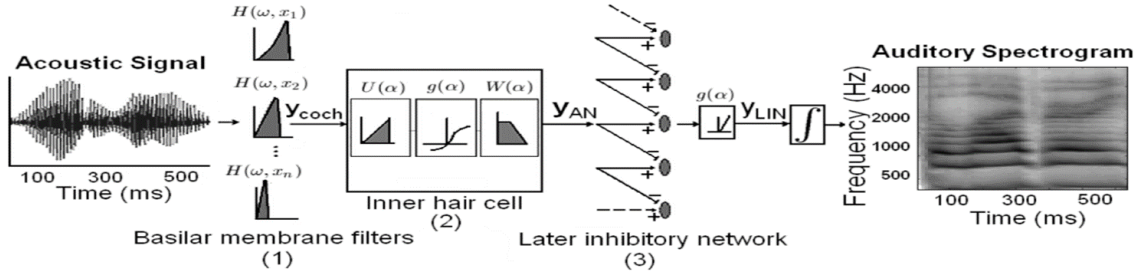

Figure 1 Schematic of the early stages of auditory processing

(1)

$$y_{\text{coch}}(t, f) = s(t) * h(t; f)$$

This period happens in the cochlea (shown in Figure 2). The convolution of the input signal "s(t)" and the impulse response of the cochlea "h (t;f)" gives the cochlea filter output "y (t,f)" in the time-frequency domain. The cochlea output represents the energy distribution of the input signal over time and frequency, providing information on how the input signal is transformed by the cochlea.

In other words, the convolution operation models the effect of the cochlea on the input signal, providing information on how the input sound is transformed into a different representation, suitable for further processing in the auditory system.

Figure 3 shows the spectrogram of real and fake speech after the gammatone filter bank, which mimics the human cochlea.

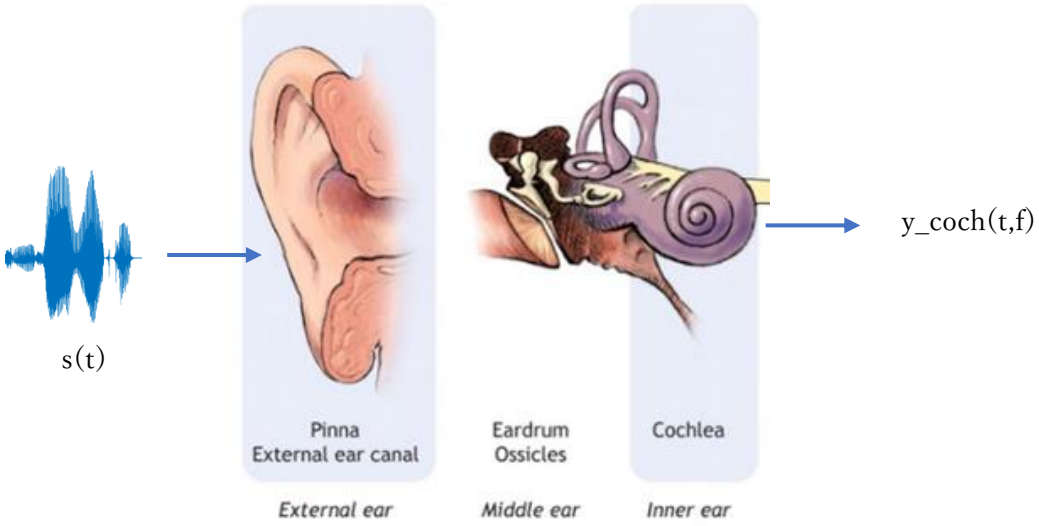

Figure 2 Auditory system structure from the external ear to the inner ear

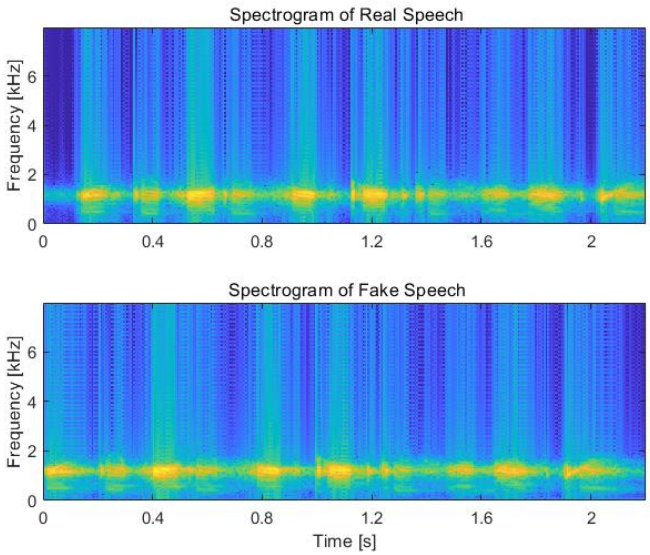

Figure 3 spectrogram of real and fake speech

(2)

$$y_{AN}(t,f) = g\left(\partial_t y_{coch}(t,f)\right) * w(t)$$

This process happens in the auditory nerve (shown in Figure 4).

$\partial_t y_{coch}(t,x)$ is the cochlea response in the time-frequency domain

$g(\partial\_t\ y\_coch\ (t,x))$ is the envelope of the cochlea response.

$w(t)$is the center frequency of the auditory filter.

The equation represents the auditory nerve (AN) response in the time-frequency domain. The

envelope of the cochlea response describes the changes in the amplitude of the cochlea response over time for a specific frequency. The center frequency of the auditory filter describes how the filter responds to different frequencies over time. The convolution of these two signals provides information on how the cochlea response is transformed by the auditory nerve, which simulates the transfer of information from the cochlea to the auditory nerve [2], then used for further processing in the auditory system.

Figure 5 shows the envelope of real and fake speech, in particular, the envelope of a speech signal contains information about the pitch, loudness and rhythmic structure. This information is processed and integrated by the auditory system to form a perceptual representation of the speech signal, which is used for speech perception, recognition, and understanding.

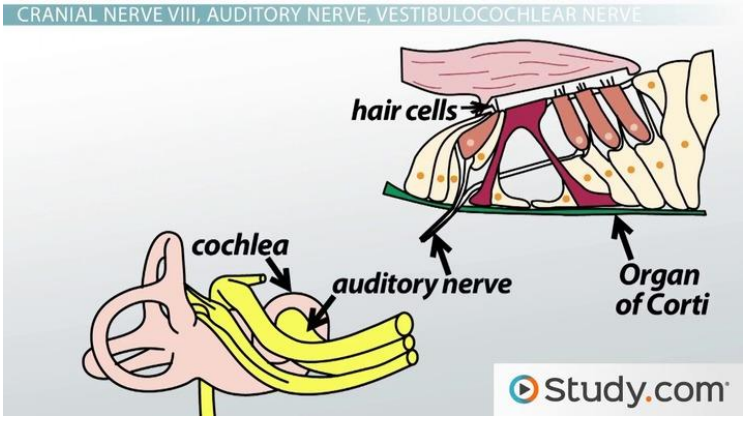

Figure 4 Auditory system structure of auditory nerve

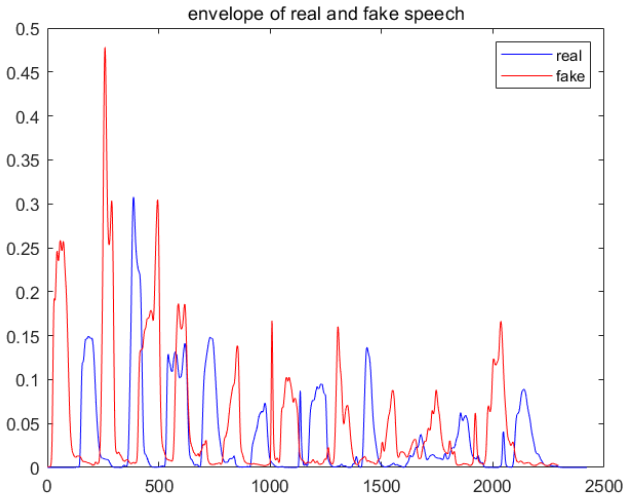

Figure 5 Envelope of real and fake speech

(3)

$$y_{LIN}(t,f) = max(\partial_x y_{AN}(t,f), 0)$$

$\partial_x y_{AN}(t,x)$is the frequency representation of the auditory nerve response, which describes the

changes in the amplitude of the auditory nerve response over time for a specific frequency.
It is a mathematical model of a stage in the auditory pathway that occurs after the cochlea, in the auditory nerve (shown in Figure 3), or the auditory brainstem. The purpose of this stage is to perform rectification, only the positive values are kept.

(4)

$$y(t,f) = y_{LIN}(t,f) * \mu(t;\tau)$$

$y_{LIN}(t,x)$ is the linearity of the auditory nerve response in the time-frequency domain.
$\mu(t;\tau)$ is the midbrain filter, which describes the response of the midbrain to different frequencies over time.
This process is simulating the processing that occurs in the midbrain. To capture the properties of the auditory processing that occurs in the midbrain, such as temporal integration and auditory filtering.
In this stage, we can get the characterization information (shown in Figure 6) of the feature. It can be used to distinguish fake speech from real speech.

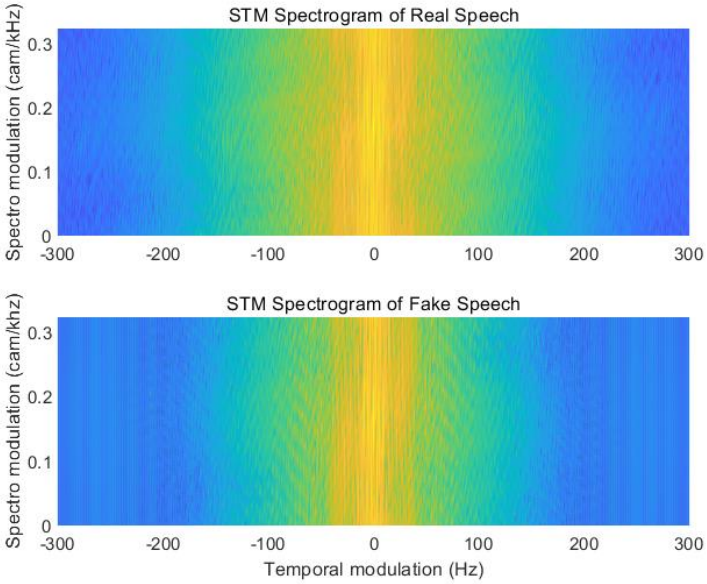

Figure 6 Spectro-temporal modulation (STM) spectrogram of real and fake speech

From the perspective of speech perception to analyze speech signals, the differences between genuine speech and fake speech can be emphasized and revealed in several ways. Here are a few examples:
1. Temporal variations: Genuine speech has natural variations in timing and duration, such as pauses, emphasis, and stress, that are difficult to replicate perfectly in synthetic speech. The feature based on auditory perception can extract features that reflect these temporal variations,

such as the rate of change of the speech signal over time, and these features can be used to distinguish between genuine and fake speech [3].

2. Spectral variations: Another main difference between genuine and fake speech is the spectral structure of the speech signal. Genuine speech has natural variations in pitch, harmonics, and formants that are also difficult to replicate perfectly in synthetic speech [4]. The feature based on auditory perception can extract features that reflect these spectral variations, such as the distribution of energy across different frequency bands and the modulation of energy across time.

3. Prosodic variations: Synthetic speech usually lacks the natural variation in prosodic features, such as speaking rate, intonation, stress, and phrasing. These variations can be used to distinguish between genuine and fake speech.

4. Non-linearity and non-stationarity variations: Real speech signals are non-linear and non-stationary, and this property is quite difficult to replicate in synthetic speech. The auditory system processes speech signals in a nonlinear manner, which means that different aspects of the speech signal can be processed differently based on the frequency and intensity of the signal [5]. For example, the basilar membrane, a key component of the inner ear, responds differently to different frequencies and amplitudes of sound, allowing us to perceive the spectral and temporal characteristics of speech.

Disadvantages:

One of the main limitations is that auditory perception is a subjective process that depends on the listener's perception and experience. This can result in variability in the ability to detect deep-fake speech, especially when the deep-fake speech is designed to mimic real speech very closely. Additionally, auditory perception is not always sufficient to detect deep-fake speech, especially when the deep-fake speech is created using advanced techniques that can manipulate both the acoustic and prosodic aspects of speech. Another disadvantage is that auditory perception is not always reliable for detecting deep-fake speech in noisy or reverberant environments. The quality of speech signals in these environments can degrade and make it difficult to distinguish between genuine and fake speech. Therefore, it is important to use multiple techniques and approaches to detect deep-fake speech, rather than relying solely on auditory perception.

Conclusion:

Overall, the use of auditory perception in deep-fake speech detection is a promising approach to ensuring the accuracy and reliability of speech verification systems. Through the analysis of specific features, listeners can identify differences between genuine speech and fake speech, helping to ensure the credibility of speech signals. While the deep-fake speech detection problem is complex, the application of auditory perception provides valuable insights into the

acoustic properties of speech and can be used to improve deep-fake speech detection algorithms.

Reference

[1] https://www.veritonevoice.com/blog/everything-you-need-to-know-about-deepfake-voice

[2] Joris, P.X., 2003. Interaural time sensitivity dominated by cochlea-induced envelope patterns. Journal of Neuroscience, 23(15), pp.6345-6350.

[3] Wu, Z., Chng, E.S. and Li, H., 2012. Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition. In Thirteenth Annual Conference of the International Speech Communication Association.

[4] Wu, Z., De Leon, P.L., Demiroglu, C., Khodabakhsh, A., King, S., Ling, Z.H., Saito, D., Stewart, B., Toda, T., Wester, M. and Yamagishi, J., 2016. Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(4), pp.768-783.

[5] Banbrook, M., 1996. Nonlinear analysis of speech from a synthesis perspective.