1 Lab Book S4

In this lab book you will work with the **WORLD** vocoder:
Morise, M., Yokomori, F., & Ozawa, K. (2016). WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. IEICE Trans. Inf. Syst., 99-D, 1877-1884.

Installing and using Python-WORLD

Installation for python WORLD vocoder can be found on Brightspace page (Week 5. Vocoders / WORLD vocoder info).

1.1 Assignment 1

[10pts] *Vocoding different speakers and speaking styles*:

Intro:

In this task you will vocode different different speakers and speaking styles.

Task:

Look for recordings of different speakers and speaking styles[a]: make sure to include male and female voices, young and old voices. Try to include diverse speaking styles, try including whispered speech and creaky voice. Please, limit the number of recordings to 5-10.

1. Using WORLD, vocode your diverse samples of speech.
2. Describe the results: what is the most natural vocoded sample? If some samples did not yield a satisfying result, in your opinion, what went wrong with the speech reconstruction?

Submission

Alongside with this LaTeX file with your reflections, commit the code you used for vocoding speech and the set of voice samples, original and vocoded.

[a] If your classmates agree, you can reuse parts of the recordings of North Wind and the Sun spoken by different people in different languages

Answer

In the assessment of vocoded speech samples, I employed a dual approach that blends subjective auditory evaluation with objective visual analysis. Utilizing the WORLD vocoder, I synthesized two variants of speech signals: one that was a direct analysis and synthesis of the original, and another that underwent pitch scaling by a factor of 1.5. The resulting samples spanned various speech types and were subjectively appraised for their naturalness on a scale from 1 to 10. This comprehensive method ensures a balanced evaluation, capturing both the perceptual nuances and the tangible modifications induced by the vocoding process.
Note: To accommodate the 20 visual plots for ease of review, I have compiled them into a single zip file rather than attaching them individually here.
Breathy Speech:
The breathy speech vocoded sample exhibits discernible electronic artifacts, leading to a subjective naturalness rating of 6 out of 10. Visually, the vocoded waveform lacks the amplitude variations present in the original, essential for conveying the characteristic breathiness. This discrepancy is likely due to the vocoder's limitations in modeling the aperiodic components intrinsic to breathy voices, a crucial aspect that contributes to their distinct quality.

Childish Speech:
The vocoded childish speech is perceived as shriller and more piercing, yet retains a reasonable degree of naturalness, with a rating of 7 out of 10. The relatively preserved spectral characteristics in the visual analysis suggest the high-pitched nature of the source is less affected by the vocoding process. However, the increased high-frequency emphasis aligns with the characteristics of phase vocoders highlighted in Kawahara's paper.

Creaky Speech:
The creaky speech vocoded sample captures low-frequency elements effectively, warranting a naturalness score of 8 out of 10. The visual comparison reveals a close match in fundamental frequencies, indicative of the WORLD vocoder's proficiency in modeling the glottal source waveform and periodicity, as suggested by Airaksinen.

Elderly Speech:
Elderly speech vocoding results in a sample reminiscent of an old-time phonograph, rated 8 out of 10 in naturalness. The visual analysis suggests a loss of detail in higher frequencies, possibly because elderly voices often exhibit micro-variations that the vocoder's deterministic plus stochastic modeling may not fully capture.

Female Speech:
The vocoded female speech sample is marked by weaker high frequencies and instability, leading to a naturalness rating of 7 out of 10. The visual data indicates a reduced spectral presence in higher frequencies, implying that the vocoder's mixed excitation model may not be as effective in replicating the rapid spectral fluctuations characteristic of female speech.

Male Speech with BGM:
The presence of BGM severely affects the vocoded male speech sample, which exhibits a significant reduction in amplitude and a piercing quality, resulting in a naturalness score of 5 out of 10. This aligns with the visual analysis, which shows a diminished waveform and spectral detail, suggesting that the vocoder's monophonic design is ill-suited for complex signals with background music.

Outdoor Female Speech:
Rated 5 out of 10, the outdoor female speech sample suffers from distortion, with auditory elements akin to wind and electrical currents absent in the original. The visual discrepancies in the spectrogram suggest that the vocoder's stochastic component fails to accurately capture the environmental noise characteristics.

Singing Speech:
The singing speech with accompaniment is vocoded with a robotic quality, receiving a naturalness score of 6 out of 10. The visual analysis reflects a loss in dynamic range and harmonic content, indicating that the WORLD vocoder's clear speech synthesis strengths may not extend to the complex harmonic structures of singing with music.

Whispered Speech:
Remarkably, the whispered speech vocoded sample is nearly flawless, with a naturalness rating of 9 out of 10. This high fidelity is visually supported by a close approximation to the original in both spectrum and spectrogram, suggesting that the absence of vocal fold vibration significantly simplifies the modeling task for the vocoder.

Young Daily Talking Speech:
The young daily talking speech vocoded sample experiences a decline in naturalness to 7 out of 10, with an altered timbre and the presence of electronic noise. The visual analysis indicates that the original's intricate details are not fully preserved, especially in the latter half of the recording, consistent with the challenges vocoders face with rapid spectral transitions.

Summary
The **whispered speech** emerges as the most natural vocoded sample, closely mirroring the original, while the **male with BGM and outdoor speech** samples exhibit the most significant departures from naturalness. The variable performance of the vocoder across different speech types underscores the influence of non-stationary noises, high-frequency components, and complex acoustic environments on the reconstruction process.