ECE6255A Digital Speech Processing – Term Project Challenges

1. Speech Segmentation

Speech segmentation is the process of identifying the temporal boundaries between words, syllables, phonemes or speech sounds that have differentiable properties in a spoken utterance. Since a speech signal encompasses sounds of varying properties to allow embedding of linguistic information, speech segmentation is often a critically important task to accomplish as speech processing may involve "signal-dependent" techniques.

Signal segmentation obviously is built upon detection of temporal change in physical properties of the signal, but the grouping can be further prescribed according to need. For example, a sequence of word segmentation points must be a subset of the sequence of phonemic segmentation, and the corresponding phonemic orthography must be "sensibly" consistent with the given lexicon (e.g., missing phonemes in real utterances occur very often and must be accommodated when the detected phonemes are grouped into word). Another example of segmentation is based on the voicing and articulation status, i.e., to segment a speech signal into voiced, unvoiced, and silent regions. Therefore, the segmentation result depends on the prescribed definition of the segmentation unit, although the underlying principle of detecting temporal change in physical properties can be considered as a common foundation. The list of key physical properties includes the power spectrum, the voicing state, the fundamental frequency, the power level, etc.

This project aims at developing a segmentation algorithm, starting at the foundational level of detecting the temporal change in physical properties. Higher level segmentations, phonemic, syllabic, and beyond are also expected, time permits.

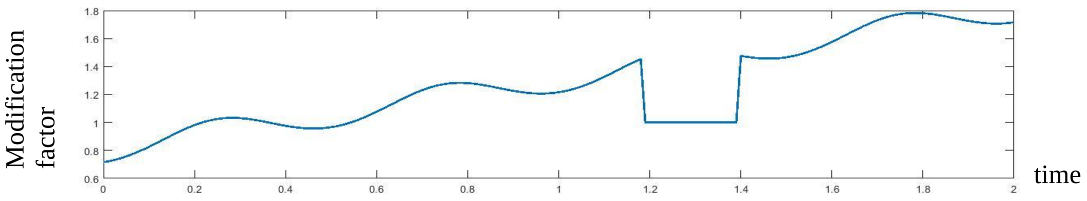
2. Arbitrary Modification of Speech Characteristics in Fundamental Frequency

The code developed for this challenge allows arbitrary modification of the fundamental frequency for any portion of a given speech signal. The modification in fundamental frequency should not result in major alteration of the speech contents and characteristics other than the f_0. By arbitrary modification it is meant that one can change the pitch of any region of the signal by specifying the starting and ending time for modification and the target fundamental frequency, which can be either a fixed value of f_0 or a scaling factor of the original fundamental frequency. Multiple modifications executed in a batch form are to be permitted. As an example, the input that specifies a single modification may include `t1, t2, f0` as arguments, where `t1, t2` are the starting and ending time (either in time or in sample index) of the interval for modification, and `f0` specifies either the target fundamental frequency, when it's > 50Hz, or a scaling factor, when < 10.

3. Arbitrary Modification of Speech Characteristics in Segmental Durations

This is similar to 2 above but targeting the duration aspect of the signal. Read 2 above first. The code developed for this challenge allows arbitrary modification of the duration of any portion of a given speech signal without changing the essential properties (e.g., pitch contour, power spectrum, etc.) of the signal. By arbitrary modification it is meant that one can change the duration of any region of the signal by specifying the starting and ending time for modification and the target duration of the specified interval, which can be either a fixed value of duration in time or a scaling factor of the original duration. Modification of multiple intervals is to be permitted.

The arbitrariness or flexibility in modification includes allowing a "contour input" which defines the durational expansion or contraction factor as a function of time, an example of which is plotted below:



4. Voice Scrambling

In speech communication, one way to maintain privacy of conversation is to employ a digital voice encoder, which turns the voice signal into a bit stream, and a digital encryption system, which encrypts the bit stream into another bit stream that cannot be decoded back into a useful bit stream without the encryption key. This is in the class of "secure communication" and requires a sophisticated encryption system to do the job right. In this challenge, we explore a much simpler type of secure voice communication, often called "voice scrambling", which is only designed to make the voice not readily intelligible. The level of sophistication in voice scrambling is not very high and thus not very secure; however, for tactical or temporary privacy, it may suffice.

In voice scrambling, the spectrum of a voice signal is split by the scrambler into a number of bands, which are then permuted to form a scrambled spectrum according to some permutation pattern. At the descrambler, which knows the permutation matrix, the received signal undergoes the reverse of the band-splitting and permutation procedure to reconstruct an intelligible voice signal.

Band-splitting is executed by filtering and the degree of privacy of a voice scrambler depends on the number of bands that is implemented in it. However, since band-splitting filters can never be perfect (of the pure brick wall type), the number of bands is limited by the amount of sidelobes, which affects the quality of the reconstructed signal. The degree of privacy also depends on how often the permutation is altered. Without time-varying permutation, some people claim to be able to learn to understand the scrambled voice without de-permutation. Time-varying permutation of the speech spectrum nonetheless induces what is called bandwidth expansion, resulting in a loss of voice quality upon reconstruction, even with perfect knowledge of the permutation matrix. One way to reduce this type of degradation is to change the permutation scheme during pauses or gaps in a speech utterance.

This challenge aims at developing a voice scrambling system, both the transmitter/scrambler and the receiver/descrambler, that allows proper time variation of the spectrum permutation. The system should allow user to specify the number of spectral channels/bands for scrambling and a nominal, permissible rate of permutation change, say from 0 (fixed permutation) to 2 per second in fractional steps (e.g., 1/2, one change in permutation every two seconds, 2/5, two changes every 5 seconds, and so on).

5. Spoken Utterance Recognition

The target system performs spoken utterance recognition, a very basic form of automatic speech recognition where each given utterance corresponds to a class identity such as a word or a phrase in the prescribed "vocabulary". (The vocabulary is thus a generalized one, containing not necessarily only words.) The code allows a user to register a "vocabulary" by recording all utterances of the intended classes. Once the vocabulary is registered, the system performs training to optimize the system parameters for actual test.

The system therefore includes three essential modules: vocabulary registration, model training, and utterance recognition. During vocabulary registration, the system prompts to the user for recording utterances with the user first entering the utterance identification (UID), followed by several rounds of recording prompts, until all intended vocabulary entries are registered and recorded. In model training, the system uses the recorded utterances to train internal representations (models) of each UID class for use during the recognition stage. Finally, during utterance recognition, the system evaluates scores for each UID class for the unknown utterance based on the previously trained models and makes decision on the most likely UID. The final output is the recognized UID.

6. Noise Suppression for Speech Signals

This challenge is the easiest to understand among all six challenges but not less difficult to accomplish. Noise is inevitable in many real-world signals and it is always desirable to reduce the amount of noise contamination to enhance the signal for final consumption. The so-called enhancement effects include less noisy, higher clarity, and/or less fatiguing to listen to.

Noise present in a speech signal could come from the ambient such as the fan noise in a car or background conversation in a bar. In research, it is also often assume to be white noise to make the search for solutions a bit more straightforward (not necessarily more successful). The code to be developed will assume a particular type of noise, i.e., white or babble noise, upon your choice. When a speech signal with simultaneous presence of the noise is given as input, the system will process it to produce the enhanced output, which will sound less noisy and hopefully easier to listen to with higher clarity.