Speech enhancement

When using speech technology in real environments, we are often faced with less than perfect signal quality. For example, if you make a phone call at cafeteria, typically you have plenty of other people speaking in the background, there could be music playing and the room itself can have reverberation. Such effects distort the desired speech signal such that the receiving end, the desired speech sounds less pleasant, requires more effort to understand or at the worst case, it becomes less intelligible. *Speech enhancement* refers to methods which try to reduce such distortions, to make speech sounds more pleasant, reduce listening effort and improve intelligibility.

The most prominent categories of speech enhancement are:

1. Noise attenuation, where we try to extract the desired speech signalm when distorted by background noise(s).
2. Echo cancellation and feedback cancellation are used when the sound played from a loudspeaker is picked up by a microphone distorting the desired signal.
3. Dereverberation refers to methods which attenuate the effect of room acoustics on the desired signal.
4. Source separation methods try to extract sounds of single sources from a mixture, for example, in the classical cocktail-party problem, we would like to isolate single speakers when multiple people are talking at the same time.
5. Beamforming refers to spatially selective methods, where the objective is isolate sounds coming from a particular direction, by using the information about the spatial separation of a set of microphones.

The objective of speech enhancement however requires a bit more consideration. In its most classical form, the objective is to extract a clean speech signal from a distorted mixture, where the distortions can be background and sensor noises, as well as room reverberation. Here the clean reference signal is considered to be that signal which would be rerecorded with a microphone close to the speaker, which does not contain said noises or reverberation. It is then clear that it will be challenging to obtain realistic data, since even a microphone close to the speaker will usually contain background noises and effect of reverberation. For development of methods, it is therefore often difficult to obtain data which would accurately correspond to a realistic situation. In any case, a typical objective would be to improve the signal to noise ratio (with or without perceptual weighting) as much as possible.

A more challenging scenario is when two or more persons are speaking in the same acoustic environment. The second speaker can then be viewed as a competing speaker (undesired source) or as a discussion partner (desired source). Even if the two speakers are in an interaction with each other, then often they will speak on top of each other, even if stereotypically we think of a dialogue as a non-overlapping back and forth exchange of non-overlapping arguments. If we want to separate between the two speakers, then overlaps are difficult, because the statistics of the both speech signals will be rather similar, whereas noise signals with distinct statistics are easier to attenuate.

Sometimes we do not want to remove all distortions entirely, but just attenuate their effect. Completely removing artefacts can sometimes make the signal sound unnatural and besides removing distortions, processing methods also almost always distorts the desired signal. Therefore, to retain a natural-sounding signal and to minimize distortion of the desired speech signal, we often limit the extent to which distortions are removed.

A further aspect of enhancement is intelligibility and pleasantness; as a starting point, observe that the speech of some people is by nature difficult to understand or otherwise just annoying (unpleasant). It then conceivable that we device some processing which improves the speech signal to better than the original. What "sounds better" is however a difficult concept, since we do not have unambiguous measures for "how good it sounds" and opinions between listeners will certainly diverge.

Intelligibility with regard to human listeners is similarly complicated as pleasantness, but luckily, we can use speech recognition engines to obtain objective measures. That is, if we give noisy and improved speech signals to a speech recognizer, we can determine the recognition performance in both cases to estimate the benefit obtained with our processing.