There are two observations for speech recognition: 1) MFCC features perform better than simple cepstrum coefficients at any of the window size, and 2) either too small or large window sizes will decrease the accuracy of recognition. The first part is probably because MFCC uses a mel scale rather than a hertz scale, which is nonlinear in frequency. It compresses high frequencies and differentiates at speech-related frequencies, so it can differentiate speeches more correctly. At small window sizes, an increase in window size can dissect the phonemes better thus introducing phoneme-related features, but a too large window (10000 above) would provide too little feature for classification.

For the speaker recognition, MFCC does a little better than cepstrum, but the difference is tiny. In contrast, an increase in window size constantly increases the accuracy of speaker recognition. This is probably the result of the cepstrum coefficients we choose (Ncc = 12 above). Different from speech recognition, the spacing between harmonics ($F\_0$) is the key difference among speeches produced by different persons, and the information about $F\_0$ lies in the high quefrency cepstrum. Shrinking the window size produces high low quefrency cepstrum (since it has less “noise” in the spectrum) that is easy for smoothing, but it makes the detection of frequency-varying $F\_0$ harder (similar to the uncertainty principle). Therefore, whether to use CC or MFCC does not make much difference (since we are always choosing the least 12 cepstrum coefficients), while increasing the window size makes $F\_0$ easier to detect. Another reason for low accuracy of speaker recognition is that all speakers in training data are males. Although humans percepts those speakers differently, males’ excitations are in certain range. In this case, either cepstrum or MFCC will struggle to detect difference with simple nearest neighbor search.

identify