# Literature Review

## Speech Processing

Automated speech recognition (ASR) systems can be broken up into numerous categories based on the various criteria they meet. The criteria for systems usually consist of; speaker dependence, speech type, and recognition type.

### Speaker dependence

Speaker dependent systems are trained for use by a single speaker, building classification models based on the speaker’s unique acoustic-phonetic model. In contrast, speaker independent systems are designed for use by numerous speakers, including those who were not involved in the system training process. Speaker independence is hard to achieve due to the feature parameterization becoming tuned to the training speaker(s), causing a speaker-specific bias in the classification [1]. Error rates for speaker independent systems tend to be 3 to 5 times larger than speaker dependent systems [2].

### Speech type

Speech can be broken into three distinct types; isolated, discontinuous, and continuous [1]. Isolated speech consists of singular words and is often considered word recognition. Discontinuous speech involves the speaker being purposefully articulate and inserting artificial pauses between consecutive words. In continuous speech (natural speech) the speaker makes no effort to alter their speech patterns. Recognition for isolated and discontinuous speech is simpler due to the clearly defined word boundaries and distinct pronunciation. Continuous speech is much harder to process due to the undefined word boundaries in addition to corrupted pronunciation introduced by co-articulation, the slurring of speech sounds, which can cause phrases like “could you” to sound like “could jou” [1]. During a standard evaluation, isolated and continuous speech achieved error rates of 3.1% and 8.7% respectively [3].

### Recognition type (word vs. phoneme)

A phoneme can be defined as “the smallest contrastive linguistic unit which may bring about a change of meaning [of a word]” [4].

Whether the system performs recognition at the word or phoneme level

## Acoustical Pre-Processing

## Feature Extraction

The two most important sections of speech characteristics are those contained in the spectral envelope (vocal tract characteristics) and those contained in the supra-segmental features (voice source characteristics) [5].

### Mel-frequency Cepstral Coefficients (MFCC)

## Classification Techniques

### Hidden Markov Models

### Artificial Neural Networks

### Dynamic Time Warping

## Robust speech recognition

## Auditory data

# References

[1] J. Tebelskis, "Speech recognition using neural networks," Carnegie Mellon University, 1995.

[2] K.-F. Lee, "On large-vocabulary speaker-independent continuous speech recognition," *Speech Communication,* vol. 7, pp. 375-379, June 21 1988.

[3] L. Bahl, R. Bakis, P. Cohen, A. Cole, F. Jelinek, B. Lewis*, et al.*, "Speech recognition of a natural text read as isolated words," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '81.*, 1981, pp. 1168-1171.

[4] A. Cruttenden, "Gimson's pronunciation of English," 7th ed: Routledge, 2013, p. 41.

[5] S. Furui, "Recent advances in speaker recognition," in *Audio- and Video-based Biometric Person Authentication*. vol. 1206, J. Bigün, G. Chollet, and G. Borgefors, Eds., 1st ed: Springer Berlin Heidelberg, 1997, pp. 235-252.