Table of Contents

1 Literature Review 1

1.1 Speech Processing 1

1.1.1 Speaker dependence 1

1.1.2 Speech type 2

1.1.3 Recognition type 2

1.2 Pre-Processing 3

1.3 Feature Extraction 4

1.3.1 Mel-frequency Cepstral Coefficients (MFCC) 4

1.3.2 Linear Predictive Coding (LPC) 4

1.4 Classification Techniques 6

1.4.1 Dynamic Time Warping 6

1.4.2 Artificial Neural Networks 6

1.4.3 Hidden Markov Models 6

1.5 Auditory data 6

1.6 Robust speech recognition 6

1.6.1 Codebook Excited Linear Prediction (CELP) 6

1.6.2 Non-Negative Matrix Factorization 6

2 References 7

# 

# Literature Review

## Speech Processing

Automated speech recognition (ASR) systems can be partitioned 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

There are two main classifications for ASR systems; word recognition and phonetic recognition. Word recognition is desirable due to the human interpretable aspect of the transcription, that is, all results are complete words. However, the main disadvantage of word based ASR is that the error rates of the system are proportional to the vocabulary size. For small sets of words it is possible to obtain <1% error rates [4], whereas vocabulary sizes of 200, 5000, or 100000 could have approximate error rates of 3%, 7%, or 45% respectively [5-7]. Additionally, the processing time for a word-based system is also proportional to the vocabulary size, making it impractical for most real-world applications.

Conversely, phonetic recognition is able to avoid both error rate and processing time inflation by breaking words down into their base components, phonemes. A phoneme can be defined as “the smallest contrastive linguistic unit which may bring about a change of meaning [of a word]” [8], which results in a classification base of approximately 42 phonemes for the English language (see Table 1‑1), with the drawback being that the transcription isn’t directly human readable (see Table 1‑2).

Table ‑: List of English Phonemes and Manners of Articulation [9]

|  |  |  |  |
| --- | --- | --- | --- |
| **Phoneme** | **Manner of Articulation** | **Phoneme** | **Manner of Articulation** |
| iy | vowel | l | liquid |
| ih | vowel | r | liquid |
| ia | vowel | m | nasal |
| ey | vowel | n | nasal |
| eh | vowel | ng | nasal |
| ae | vowel | f | fricative |
| ea | vowel | v | fricative |
| aa | vowel | th | fricative |
| ao | vowel | dh | fricative |
| ow | vowel | s | fricative |
| uh | vowel | z | fricative |
| uw | vowel | sh | fricative |
| ua | vowel | zh | fricative |
| ah | vowel | hh | fricative |
| er | vowel | p | stop |
| ax | vowel | b | stop |
| ay | diphthong | t | stop |
| oy | diphthong | d | stop |
| oh | diphthong | k | stop |
| aw | diphthong | g | stop |
| y | glide | ch | affricate |
| w | glide | jh | affricate |

Table ‑: Example of the Phonetic Decomposition of Various Words/Phrases [9]

|  |  |
| --- | --- |
| **Word/Phrase** | **Phonetic Decomposition** |
| call | k ao l |
| dial | d ay ax l |
| seven | s eh v n |
| recognise speech | r eh k ao g n ay z s p iy ch |
| wreck a nice beach | r eh k ay n ay s b iy ch |

## Pre-Processing

Pre-processing is used as a means of cleaning and normalizing a speech signal to allow for easier and more reliable feature extraction, as most speaker independent ASR systems perform very poorly when tested in environments different from the one in which they were trained [10]. Two major factors that contribute to signal distortion are additive noise and signal convolution with an unknown linear system [11].

## Feature Extraction

Feature extraction is the process of parsing an input signal and parameterizing it into a vector, known as a feature vector, which contains the important characteristics of the signal, as deemed by the chosen parameterization process. 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) [12].

* Cepstral and delta cepstral coefficients
* Non-negative matrix factorization
* Discrete fourier transform methods
* Discrete wavelet packet transforms
* Linear Predictive Coding
* Perceptual Linear Prediction

### Mel-frequency Cepstral Coefficients (MFCC)

### Linear Predictive Coding (LPC)

Linear Predictive Coding is a method of feature extraction that attempts to predict future values of an input signal based on the past values of that signal [13].

The features are coefficients of linear filters.

LPC can be broken into two distinct segments, analysis/encoding and synthesis/decoding. During the encoding stage the speech signal is broken in blocks or frames, which are then processed to determine the corresponding filter coefficients that would be capable of reproducing that frame of speech. The decoding stage involves rebuilding the speech signal from the received filter coefficients. For the purposes of this thesis, only the encoding stage is considered, as speech reconstruction is not necessary for classification.

The encoding phase of LPC feature extraction involves two main steps:

Determining whether a block is voiced or unvoiced

This is done by taking the assumption that voiced sounds have a high average energy (thus large amplitudes) as well as having distinct formant frequencies.

Unvoiced sounds have “random” waveforms of a lower amplitude but higher frequency.

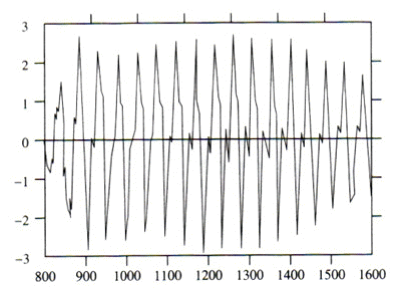


Figure : Voiced sound - Letter 'e' in the word 'test' [13]

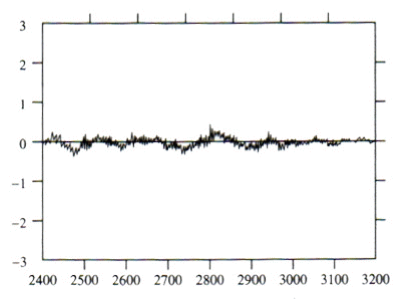


Figure : Unvoiced sound - Letter 's' in the word 'test' [13]

Due to higher frequencies, unvoiced speech must cross the x-axis more times than voiced speech.

Neighboring segments are also taken into consideration when determining the vocal nature of the current segment, as there are likely correlations due to the nature of human speech, e.g. it’s improbable to have unvoiced segments contained in a group of voiced segments and vice versa.

LPC isn’t very robust as it’s classification methods are based on assumptions that are easily interfered with by additive noise. Background noise can increase the average energy of an unvoiced segment enough to misclassify, as much as it can increase the frequency of a voiced segment.

Pitch Period determination

## Classification Techniques

### Dynamic Time Warping

### Artificial Neural Networks

### Hidden Markov Models

## Auditory data

## Robust speech recognition

### Codebook Excited Linear Prediction (CELP)

### Non-Negative Matrix Factorization

# 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] G. R. Doddington, "Phonetically sensitive discriminants for improved speech recognition," in *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, 1989, pp. 556-559.

[5] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 23, pp. 67-72, 1975.

[6] M. Miyatake, H. Sawai, Y. Minami, and K. Shikano, "Integrated training for spotting japanese phonemes using large phonemic time-delay neural networks," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, 1990, pp. 449-452.

[7] S. Kimura, "100000-word recognition using acoustic-segment networks," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, 1990, pp. 61-64.

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

[9] A. Bath, "Robust Speech Recognition via Human Auditory Modeling," Bachelor of Engineering (Electrical) / Bachelor of Science (Computer Science), School of Engineering and Physical Sciences, James Cook University, 2010.

[10] R. M. Stern, F.-H. Liu, Y. Ohshima, T. M. Sullivan, and A. Acero, "Multiple approaches to robust speech recognition," in *Proceedings of the workshop on Speech and Natural Language*, 1992, pp. 274-279.

[11] A. Acero and R. Stern, "Environmental robustness in automatic speech recognition," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, 1990, pp. 849-852 vol.2.

[12] 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.

[13] J. Bradbury, "Linear predictive coding," *Mc G. Hill,* 2000.