

Chapter 11

On-Line Adaptive Speech Recognition

Prof. Nik Kasabov

Nkasabov@aut.ac.nz

<http://www.kedri.info>

Overview

- Introduction to the adaptive speech recognition problems.
- A framework of an evolving connectionist system for adaptive speech recognition.
- On-line, adaptive phoneme-based speech recognition.
- On-line, adaptive whole word and phrases recognition.
- Feature selection and feature evaluation for on-line adaptive speech recognition systems
- Natural language understanding for adaptive, intelligent human computer interfaces.

Introduction

- Need new methods that deal with the problems of noise and adaptation in order for these technologies to become common tools for communication and information processing.
- An adaptive system can learn spoken phonemes, words and phrases.
- New words, pronunciations, and languages can be introduced to the system in an incremental, adaptive way.

Introduction to the Adaptive Speech Recognition Problems

- Speech recognition is one of the most challenging applications of signal processing
- Speech is a sequence of waves that are transmitted over time through a medium and are characterised by some features, among them - intensity and frequency
- Biological background of speech recognition is used by many researchers to develop human-like Automatic Speech Recognition
- Speech can be represented on different scales:
 - » time scale, which representation is called waveform representation;
 - » frequency scale, which representation is called spectrum;
 - » both time and frequency scale - this is the spectrogram of the speech signal.

Introduction to the Adaptive Speech Recognition Problems

- 3 factors which provide the easiest method of differentiating speech sounds are the perceptual features of loudness, pitch and quality
- A spectrogram of a speech signal shows how the spectrum of speech changes over time.
 - » The horizontal axis shows time and the vertical axis shows frequency
- Speech signal is highly variable due to
 - » different speakers
 - » different speaking rates
 - » different contexts and
 - » different acoustic conditions.
- The speech signal is very dependent on the physical characteristics of the vocal tract, which in turn are dependent
 - » on age and gender
 - » country of origin of the speaker

Introduction to the Adaptive Speech Recognition Problems

- Different rhythm and intonation due to different accents
- If English is the second language of a speaker, there can be an even greater degree of variability in the speech
- variability in the way they speak, depending on whether it is a formal or informal situation.
- Speed of speech varies due to such things as the situation and emotions of the speaker
- variability of speech requires robust and adaptive systems
- adaptive speech recognition problem is concerned with the development of methods and systems for
 - » speaker-independent recognition
 - » high accuracy
 - » capable to adapt fast to new words, new accents, new speakers for a small-, medium-, to large vocabulary of words, phrases and sentences.

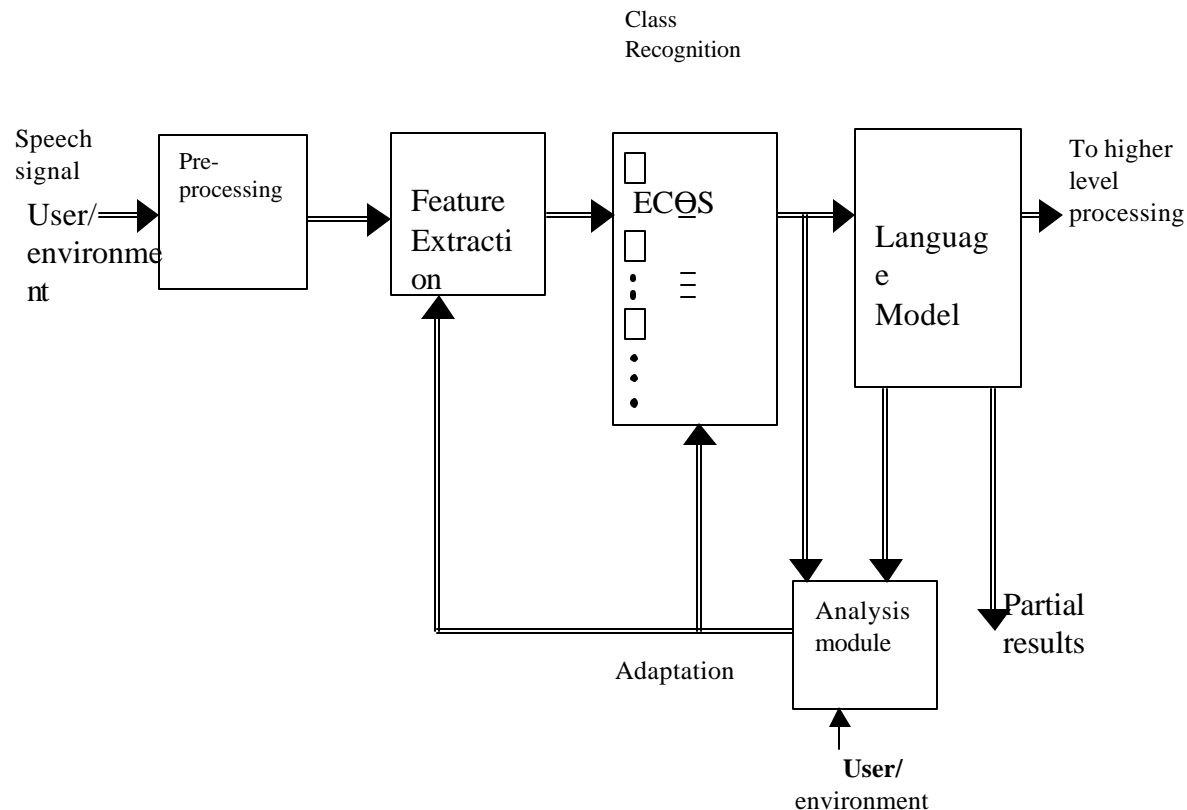
Introduction to the Adaptive Speech Recognition Problems

- Humans can adapt to different accents of English, e.g. American, Scottish, New Zealand, Indian
- Spoken language modules in the human brain evolve continuously
- every time we speak, we pronounce same sounds of the same language at least slightly differently

A Framework of Evolving Connectionist Systems for Adaptive Speech Recognition

- It consists of the following modules and procedures:
 - » Pre-processing module
 - » Feature extraction module
 - » Pattern classification (modelling) module
 - » Language module
 - » Analysis module
- The set of features selected, depends on the organization and on the function of the pattern classifier module (e.g. phoneme recognition, whole word recognition, etc)
- Pattern (class) recognition module can be trained to recognize phonemes, or words, or other elements of a spoken language.
- New words and phrases can be added or deleted from the system at any time of its operation
- Recognized words and phrases at consecutive time moments are stored in a temporal buffer.
- Temporal buffer is fed into a sentence recognition module where multiple-word sequences (or sentence) are recognized

A Framework of Evolving Connectionist Systems for Adaptive Speech Recognition



A block diagram of an adaptive speech recognition system framework that utilises ECOS in the recognition part (fig. 11.2)

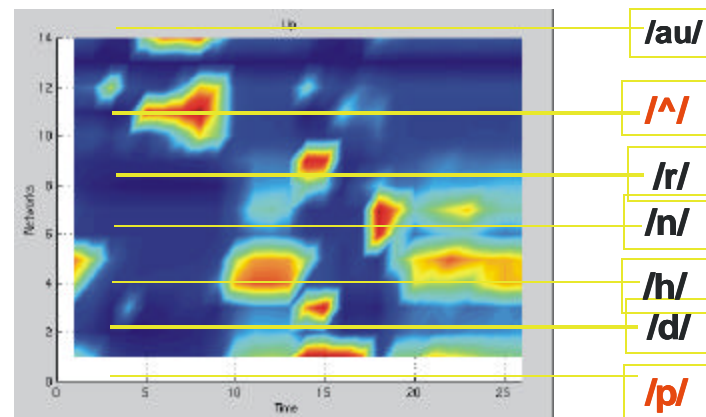
On-line Adaptive Phoneme Based Speech Recognition

- Recognising phonemes from a spoken language is a difficult, but important problem.
- Can recognize the words and the sentences of a spoken language.
- Pronounced vowels and consonants differ depending on the accent, dialect, health status, etc of the person
- There are different neural network (NN)-based models for speech recognition that utilise
 - » MLP
 - » SOM
 - » RBF networks
 - » time-delay NN (Weibel et al, 1989; Picone, 1993),
 - » hybrid NN and hidden Markov models (Rabiner, 1989; Trentin, 2001)
- All these models use usually one NN for the classification of all phonemes and they work in an off line mode.

On-line Adaptive Phoneme Based Speech Recognition

- An approach is used, where each NN module from a multi-modular system is trained on a single phoneme data and the training is in an on-line mode
- Single phoneme NN can be adapted to different accents and pronunciations without necessarily re-training the whole system

The activation of seven phoneme NN modules, trained on their corresponding phoneme data when an input signal of “up” is submitted (Fig 11.5)

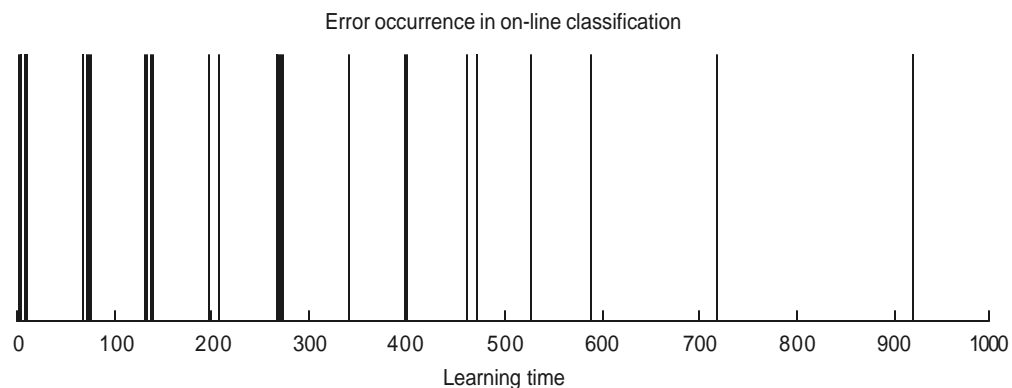


On-line Adaptive Phoneme Based Speech Recognition

- Phoneme modules miss-activation problem can be overcome through analysis of the sequence of the recognised phonemes and forming the recognized word through a matching process using a dictionary of words.
- To improve the recognition rate, the wrongly activated phoneme NN modules can be further trained not to react positively on the problematic for them phoneme sounds.
- Each of the phoneme NN module can be further adapted to a new accent, e.g. Australian English

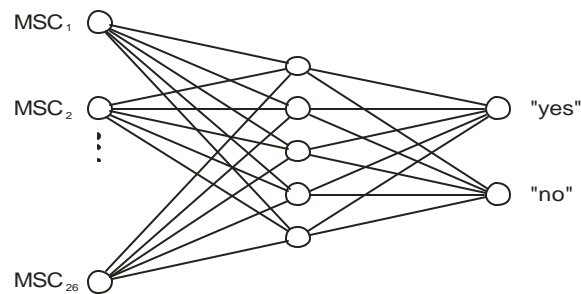
On-line Adaptive Phoneme Based Speech Recognition

- ESOM was used for the classification of phoneme data
- advantage of ESOMs as classifiers is that they can be trained (evolved) in a life-long mode
- Longer system is training, the lower the error rate
- *Error rate of an ESOM system trained in an online learning mode (Fig. 11.6)*



On-line, Adaptive Whole Word and Phrases Recognition

- Using a NN for the recognition of a whole word
- As inputs, 26 mel-scale cepstrum coefficients taken from the whole word signal, are used
- Each word is an output in the classification system
- *An illustration of an NN for a whole-word recognition problem on the recognition of two words – “yes” and “no” (Fig. 11.7)*



On-line, Adaptive Whole Word and Phrases Recognition

- Speech signal is processed so that the segment that represents a spoken word is extracted from the rest of the signal
- Problems with ambiguity of speech
- Ambiguity is resolved by humans through some higher-level processing
- Ambiguity can be caused by:
 - » Homophones - words with different spellings (for example "to, too, two" or "hear, hair, here").
 - » Word boundaries - extracting whole words from a continuous speech signal may lead to ambiguities, for example /greiteip/ could be interpreted as "grey tape" or "great ape".
 - » Syntactic ambiguity – the phrase 'the boy jumped over the stream with the fish' - means either the boy with the fish jumped over the stream, or the boy jumped over the stream with a fish in it

On-line, Adaptive Whole Word and Phrases Recognition

- The complexity is basically affected by:
 - » vocabulary size and word complexity.
 - small, tens of words;
 - medium, hundreds of words
 - large, thousands of words
 - very large, tens of thousands of words;
 - » format of the input speech data entered to the system:
 - isolated words (phrases);
 - connected words; this represents fluent speech but in a highly constrained vocabulary, e.g. digit dialling;
 - continuous speech.
 - » The degree of speaker dependence of the system:
 - speaker dependent
 - multiple speakers
 - speaker independent

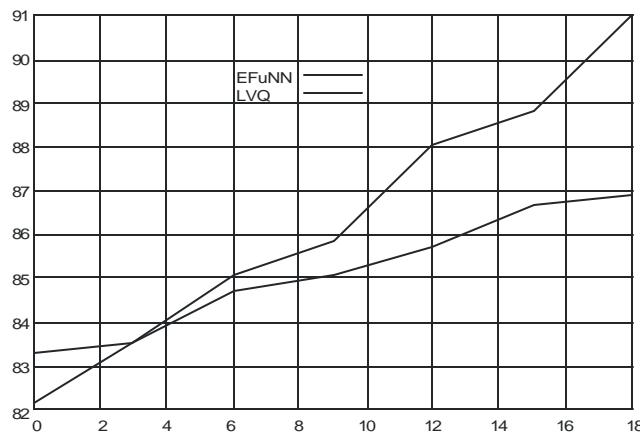
A Case Study of Adaptive On-line Digit Recognition – English Digits

- Recognition of speaker independent pronunciations of English digits
- 17 speakers (12 males and 5 females) are used for training,
- 17 other speakers (12 males and 5 females) are used for testing
- EFuNN-based classification system
- Comparison with the Linear Vector Quantization (LVQ) method
- First instance, car noise is added to the clean speech
- Second instance office noise is introduced over the clean signal

A Case Study of Adaptive On-line Digit Recognition – English Digits

- The EFuNN method outperforms the LVQ method

Word recognition rate (WRR) of two speech recognition systems (LVQ, EFuNN) when car noise is added (Fig. 11.8)



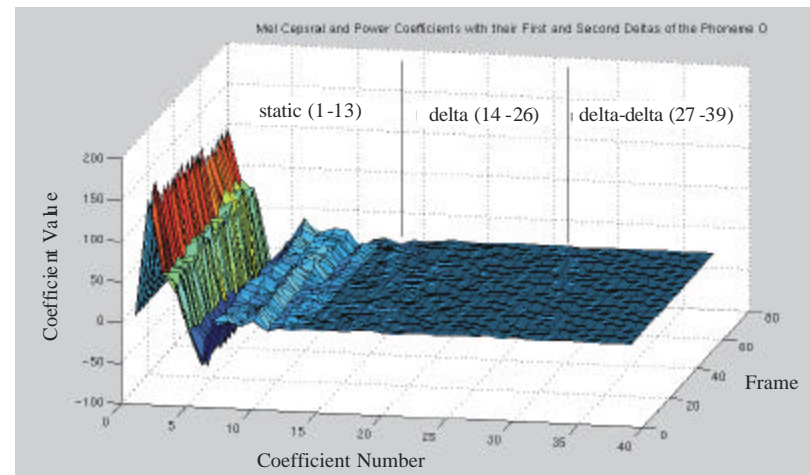
Feature Selection and Feature Evaluation for On-line Adaptive Speech Recognition Systems

- Feature selection process is an extremely important issue for every speech recognition system
- Many current approaches towards speech recognition systems use Mel frequency cepstral coefficients (MCCs) vectors to represent each 10 to 50 ms window of speech samples, taken each 5 to 25 ms, by a single vector of certain dimension
- For many applications the most effective components of the Mel scale features are the first 12 coefficients (static coefficients)
- MCC have been very successfully used for off-line learning, static speech recognition systems
- For on-line learning adaptive systems more appropriate set of features are combinations between static and dynamic features

Feature Selection and Feature Evaluation for On-line Adaptive Speech Recognition Systems

- Recently shown that the speech recognition rate is noticeably improved when using additional coefficients representing the dynamic behaviour of the signal
- Coefficients are the first and second derivatives of the cepstral coefficients of the static feature vectors.
- Power coefficients, and their first and second derivatives, also have important roles to be included in the representation of the feature vectors.
- Static coefficients will dominate the effect of the dynamic coefficients
- Using dynamic features also increases the dimensionality of the feature vectors.

Feature Selection and Feature Evaluation for On-line Adaptive Speech Recognition Systems



The power and 12 Mel scale coefficients, with their first and second derivatives, of a phoneme /o/ sound (Fig. 11.10)

Feature Selection and Feature Evaluation for On-line Adaptive Speech Recognition Systems

- Other features that account for dynamic changes of the speech signal are
 - » Wavelets
 - » Gamatonne feature vectors
- It is appropriate to use different sets of features in different modules if a modular speech recognition system is built

Natural Language Understanding for Adaptive, Intelligent Human Computer Interfaces

- Speech recognition and language modelling systems can be developed as main parts of an intelligent human computer interface to a database.
- Data entry and a query to the database can be done through a voice input
- Natural language understanding is an extremely complex phenomenon.
- Involves recognition of sounds, words and phrases, as well as their comprehension and usage.

Natural Language Understanding for Adaptive, Intelligent Human Computer Interfaces

- Various levels in the process of language analysis
 - » prosody - deals with rhythm and intonation;
 - » phonetics - deals with the main sound units of speech (phonemes) and their correct combination;
 - » lexicology - deals with the lexical content of a language;
 - » semantics - deals with the meaning of words and phrases seen as a function of the meaning of their constituents;
 - » morphology - deals with the semantic components of words (morphemes);
 - » syntax - deals with the rules, which are applied to form sentences;
 - » pragmatics - deals with the language usage and its impact on the listener.
- Importance of language understanding in communication between humans and computers, which was the essence of the Alan Turing's test for AI

Natural Language Understanding for Adaptive, Intelligent Human Computer Interfaces

- Computer systems for language understanding require methods that can represent
 - » Ambiguity
 - » common sense knowledge
 - » hierarchical structures.
- Humans, when they communicate between each other, share a lot of common sense knowledge which is inherited and learned in a natural way.
- Humans use face expressions, body language, gestures and eye movement when they communicate between each other
- Computer systems which analyse speech signals, gestures and face expressions when communicate with users are called multi-modal systems

Summary

- The applicability of evolving, adaptive speech recognition systems is broad and spans across all application areas of computer and information science
- Where systems that communicate with humans in a spoken language ('hands-free and eyes-free environment'). This includes:
 - » Voice dialling, especially when combined with "hands-free" operation of a telephone system (e.g. a cell phone) installed in a car. Here a simple vocabulary that includes spoken digits and some other commands would be sufficient.
 - » Voice control of industrial processes.
 - » Voice command execution – the controlled device could be any terminal in an office. This provides a means for people with disabilities to perform simple tasks in an office environment.
 - » Voice control in an aircraft.

Further Readings

- Reviews on speech recognition problems, methods, and systems (Cole, 1995; Lippman, 1989; Rabiner, 1989; Kasabov, 1996);
- Signal processing (Owens, 1993; Picone, 1993)
- Neural network models and systems for speech recognition (Morgan and Scofield, 1991).
- Phoneme recognition using time-delay neural networks (Waibel, et al, 1989).
- Phoneme classification using radial basis functions (Renals and Rohwer, 1989).
- Hybrid NN- HMM models for speech recognition (Trentin, 2001).
- A study on acoustic difference between RP English, Australian English and NZ English (Maclagan, 1982).
- Evolving fuzzy neural networks for phoneme recognition (Kilgour, 2001; Kasabov, 1998, 2000).
- Evolving fuzzy neural networks for whole word recognition – English and Italian digits (Kasabov and Iliev, 2000).
- Evolving self-organising maps for adaptive on-line vowel classification (Deng and Kasabov, 2002).