Seminar I

(2110716)

Nuttakorn Thubthong

ID. 4171815021

Department of Computer Engineering
Faculty of Science
Chulalongkorn University
1998

Speech Recognition Review

Speech is a natural mode of communication for people. We learn all the relevant skills during early childhood, without instruction, and we continue to rely on speech communication throughout our lives. It comes so naturally to us that we don't realize how complex a phenomenon speech is. The human vocal tract and articulators are biological organs with nonlinear properties, whose operation is not just under conscious control but also affected by factors ranging from gender to upbringing to emotional state. As a result, vocalizations can vary widely in terms of their accent, pronunciation, articulation, roughness, nasality, pitch, volume, and speed; moreover, during transmission, our irregular speech patterns can be further distorted by background noise and echoes, as well as electrical characteristics (if telephones or other electronic equipment are used). All these sources of variability make speech recognition, even more than speech generation, a very complex problem.

The current state of the art in speech recognition depends on the conditions under which it is evaluated: under sufficiently narrow conditions almost any system can attain human-like accuracy, but it's much harder to achieve good accuracy under general conditions. The conditions of valuation-and hence the accuracy of any system-can vary along the following dimensions:

- Vocabulary size and confusability
    - 10 digits can be recognized essentially perfectly (Doddington 1989)
    - Vocabulary sizes of 200, 5000, or 100000 may have error rates of 3%, 7%, or 45% (Itakura 1997, Miyatake 1990, Kimura 1990)

- Speaker dependence vs. Independence
    - Single speaker system is intended for use by a single speaker
    - Multi-speaker system is intended for use by a small group of people
    - Speaker-adaptive system can be tune themselves to any speaker

- Isolated, discontinuous, or continuous speech
    - Isolated speech means single words.
    - Discontinuous speech means full sentences in which words are artificially separated by silence.
    - Continuous speech means naturally spoken sentences.

- Task and language constraints
    - Task-dependent
    - Semantic
    - Syntactic
    - Grammar

- Read vs. Spontaneous speech
    - Spontaneous speech is vastly more difficult, because it tends to be peppered with disfluencies like “uh” and “um”, false states, incomplete sentences, stuttering, coughing, and laughter; and moreover, the vocabulary is essentially unlimited.

- Adverse conditions
    - environmental noise
    - acoustical distortions
    - different microphones
    - limited frequency bandwidth
    - altered speaking manner

The central issue in speech recognition is dealing with variability. Currently, speech recognition systems distinguish between two kinds of variability:

1. Acoustics variability

Covers different accents, pronunciations, pitches, volumes, etc.

2. Temporal variability

Covers different speaking rates

Acoustic variability is more difficult to model, partly because it is so heterogeneous in nature. Consequently, research in speech recognition has largely focused on efforts to model acoustic variability. Past approaches to speech recognition have fallen into three main categories:

1. Template-based approaches, is which unknown speech is compared against a set of prerecorded words (templates), in order to find the best match. This has the advantage of using perfectly accurate word models; but it also has the disadvantage that the prerecorded templates are fixed, so variations in speech can only be modeled by using many templates per word, which eventually becomes impractical.
2. Knowledge-based approaches, in which “expert” knowledge about variations in speech is hand-coded into a system. This has the advantage of explicitly modeling variations in speech; but unfortunately such expert knowledge is difficult to obtain and use successfully, so this approach was judged to be impractical, and automatic learning procedures were sought instead.
3. Statistical-based approaches, in which variations in speech are modeled statistically (e.g., by Hidden Markov Models, or HMMs), using automatic learning procedures. This approach represents the current state of the art. The main disadvantage of statistical models is that they must make a priori modeling assumptions, which are liable to be inaccurate, handicapping the system’s performance. We will see that neural networks help to avoid this problem

In English language, there are several possible choices for subword units that can be used to model a speech system, including the following:
(Rabiner and Juang 1993)

- Phoneme-like units 50 PLUs
- Syllable-like units 10,000 syllables
- Demisylable-like units 2,000 demisyllable-like unit

In past, phoneme-based have been often used because of the less numbers of phoneme-like units. Phoneme-based recognition has been attempted using both frame-based and segment based approaches, in which systems also have disadvantages and advantages. But the frame-based systems are currently more popular since they do not require explicit of segment boundaries. Evidence suggesting that the transitional part of speech carries important information for speech perception exists, therefore, transition based have been used. (Hu 1995)

One idea arising again recently is the proposal of a paradigm shift from phoneme-based to syllable-based recognition system. (Hu 1996; Pfitzinger 1996; Ganapathiraju 1997)

Some of the potential advantages of syllable-base ASR are: (Hauenstein 1996)

- The human auditory system integrates time spans of about 200 ms of speech, which corresponds roughly to the duration of syllables. Thus the very robust human perception may be modelled more accurately by use of syllables instead of phonemes
- The relative duration of syllables is less dependent on variations in speaking rate than the relative duration of phonemes. Therefore the mismatch between the observation window for classification (feature vectors including multiple frames and Δ-components) and the duration of the unit classified is reduced for speakers whose speaking rate varies from the average.
- It was shown that time spans of 250 ms are suitable for methods of *cepstral mean subtraction* in order to suppress convolutional noise.

A robustness speech recognition system is currently the one interesting topics, such as a speech recognition system in noise and echo environments.(Barnard 1995; Kingsbury 1997) The past feature extraction method were not able to solve the noise in the echo environment problem. Therefore, many methods have been designed to resolve these system, namely RASTA-PLP. (Hermansky and Morgan 1995; Shire 1997)

A *dynamic time wraping* was a well-known method for pattern comparisons, after that *Hidden Markov Model* (HMM) and *Neural Network* (NN) have been applied to speech recognition system.(Vermeulen 1995) Recently, both systems have been combined as a hybrid systems to achieve the accuracy rates.(Tebelskis 1995)

Over the past few years, Support Vector Machine (SVM) is a new technique that has been created a lot of interest in the pattern recognition community (Ganapathiraju 1998), however, the results of these system are still not satisfied.

Refferences

1. Ahkuputra,W., et al. (1997). “A Speaker-Independent Thai Polysyllabic Word Recognition System Using Hidden Markov Model”, *Proceedings of the Natural Language Processing Pacific Rim Symposium* 1997, pp.281-286.

2. Arai, T. and Greenberg, S. (1997). “The Temporal Properties of Spoken Japanese Are Similar to Those of English”, *EUROSPEECH’97*, Rhodes, Vol. 2, pp.1011-1014.

3. Barnard, E., Cole, R., et al. (1995). “Real-world speech recognition with neural networks” Proceedings of the International Symposium on Aerospace/Defense Sensing & Control and Dual-Use Photonics, *International Society for Optical Engineering, Technical Conference no. 2492*, Orlando, FL, April 17-21, 1995.

4. Cole, R., Yan, Y., et al. (1996). “The Contribution of Consonants Versus Vowels to Word Recognition in Fluent Speech”, *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, Atlanta, GA, May, 1996.

5. Gannaparthiraju, A., Goel, V., et al. (1997). “Syllable-A Promising Recognition Unit for LVCSR” , *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, pp.207-214, Santa Barbara, California, USA, December 1997.

6. Ganapathiraju, A., Hamaker, J., and Picone, J. (1998). “Support Vector Machines for Speech Recognition”, *Submitted to the International Conference on Spoken Language Processing*, Sydney, Australia, October 1998.

7. Godfrey, J., Ganaphthiraju, A., and Picone, J. (1997). “Microsegment-Based Connected Digit Recognition”, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol.3*, pp 1755-1758, Munich Germany, April 1997.

8. Greenberg, S. (1997). “On The Origins of Speech Intelligibility in The Real World”, *Proceeding of the ESCA Workshop on Robust Speech Recognition of Unknow Communication Channels*, April, 1997.

9. Hamaker, J., Ganapathiraju, A., Picone, J., and Godfrey, J. (1998). “Advances in Alphadigit Recognition Using Syllables”, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol.1, pp421-424, Seattle, Washington, USA, May 1998.

10. Hauenstein, A. (1996). “The Syllable Re-revisited”, *ICSI Technical Report*, tr-96-035, Berkeley, California, August 1996.

11. Hermansky, H. (1990). “Perceptual linear predictive (PLP) analysis of speech”, *Journal of the Acoustical Society of America*, **87**(4):1738-1752, April 1990.

12. Hermansky, H. and Morgan, N. (1994). “RASTA processing of speech”, *IEEE Transactions on Speech and Audio Processing*, **2**(4):578-589, October 1994.

13. Hosom, J., and Cole, R. (1997). “A Diphone-Based Digit Recognition System Using Neural Networks”, *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, Munich, April 1997.

14. Hu, Z., Barnard, E., and Cole, R. (1995). “Transition-based Feature Extraction within Frame-based Recognition”, *EUROSPEECH'95*.

15. Hu, Z. Schalkwyk, J. and et al. (1996). “Speech Recognition Using Syllable-Like Units”, *Proceedings of the International Conference on Spoken Language Processing*, Philadelphia, PA, October, 1996.

16. Hu, Z., and Barnard, E. (1997). “Efficient Estimation of Perceptual Features of Speech Recognition”, *EUROSPEECH'97*.

17. Hu, Z., and Barnard, E. (1997). “Smoothness Analysis for Trajectory Features”, *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, Munich, April 1997.

18. Kingsbury, B., and Morrgan, N. (1997). “Recognizing Reverberant Speech with RASTA-PLP”, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Munich Germany, April 1997.

19. Kingsbury, B., Morgan, N. and Greenberg, S. (1997). “Improving ASR Performance for Reverberant Speech”, *Proceeding of the ESCA Workshop on Robust Speech Recognition of Unknow Communication Channels*, April, 1997

20. Lee, T., P.C.Ching, and et al. (1995). “Tone Recognition of Isolated Cantonese Syllables”, *IEEE Transactions on Speech and Audio Processing*, **3**(3), pp.204-209, May 1995.

21. Mak, B. (1997). “Combining ANNS to Improve Phone Recognition”, *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, Munich, April 1997.

22. Maneenoi E. and et al, (1997). “Modification of BP Algorithm for Thai Speech Recognition”, *Proceedings of the Natural Language Processing Pacific Rim Symposium 1997*, pp.287-291.

23. Pfitzinger, H., Burger, S. and Heid, S. (1996). “Syllable Detection in Read and Spontaneous Speech”, *Proceedings of the International Conference on Spoken Language Processing*, Philadelphia, PA, Oct., 1996.

24. Rabiner, L. and Juang, B. (1993). *"Fundamentals of Speech Recognition"*, Englewood Cliffs, NJ:Prentice Hall, 1993

25. Shire, M. (1997). "Deployment of RASTA-PLP with the Siemens ZT Speech Recognition System", *ICSI Technical Report*, tr-97-057, Berkeley, California, December 1997.

26. Tebelskis, J. (1995). "Speech Recognition using Neural Networks", Ph.D. Thesis, School of Computer Science, Carnegie Mellon University.

27. Thubthong, N. (1995). "A Thai Tone Recognition System Based on Phonemic Distinctive Features", *Proceedings of SNLP'95 The 2nd Symposium on Natural Language Processing*, pp.379-386.

28. van Vuuren, S. (1996). "Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch", *Proceedings of the International Conference on Spoken Language Processing, Philadelphia*, PA, October 1996.

29. Vermeulen, P., Barnard, E. et al. (1995). "A Comparison of HMM and Neural Network Approaches to Real World Telephone Speech Applications", *Proceedings of the International Conference on Neural Networks and Signal Processing*, Nanjing, P.R. China, December, 1995.

30. Yan, Y. Fanty M. and Cole, R. (1997). "Speech Recognition Using Neural Networks with Forward-backward Probability Generated Targets." *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, Munich, April 1997.