

論文 / 著書情報
Article / Book Information

Title	Feature Analysis Based on Articulatory and Perceptual Models
Author	SADAOKI FURUI
Journal/Book name	Proc. 1993 IEEE Workshop on Automatic Speech Recognition, Vol. , No. , pp. 63-64
発行日 / Issue date	1993, 12
権利情報 / Copyright	(c)1993 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Feature Analysis Based on Articulatory and Perceptual Models

Sadaoki Furui

NTT Human Interface Laboratories
3-9-11, Midori-cho, Musashino-shi, Tokyo, 180 Japan

Speech research is fundamentally and intrinsically supported by a wide range of sciences. The intensification of speech research continues to underscore an even greater interrelationship between scientific and technological interests [1]. Although it is not always necessary or efficient for speech recognition systems to directly imitate human speech production and perception mechanisms, it will become more important in the near future to build mathematical models based on these mechanisms to improve performance [2,3].

For example, when sequences of phonemes and syllables are produced by human articulatory organs, such as the tongue, jaw, and lips, these organs move in parallel and asynchronously, but still systematically. Present speech analysis methods, however, convert speech signals into a single sequence of instantaneous spectra. It will become important to decompose speech signals or spectra into multiple/parallel/overlapping sources based on concealed production mechanisms [4]. This approach seems to be essential for solving the coarticulation problem, one of the most important problems in speech recognition.

Psychological and physiological research into human speech perception mechanisms shows that the human hearing organs are highly sensitive to changes in sounds, that is, to transitional (dynamic) sounds, and that the transitional features of the speech spectrum and the speech wave play crucially important roles in phoneme perception [5]. It can even be said that only changes bear information. The length of the time windows in which transitions of sounds are perceived have a hierarchical structure ranging from the order of several milliseconds to several seconds. The hierarchical layers correspond to various speech features, such as phonemes, syllables, and prosodic features.

The representation of the dynamic characteristics of speech waves and spectra has been studied, and several useful methods have been proposed. These methods include transitional cepstra [6], dynamics-emphasized cepstra [7], hierarchical transitional cepstra [8], RASTA [9], cepstral blind deconvolution [10], and neural prediction models [11]. The discovery of good methods for representing dynamics of speech associated with various time lengths will have a substantial impact on the course of speech recognition research. This research is closely related to the analysis method based on the speech production mechanism described above.

Studies on the auditory compensation mechanism, in which coarticulated acoustic patterns are inversely transformed into the ideal sequence of categorical phonemic information based on modeling of the coarticulation mechanisms, have been performed using various approaches. A mathematical model of the auditory system, which predicts the spectral target value based on the spectral transitional information extracted over a period of around 50 ms was proposed, and experimental results indicated that the target was reliably predicted and categorically changed at the maximum spectral transition position [12].

Robustness is one of the most important issues in speech recognition [13,14]. The human hearing system is far more robust than machine systems - more robust not only against the direct influence of additive noise but also against speech variations (that is, the indirect influence of noise), even if the noise is very inconsistent. Speech recognizers should therefore become more robust when the front end is based on models of human hearing. This can be done by imitating the physiological organs [15] or by reproducing psychoacoustic characteristics [16]. It was confirmed that transitional representations of spectra are robust against additive noise and speech variations due to microphone changes.

Recognition performance under noisy conditions is often impaired by variations in the amount of speech quality degradation rather than by the degradation itself. Problems are created, for example, by the variation in noise level associated with variations in the distance between the speaker and the microphone. To cope with these variations, it is essential to develop methods for automatically adapting to and normalizing these effects. The human hearing system is highly adaptable to various variations. A spectral mapping method based on hierarchical spectral clustering, which was originally proposed for speaker adaptation, has recently been applied to the spectral normalization of speech distortion by various additive

noises, and its usefulness has been confirmed [17].

Basic speech units for speech recognition should be studied from the following perspectives: (1) Linguistic units (e.g., phonemes and syllables), (2) Articulatory units (e.g., positions and motion targets for the jaw and tongue), (3) Perceptual units (e.g., targets and loci of spectral movement, and distinctive features), (4) Visual units (features used in spectrogram reading), and (5) Physical units (e.g., centroids in vector/segment quantization). It will be important to establish new units based on combinations of these viewpoints [18].

When recognizing spontaneous speech in dialogs, it is necessary to deal with variations that are not encountered when recognizing speech that is read from text. These variations include extraneous words, out-of-vocabulary words, ungrammatical sentences, botched utterances, restarts, repetitions, and style shifts. It is crucially important to develop robust and flexible parsing algorithms that simulate the characteristics of human beings.

References

- [1] J. L. Flanagan: "Overview of speech communication," Proc. NAS Colloquium on Human-Machine Communication by Voice, Session I (1993).
- [2] S. Furui: "Digital speech processing, synthesis, and recognition," Marcel Dekker, Inc., New York (1989).
- [3] B. S. Atal: "Speech technology in 2001: new research directions," Proc. NAS Colloquium, Session VIII (1993).
- [4] B. S. Atal: "Efficient coding of LPC parameters by temporal decomposition," Proc. ICASSP, Boston, 2.6 (1983).
- [5] S. Furui: "On the role of spectral transition for speech perception," J. Acoust. Soc. Am., 80, pp. 1016-1025 (1986).
- [6] S. Furui: "Speaker-independent isolated word recognition using dynamic features of speech spectrum," IEEE Trans. ASSP-34, pp. 52-59 (1986).
- [7] S. Furui: "Speaker-independent isolated word recognition based on emphasized spectral dynamics," Proc. ICASSP, Tokyo, 14.17 (1986).
- [8] S. Furui: "On the use of hierarchical spectral dynamics in speech recognition," Proc. ICASSP, Albuquerque, S15a.10 (1990).
- [9] H. Hermansky et al.: "Recognition of speech in additive and convolutional noise based on RASTA spectral processing," Proc. ICASSP, Minneapolis, II-83-86 (1993).
- [10] B. S. Atal: "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," J. Acoust. Soc. Am., 55, pp. 1304-1312 (1974).
- [11] K. Iso and T. Watanabe: "Large vocabulary speech recognition using neural prediction model," Proc. ICASSP, Toronto, S2.5 (1991).
- [12] S. Furui and M. Akagi: "On the role of spectral transition in phoneme perception and its modeling," Proc. 12th Int. Cong. Acoust., A2-6 (1986).
- [13] B. H. Juang: "Speech recognition in adverse environments," Computer Speech and Language, 5, pp. 275-294 (1991).
- [14] S. Furui: "Toward robust speech recognition under adverse conditions," ESCA Workshop on Speech Processing in Adverse Conditions, Cannes, pp. 31-42 (1992).
- [15] O. Ghitza: "Auditory nerve representation as a basis for speech processing," in Advances in Speech Signal Processing, Marcel Dekker, New York, pp. 453-485 (1992).
- [16] H. Hermansky: "Perceptual linear predictive (PLP) analysis of speech," J. Acoust. Soc. Am., 87, pp. 1738-1752 (1990).
- [17] H. M. Cung and Y. Normandin: "Noise adaptation algorithms for robust speech recognition," Speech Communication, 12, pp. 267-276 (1993).
- [18] S. Furui: "Towards the ultimate synthesis/recognition system," Proc. NAS Colloquium, Session VIII (1993).