

論文 / 著書情報
Article / Book Information

Title	Text-Independent Speaker Adaptation Based on Hierarchical Spectral Clustering
Author	SADAOKI FURUI
Journal/Book name	Proc. IEEE Workshop on Speech Recognition, Vol. , No. , pp.
発行日 / Issue date	1988,
権利情報 / Copyright	(c)1988 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.

TEXT-INDEPENDENT SPEAKER ADAPTATION BASED ON HIERARCHICAL SPECTRAL CLUSTERING

SADAOKI FURUI

NTT Human Interface Laboratories
Musashino-shi, Tokyo, 180 Japan

This paper proposes a new automatic speaker adaptation algorithm in which only a small amount of training material of unspecified text is used. The algorithm, which is based on an adaptation method for a segment vocoder [1], is easily applicable to a SPLIT word recognition system based on the combination of vector quantization (VQ) and dynamic time-warping techniques. In this adaptation algorithm, only a VQ codebook is adapted for each new speaker, whereas a word dictionary, which represents each word as a sequence of codebook entries, is universally used for all speakers. The important point of this algorithm is that a set of training frame spectra and the codebook entries are clustered hierarchically. Based on the distances between centroids of the training frame clusters and the corresponding codebook clusters, adaptation is performed hierarchically from small to large numbers of clusters. The spectral resolution of the adaptation process is accordingly improved.

The first step in the process involves setting the number of clusters, j , at 1, with the centroid for the training frames and for the codebook entries being calculated. All codebook elements are then shifted so that their centroid coincides with the centroid of the training frame cluster based on the distance between these centroids. Next, j is doubled, all training frames are split into j clusters, and the corresponding centroids are calculated. These centroids are subsequently used to cluster the codebook entries, with the centroids of the resultant clusters being computed accordingly. The distances between the centroids of the training frames and their corresponding codebook centroids are then calculated. The codebook entries are shifted again based on these distances, while the continuity between adjacent clusters is maintained. This process is repeated, by doubling the number of clusters each time and creating a hierarchical spectral mapping between the standard spectral space and the new speaker's spectral space till the number of clusters reaches a fixed number, $j_{\max}$.

Recognition experiments were performed using utterances of 100 Japanese city names represented by sequences of a 10 dimensional LPC cepstral vector every 16 ms. An initial 1024 codebook entries were obtained from the utterances of four male speakers representing the range of speaker variation. Four word templates were constructed from the utterances of these four speakers and stored for each word. The utterances of 20 different male speakers were used for testing. Ten arbitrary words recorded by each individual test speaker one week before the test were used for adaptation.

The experimental results indicated that the spectral distortions between the test utterances and reference templates decrease with the progress of the adaptation process. Recognition experiments showed that the mean word recognition error rate is reduced from 4.9 % to 3.1 % by adaptation. The error rate for speaker-dependent recognition was 2.2 %. These results clearly confirm that the adaptation method proposed in this paper is highly effective. Furthermore, we believe that this method is general enough to be equally applicable to other speech recognition systems.

[1] Y. Shiraki and M. Honda: "Speaker adaptation algorithms for segment vocoder", Trans. Committee of Speech Res., Acoust. Soc. Jap., SP87-67 (1987)