Study on in influence of dynamic features on speaker identification

Tsuyoshi Izumida (1010003)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 6, 2011

Keywords: Speaker identification, Three-layer model, Dynamic feature, Hearing impression, Adjective.

Speech comunication is a priamry and basic human business. However, a question that "how human perceive linguistic information ?" has not been solved yet. A factor that has not been able to get elucidation, is differences among speakers (speaker individuality). Acoustic features are different among speakers even in the same utterances. However, human can extract the same linguistic information, even though different speakers. Thus, human pick up linguistic information by normalizing or adapting speaker individualities. On the other hand, human can perceive who speaks by using speaker individuality. However, a question that "how human perceive who speaks ?" has not been solved yet. The two questions that "how human perceive linguistic information ?" and "how human perceive who speaks ?" are basic problems on speech science. In order to study the questions, it is necessary to study what acoustic features in speech become cue of speaker individuality.

Previous studies on speaker identification reported that a variety of acoustical features contribute to perception of speaker individuality. Features in these studies are categorized into two groups, that is, averaged amount (static features) and varied amount (dynamic features). However, it is difficult to say in current research that relationships between speaker identification and dynamic features have been investigated enough. The dynamic features are derived from movements of speech organs. Acoustic features related to the movements also vary each other. Thus, it is necessary to consider combinations of several acoustic features to investigate the relationships between speaker identification and dynamic features. Focusing on hearing impressions of speech such as voice quality and speaking style, is beneficial to integrate several acoustic features. Hearing impressions were described using adjectives. For example, relationships between perception

and acoustic features on non-linguistic areas such as emotional speech and singing voice were modeled using three-layer models. This paper discussed about relationships between speaker identification and dynamic features using a three-layered model, in which relationships between speaker identification (first layer) and hearing impression (second layer), and the second layer and acoustic features (third layer) are constructed from top to bottom. Furthermore, influences on speaker identification in the first layer from varied acoustic features in the third layer are evaluated from bottom to top. This paper report these results.

Relationships between speaker identification (first layer) and hearing impressions (second layer) are obtained by taking the following two steps. First, a perceptual space for speakers is estimated from similarity measurements of speakers' characteristics using the multi dimensional scaling. Next, degrees of speaker impressions are estimated by the Semantic Differential test (SD test). The results show taht, "brisk" is a major factor in hearing impression of speaker identification. The relationships between hearing impression (second layer) and acousitic features (third layer) are found out by the correlation analysis between the acoustic features and the degrees of hearing impression. Extracted acoustic features are fundamental frequency (F_0), power, spectra, durations. Results of correlation analysis show that, average, maximum and slope of F_0 are correlated with the degrees of "brisk." In addition, maximum and dynamic range of spectral tilts were correlated with "brisk." Slope of F_0 and dynamic range of spectral tilts are amount of dynamic features. Therefore, "brisk" is a hearing impression of speaker identification, correlating with dyanamic features.

A three-layer model corresponding to "brisk" constructed by analysis is evaluated from bottom to top. Stimuli are synthesized controlling of phased degrees of "brisk" by controlling slope of F_0 and dynamic range of spectral tilt to evaluate the model. First, it is checked that the stimuli control the hearing impression in second layer. Next, influence on speaker identification by varying degrees of "brisk," was evaluated. The results show that, varying acoustic features for "brisk" affected speaker identification. Thus, amount of dynamic features affect speaker identification. Additionally, it is suggested that degrees of hearing impressions affect speaker identification. Methods and findings on this study are though of as leading to the elucidation of the major questions that "how human perceive linguistic information ?" and "how human perceive who speaks ?"