ICASSP 2003, Hong Kong

AN INFORMATION-THEORETIC ANALYSIS OF DEVELOPMENTAL CHANGES IN SPEECH[1]

Serdar Yildirim, Shrikanth S. Narayanan

Speech Analysis and Interpretation Laboratory, http://sail.usc.edu
Department of Electrical Engineering and Integrated Media Systems Center
University of Southern California, Los Angeles, CA 90089, USA

[yildirim, shri]@sipi.usc.edu

ABSTRACT

Developmental changes in the human speech production system signal age-dependent variability in the speech signal properties. In this paper, an information-theoretic analysis of developmental changes in the speech signal is presented. The effects of age and signal bandwidth on speech signal features are analyzed especially motivated by implications to automatic recognition of children's speech. Mutual information is calculated between cepstral features and the vowel phonetic class for different age groups and signal bandwidths. The results show that information contained in cepstral features about phonetic classes increases as bandwidth increases for all ages. Cepstral features of adult speech convey more information compared to that of children's speech for both genders. Information increases rapidly between bandwidths 500Hz and 4500Hz. These findings based on mutual information correspond well with vowel recognition (classification) experiments. The vowel recognition experiment shows that as bandwidth increases recognition accuracy increases as well.

1. INTRODUCTION

Most automatic speech recognition systems' (ASR) performances degrade when the training and testing conditions are not similar. One reason for the acoustic mismatch between training and testing data is speaker variability. Typically, female speech is different from male speech. So do children differ from adults. Previous research has shown that spectral and temporal variability in children's speech is greater than that of adult's speech [2]. This study has also shown that children's speech has higher pitch and formant frequencies, and longer segmental durations. This age dependence causes a serious degradation on ASR performance especially if, the ASR models are trained using speech from different age groups than those encountered during testing. Results have shown that in such cases the word error rates are typically two to five times worse for children speech than for adult speech. Vocal tract normalization based on frequency warping and model adaptation methods have been shown to improve recognition performance for children's speech [1]. However, these results show that there is still age-dependent performance variability even under matched or normalized model conditions. It should be noted that there are also other reasons for degraded performance of children's ASR. For example, a recent study by Li [6] shows that the child's speaking proficiency effects recognizer performance. The error rates four times worse for children speech judged to be good than for children speech judged to be poor. This study however focuses on the basic acoustic variability problem related to developmental changes.

Previous studies have also shown that the effect of signal bandwidth on recognition performance is significant especially in dealing with children's speech [4,6]. Recognition performance degrades rapidly as bandwidth is reduced to less than 6kHz [4]. The basic underlying reason here is that, for a given bandwidth, the amount of spectral resonance information available for ASR in a child's speech is much less than for an adult due to the shorter vocal tract length. This paper attempts to quantify, and explain, this developmental effect using a simple information analysis.

Mel frequency cepstral coefficients (MFCCs) are widely used in speech recognition systems. These coefficients are computed from a window of speech segment in overlapping short time intervals. This paper uses cepstral features as a means to study acoustic variability between different age groups. Specifically, we present a measure based on mutual information to determine effects of age and bandwidth on cepstral features. This can be achieved by comparing relative information between cepstral features and phonetic class units for different ages and bandwidths. In this paper, we also analyze the effects of bandwidth of the signal on recognizer performance across different children's ages.

[1] Research supported in part by NSF ERC, Cooperative agreement No. EEC-9529152 and a USC Zumberge Interdisciplinary Research grant

The rest of the paper is organized as follows. In Section 2, the speech data corpus used in experiments is explained briefly. The basic idea behind mutual information method is given in Section 3. Experiments and results are presented in Section 4 and Section 5 respectively. Finally, Section 6 provides our conclusions.

2. SPEECH DATA CORPUS

The speech data (16kHz, NIST) used in the experiments were obtained from 436 children (ages 5-18) resolution of 1 year of age and from 56 (ages 25-50) adult speakers [3]. The distribution of male and female speakers is 258 and 234 respectively. The database contains ten monophthongal and five diphthogonal vowels and five phonetically rich meaningful sentences. To enable direct comparisons with the analyses in [1], only the ten monophthongal vowels were analyzed in this work. To obtain the acoustic features, ten monophthongal vowels from bead (/IY/), bit (/IH/), bet (/EH/), bat (AE/), pot (/AA/), ball (/AO/), but (/AH/), put (/UH/), boot (/UW/), and bird (/ER/) target words were segmented using label files which contain phonetic segments labeled by HMMs. Each gender and age cepstral features were calculated from phonetic segments for the various bandwidths (obtained by appropriate signal downsampling).

3. MUTUAL INFORMATION: FEATURES & PHONETIC UNITS

Mutual information is a measure of the amount of information that one random variable contains about another random variable, or how much information one random variable tells us about another one [5]. Given phonetic class units as samples of a discrete-valued random variable C, and cepstral feature vectors as samples of a continuous-valued random variable Z, mutual information between C and Z is

$$I(C,Z) = H(C) - H(C \mid Z). \tag{1}$$

It is basically a reduction in uncertainty about random variable C due to knowing of Z.
The entropy of phonetic class C, H(C), is defined by

$$H(C) = -\sum_{c} p(c)\log(p(c)) \,. \tag{2}$$

Where p(c) are the prior probabilities of random variable C. After having observed the acoustic feature vector Z, the entropy of phonetic class C is defined by

$$H(C \mid Z) = -\int_{z} p(z)\left(\sum_{c} p(c \mid z)\log(p(c \mid z))\right)dz. \tag{3}$$

After applying the Bayes rule,

$$p(c,z) = p(c \mid z)\,p(z). \tag{4}$$

and the identity

$$\mathrm{p(c)} = \int_{z} \mathrm{p(c,\ z)dz}\,. \tag{5}$$

The mutual information between phonetic class units and cepstral feature vectors, I(C, Z), can be calculated as

$$\begin{aligned} I(C,Z) &= H(C) - H(C \mid Z) \\ &= \sum_{c} p(c)\int_{z} p(z \mid c)\log\frac{p(z \mid c)}{p(z)}dz. \end{aligned} \tag{6}$$

After vector quantization of Z, (6) can be rewritten as given in [7]

$$I(C,Z) = \sum_{c} p(c)\sum_{z_j} p(z_j \mid c)\log\frac{p(z_j \mid c)}{p(z_j)}. \tag{7}$$

4. EXPERIMENTS

First, 13 dimensional cepstral observation vectors were calculated from phonetic segments for each gender and age groups for various bandwidths by using 25msec Hamming window every 10msec time interval. The total number of feature vectors from a phonetic class per age was limited to 4000. In order to apply Eq. 7, we need to estimate the required probability density functions of the variables involved. After quantizing the feature vectors, we employed a histogram-based method to estimate these functions.

Lets assume we have D_c samples for each phonetic class c, c $\in \{1, 2, , N_c\}$, and then class prior probabilities are

$$p(c) = \frac{D_c}{N}. \text{ , } N = \sum_{c=1}^{N_c} D_c \tag{8}$$

After k-means clustering the N total samples to J codewords, the prior probabilities of quantized feature vectors are calculated as,

$$p(z_j) = \frac{J_j}{N}. \text{, } j \in \{1,2,\ldots,J\} \tag{9}$$

where J_j is the total number of feature vectors in that cluster belongs to same class as $z_{j.}$ Probabilities of quantized feature vectors when class is given are calculated as

$$p(z_j \mid c) = \frac{J_{j|c}}{J_j}. \tag{10}$$

where $J_{j|c}$ is the number of feature vectors belongs to class c in cluster j.

After estimating required probabilities, Eq. (7) was used to compute the mutual information between the phoneme class and the quantized feature vectors for different ages and bandwidths. In this study, 10 vowels, /IY/, /IH/, /EH/, /AE/, /AA/, /AO/, /AH/, /UH/, /UW/, and /ER/ are considered as elements of phonetic class. The entropy of class distribution is 3 bits, and the mutual information associated with the cepstral coefficients for different ages and bandwidths are given in the results section.

In order to determine effects of bandwidth reduction on an automatic recognition system, we created an HMM-based recognizer using the HTK 3.0 toolkit [8]. Each HMM model had 3 states and 5 Gaussian mixtures. 39 dimensional feature vectors (13 MFCC's, 13 delta, and 13 delta-delta) were calculated by using 25msec Hamming window every 10msec time interval. The original 16kHz speech signal was decimated to 500Hz, 1kHz, 2kHz, 3kHz, 4kHz, 6kHz, and 8kHz, and then the HMM training and testing experiments were repeated for each bandwidth. Due to limited amount of data for each age, a "leave-one-out" strategy was used. One speaker data was left out for testing and others were used for training. This procedure was repeated for each speaker.

5. RESULTS

Information contained in acoustic features about phonetic class for different ages and bandwidths is given in Figure 1. and Figure 2. for male and female speakers, respectively. It can be seen from the figures that information increases as bandwidth increases for all ages. It can easily be observed that cepstral features of adult speech convey more information compared to that of children's speech for both genders. It should also be noted that between bandwidths 1500 and 3500 kHz, information difference between adult and lower age groups of male speakers are more significant compared to that of female speakers. Even there is an almost linear information increment as bandwidth increases across ages; Information contained in children speech cepstral features never reaches to adult level for both genders.

When results are examined with respect to ages, it is easily observed that cepstral features from adult speech convey much more information than that of children speech, ages 5 to 13 years, for both male and female speakers. Even though, there is significant information difference between children's and adult's speech cepstral features, one can not conclude that there is a linear increment between ages.

The change in mutual information averaged across all ages as a function of bandwidth is given in Figure 3 for male and female speakers. In Figure 3, it can be seen that the information increases almost exponentially between 500Hz and 4500 Hz, increment is relatively small for bandwidths above 4500Hz. It is interesting to note that; information values from female speakers and from male speakers averaged across all ages for any given bandwidth are almost the same.

The vowel recognition results are given in Figure 4., and Figure 5., for male and female speakers respectively. It can be seen from the figures that accuracy rates increase as bandwidth increases. This can be explained by the (phonetic class dependent) information increase contained in the cepstral features as bandwidth increases. The accuracy rates degrade more rapidly for small ages compared to ages 16 and above, for both genders, as bandwidth decreases. It can also be concluded from these figures that recognizers that are trained with children speech are much more sensitive to bandwidth changes compared to recognizers that are trained with adult's speech.

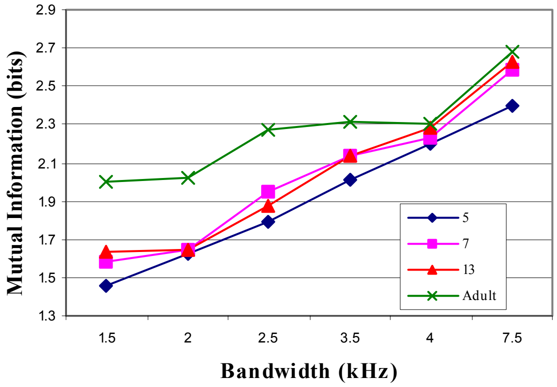

Figure 1. Information changes for different ages (years) with respect to different bandwidths (kHz) for male speakers.

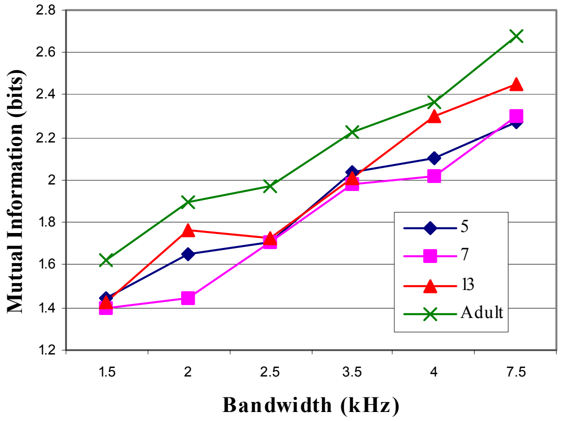

Figure 2. Information changes for different ages (years) with respect to different bandwidths (kHz) for female speakers.

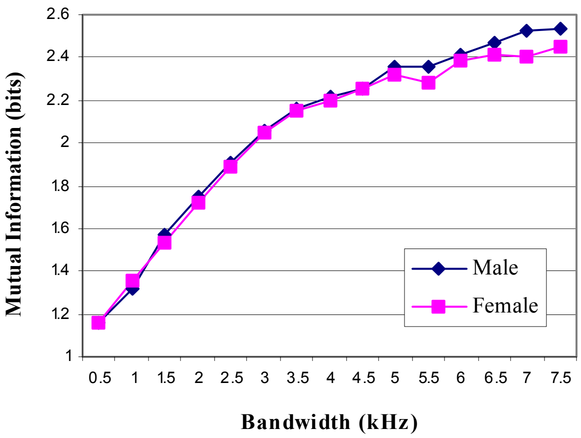

Figure 3. Information changes averaged across all ages for both male and female speakers

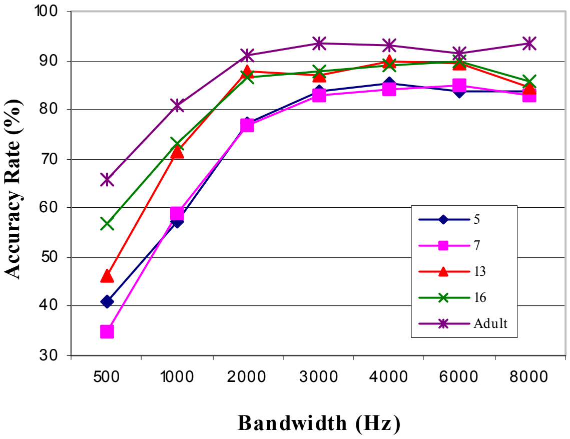

Figure 4. Vowel recognition accuracy results for male speakers.

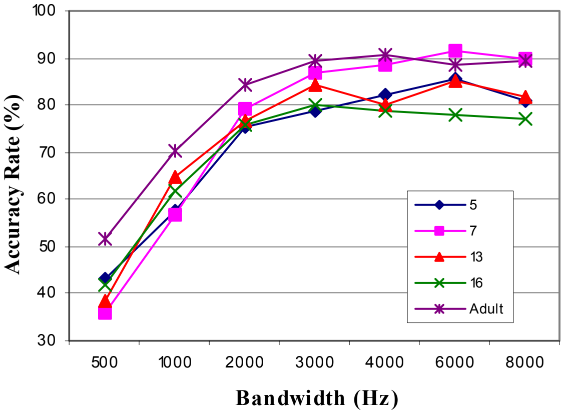

Figure 5. Vowel recognition accuracy results for female speakers.

6. CONCLUSIONS

In this study, we explored the effects of age and bandwidth changes on cepstral features. We employed mutual information based method to determine age and bandwidth effects on cepstral features. Our results clearly showed that as bandwidth increases information in the cepstral features about the vowel class increases. This result explains the increase in vowel recognition accuracy as bandwidth increases.

REFERENCES

[1] S. Narayanan and A. Potamianos, ``Creating conversational interfaces for children," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 2, pp. 65-78, 2002.

[2] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," J. Acoust. Soc. Am., vol. 105, pp. 1455-1468, Mar. 1999.

[3] Miller, J. D., Lee, S., Uchanski, R. M., Heidbreder, A. H., Richman, B. B.,and Tadlock, J., "Creation of two children's speech database," in Proc. of ICASSP, (Atlanta, GA), 1996.

[4] Li, Quan and Russell, Martin J., "Why is Automatic Recognition of Children's Speech Difficult?", Eurospeech 2001, Scandinavia.

[5] T.M. Cover and J. A. Thomas, "Elements of Information Theory,"(Wiley Series in Communications), John Wiley & Sons Inc., New York, 1991.

[6] Li, Quan and Russell, Martin, "An Analysis of the Causes of Increased Error rates in Children's speech Recognition," in Proc. of ICSLP, (Denver, CO), 2002.

[7] Padmanabhan, M., "Use of Spectral Peak Information in Speech Recognition", Speech Transcription Workshop, NIST, University of Maryland, 2000.

[8] Young S. J, Odell J., Ollason D., Valtchev V., Woodland P., "HTK Book" Cambridge Research Laboratory, 1997.