ACOUSTIC –PHONETIC SPEECH PARAMETERS FOR SPEAKER-INDEPENDENT SPEECH RECOGNITION

Om Deshmukh, Carol Y. Espy-Wilson and Amit Juneja
{omdesh,espy,juneja}@Glue.umd.edu
http://www.ece.umd.edu/~omdesh/ICASSP2002/
Speech Communication Lab, ECE Dept, University of Maryland, College Park

Introduction

We discuss a method that compensates the effect of inter-speaker variability in Automatic Speech Recognizers. In particular, we demonstrate that when a Hidden Markov Model (HMM) based system is used as the back-end, a Knowledge-Based Front End (KBFE) can outperform the traditional Mel-Frequency Cepstral Coefficients (MFCCs), particularly when there is a mismatch in the gender and ages of the subjects used to train and test the recognizer.

Methodology

Front End:

1.KBFE: The acoustic parameters (APs) that make up KBFE are acoustic correlates of the phonetic features proposed by linguists [1]. Thus, they are designed to target only the linguistic information in the speech signal. Theoretically, there are 20-odd phonetic features that describe all of the languages of the world. At present, we have developed APs for 13 of the phonetic features. In particular, we have APs for the manner phonetic features: *sonorant, syllabic, continuant, voiced* and *strident*, in addition to silence. In addition, we have APs for the place phonetic features: *anterior* for strident fricatives, *alveolar* and *labial* for stop consonants, *high*, *low*, *front* and *back* for vowels, and *rhotic* for /r/.

In previous work [1], we showed the importance of using relative measures to reduce the speaker dependency of the APs. These relative measures are obtained by normalizing the APs over time and/or over frequency. An example of a frequency-normalized parameters is our measure for characterizing the spectral shape of strident fricatives. In this case, we base the frequency bands on the third formant (F3) estimated from each utterance. An example of a time-normalized parameters is our onset/offset parameter which is the sum of the first differences computed across the short-time Fourier transform channels.

2. ***MFCCs***: The standard front end consisted of 13 MFCCs along with their first and second derivatives, normalized to zero mean. The MFCCs were computed every 5ms.

Back End:

HTK [2], a HMM-based recognition system, was used for the back end processing. 10 different HMM models, one for each digit, were developed. Each of the digits and the E-set phonemes was modeled as a three state HMM with eight Gaussian mixtures per state, each of which was initialized as zero mean and unit diagonal covariance. All the mixtures weights were initialized at the same value. Left-to-right state transition with no skips was incorporated with the additional constraint that each model has to start with the first state. All of the allowable transitions were initialized as equi-probable.

Databases:

Recognition experiments were run on the highly confusable E-set (/b/,/c/,/đ/,/e/,/g/,/p/,/t/,/v/,/z/) from the TI46 [3] (adult speech), isolated digits from the TI-46 (adult speech) and isolated digits from the TIDIGITS [3] (children speech). Boys were between the ages of 6-14 and girls were between the ages of 7-15. The TIDIGITS training data consisted of 102 utterances of each digit, 2 repetitions by each of the 25 boys and 26 girls. The TIDIGITS test data consisted of 100 utterances of each digit, 2 repetitions by each of a different set of 25 boys and 25 girls. The TI46 training data consisted of 160 productions of each alphabet and digit, with 10 repetitions by 8 males and 8 females. The TI46 test data consisted of 256 utterances of each alphabet and digit, with 16 repetitions by the same 8 males and 8 females. .

Results

We performed a set of experiments where we trained and tested on similar data, and where we trained and tested on data that differed in terms of the gender and age group that produced it.

Tables 1 shows the overall phoneme accuracy results (in %accuracy) for the E-set data whereas table 2-4 show the overall word accuracy for the digit dataset. 'Tr' specifies the training data & 'Te' specifies the test data.

Table 1. Results for gender variability in adults (E-set data)

	Tr: Adult Te: Adult	Tr: Males Te: Females	Tr: Females Te: Males
MFCC (39 pars)	82.45	67.14	71.23
KBFE (21 pars)	85.14	79.96	82.47

Table 2. Results for gender variability in adults (digits data)

	Tr: Adult Te: Adult	Tr: males Te: females	Tr: Females Te: Males
MFCCs (39 pars)	99.88	70.27	68.29
KBFE (28 pars)	97.26	91.67	78.33

Table 3. Results for gender variability in children (digits data)

	Tr: children Te: children	Tr: boys Te: girls	Tr: girls Te: boys
MFCCs (39 pars)	98.30	95.60	95.00
KBFE (28 pars)	94.71	92.33	96.56

Table 4. Results for age variability (digits data)

	Tr:adult Te:adult	Tr:child Te:child	Tr:adult Te:child	Tr:child Te:adult
MFCCs (39 pars)	99.88	98.30	60.20	62.37
KBFE (28 pars)	97.26	94.71	85.15	85.88

When the training and test data are similar, the results using KBFE are better than those for MFCCs for the E-set task (column 1 in Table 1), but not as good as the results obtained with MFCCs for the digit task(column 1 in Table 2 & 3). This difference in performance is due to the fact that fewer phonetic features are needed to distinguish among the sounds that make up the E-set. In particular, we are missing some key APs for the digit task like those that extract information relevant for nasals. Thus, KBFE is not yet a full linguistic representation. It consists of only 21 parameters for only 13 of the 20-odd phonetic features in the case of the E-set task. Additional parameters were added for the digit task to capture information about the formant frequencies. However, the MFCCs are a full representation of the speech signal.

The more important results are those obtained when the recognizer is tested on data different from the data it was trained with (columns 2 & 3 in Tables 1, 2 & 3, and columns 3 & 4 in Table 4). Relative to the results obtained with the MFCCs, the results with KBFE show an error reduction of about 39% when there is a gender difference between the subjects used to train and test the system (Table 1, 2 & 3). Similarly, KBFE shows an error reduction of about 63% when there is a considerable age difference between the subjects used to train and test the recognizer (Table 4). There is very little change in the results of Table 3. Such consistent results suggest that boys and girls between 7 and 15 years of age have very similar vocal tract lengths.

References

[1]Bitar, N. and Espy-Wilson, C., "The design of acoustic parameters for speaker-independent speech recognition", Eurospeech 97, Vol 3., 1239-1242

[2] http://htk.eng.cam.ac.uk/

[3] http://www.ldc.upenn.edu/

Acknowledgements

This work was supported by NSF grant #SBR-9729688 and NIH grant #1K02DC00149.