USE OF PERIODICITY AND JITTER AS SPEECH RECOGNITION FEATURES

David L. Thomson and Rathinavelu Chengalvarayan

Speech Processing Group
Bell Labs, Lucent Technologies
Naperville, Illinois 60566, USA
Email: davidt@lucent.com, rathi@lucent.com

ABSTRACT

We investigate a class of features related to voicing parameters that indicate whether the vocal chords are vibrating. Features describing voicing characteristics of speech signals are integrated with an existing 38-dimensional feature vector consisting of first and second order time derivatives of the frame energy and of the cepstral coefficients with their first and second derivatives. HMM-based connected digit recognition experiments comparing the traditional and extended feature sets show that voicing features and spectral information are complementary and that improved speech recognition performance is obtained by combining the two sources of information.

1. INTRODUCTION

Pitch and voicing are widely used in speech coding [3, 7] but not in speech recognition. Methods for making the voiced-unvoiced decision usually work in conjunction with pitch analysis. For speech recognition, voicing features are useful in distinguishing vowels from consonants and in distinguishing consonants such as /d/ and /t/ from each another. For example, one difference between plosives /b/ and /p/ is that voicing begins earlier in /b/. This observation suggests that prosodic information such as voicing may be useful in speech discrimination.

Voicing can be determined with reasonable accuracy from spectral coefficients, since unvoiced speech tends to contain stronger high frequency components than voiced speech. However, our experiments have not shown voicing features derived from spectral coefficients to improve error rates over using spectral coefficients alone. The voicing features described in this study contain information not present in the spectrum, and are derived from the time signal. In this study, we describe two voicing parameters called periodicity and jitter. Periodicity is a measure of the periodic structure of speech. Jitter is the small fluctuations in glottal cycle lengths and has been studied recently by means of a statistical time series model [5].

Over the last several years, a major factor in reducing the error rate in speech recognition systems has been the addition of new feature components to the frame vectors. In this work, periodicity and jitter metrics are combined with a 38-dimensional feature vector consisting of first and second order time derivatives of the frame energy and of the cepstral coefficients with their first and second derivatives. Discriminative training is necessary because of the strong correlation between voicing and the first spectral coefficient. We report several connected digit recognition results comparing the traditional maximum likelihood (ML) method and the minimum string error (MSE) training method to study the effects of including voicing features in the signal representation. We have noted that the addition of voicing features makes the system more robust because these features are relatively insensitive to differences in transmission conditions.

2. INCORPORATION OF VOICING FEATURES

In this section, we describe two voicing parameters: periodicity and jitter. Both are derived from pitch analysis. There are variety of methods for pitch estimation of speech signals described in the literature [3]. The pitch estimation algorithm adopted in this study is based on the short-time autocorrelation function. Let X_n correspond to the rectangular windowed input speech sample. The short-time autocorrelation function is given by

$$R_i(m) = \frac{1}{N-m} \sum_{i=0}^{N-m-1} X_{n+i} X_{n+i+m},$$

where i is the index of the starting sample of the frame and N (corresponding to 30 msec) is the frame length. In general, female speech has higher pitch (120 to 200Hz) than the male speech (60 to 120 Hz). The range of delays considered spans the pitch period values most likely to occur in speech (20 to 120 samples, or 66Hz to 400Hz). The autocorrelation function is normalized with the peak at $m = 0$ so that the ratio lies between 0 and 1. The largest peak in the normalized function is chosen as the estimate of the pitch period and the value of the peak becomes the periodicity measure

$$Periodicity = \max_m \left\{ \frac{R_i(m)}{R_i(0)} \right\}, \quad 20 \le m \le 120 \quad (1)$$

This voicing function is a measure of how strongly periodic the speech frame is. It is often used to make a voiced/unvoiced decision by applying a threshold. For speech recognition, we treat it as an indicator of the probability that a given frame is voiced. Voicing is computed every 10 msec to match the frame rate of the speech recognizer.

Another voicing parameter useful in speech recognition is the variation in estimated pitch between frames. Whereas the pitch in voiced speech is relatively constant, the measured pitch of an unvoiced frame is essentially random, since most unvoiced speech consists of noise and other aperiodic signals. The change in pitch between frames, therefore, is an indicator of voicing. As a measure of change of pitch, we define a variation function

$$V_n \quad = \quad |P_n - P_{n-1}|\,,$$

where n is the index of the current frame and P is the measured pitch period for that frame.

One complication in measuring pitch variation is pitch multiplication and division. If the peak at the nth sample in the the autocorrelation function corresponds to the pitch period, there are usually also peaks at $k \times n$, where k is an integer. Peaks at $k \times n$ are sometimes larger than the peak at n, and can be chosen as the estimate of the pitch period. While this does not significantly affect the periodicity measure, it must be taken into account when estimating jitter. If the pitch period changes from n to $2 \times n$, for example, we should generally consider that the pitch variation is zero. We redefine the variation function to allow for pitch multiplication and division:

$$V_n \quad = \quad \min_{j,k} \left\{ \left| \frac{P_{n-1}}{j} - \frac{P_n}{k} \right| \right\},$$

where j and k are integers corresponding to the pitch multipliers for the previous and current frames, respectively. The range of values allowed for j and k are selected to minimize the expected variation function for voiced speech and maximize its expected value for unvoiced speech. A set of values that effectively separate voiced from unvoiced speech were determined experimentally to be

$$(j,k) \quad \in \quad \{(1,1),(1,2),(2,1),(3,1),(1,3)\}\,.$$

These values provide for pitch doubling and tripling. We also allow the pitch multiplier to change from double to triple and vice versa by permitting the following additional values:

$$(j,k) \quad \in \quad \begin{cases} (3,2) & \text{if } (j^\star,k^\star) = (1,3) \\ (2,3) & \text{if } (j^\star,k^\star) = (1,2) \end{cases}$$

where $j^\star$ and $k^\star$ are the values of j and k from the previous frame pair $n-1$ and $n-2$.

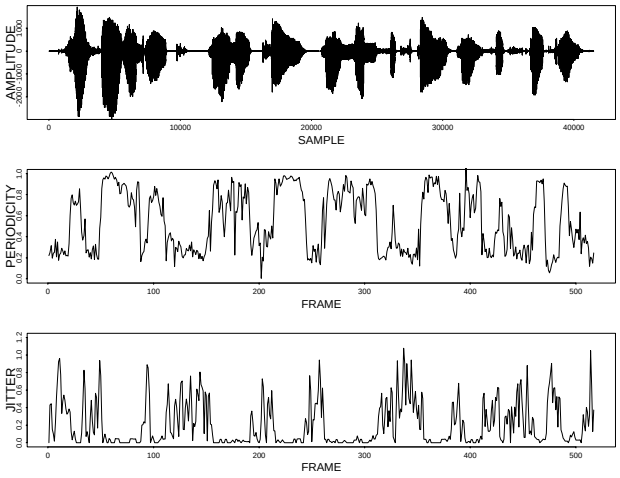

Figure 1. Typical voicing measurement contours for the utterance "341898291659603" spoken by a male speaker

More combinations are possible, but we limit the number because if too many are permitted, unvoiced speech is increasingly likely to yield a small value for the variation function. Once the variation function is computed between frame n and the two adjacent frames $n-1$ and $n+1$, we compute the jitter as the average of the two variation functions, normalized by the average pitch for the three frames

$$Jitter \quad = \quad \frac{\frac{1}{2}[V_n + V_{n+1}]}{\frac{1}{3}[P_{n-1} + P_n + P_{n+1}]}. \qquad (2)$$

Figure 1 illustrates the measured of jitter and periodicity for a typical digit string spoken by a female speaker. It is observed that the periodicity is about 1.0 and jitter is about zero for voiced speech. For unvoiced speech, periodicity is between zero and 0.5 and jitter is a random variable between about 0 and 1. (Silence is considered unvoiced.) Figure 1 suggests speech segments can be reliably classified as voiced or unvoiced based on periodicity and jitter measurements.

3. DISCRIMINATIVE MODEL PARAMETER ESTIMATION

We have used two methods for obtaining estimates of the HMM parameters namely the conventional maximum likelihood (ML) algorithm, and a more effective minimum string error (MSE) training procedure. For ML training, the segmental k-means training procedure was used [4]. The MSE training directly applies discriminative analysis techniques to string level acoustic model matching, thereby allowing minimum error rate training to be implemented at the string level [1]. A brief formulation of the MSE algorithm using generalized probabilistic descent (GPD) method is as follows:

- A discriminant function in MSE training is defined as

$$g(O, S_k, \Lambda) \quad = \quad \log \ f(O, \Theta_{S_k}, S_k|\Lambda),$$

where S_k is the k-th best string, Λ is the HMM set used in the N-best decoding, Θ_k is the optimal state sequence of the k-th string given the model set Λ, and $\log \ f(O, \Theta_{S_k}, S_k|\Lambda)$ is the related log-likelihood score on the optimal path of the k-th string.

- The misclassification measure is determined by

$$d(O, \Lambda) = -g(O, S_c, \Lambda) + \log \left(\frac{1}{N-1} \sum_{S_k \neq S_c} e^{g(O, S_k, \Lambda)} \right)$$

which provides an acoustic confusability measure between the correct and competing string models.

- The loss function is defined as

$$l(O, \Lambda) \quad = \quad \frac{1}{1 + e^{-\gamma d(O, \Lambda)}},$$

where γ is a positive constant, which controls the slope of the sigmoid function.

- The model parameters are updated sequentially according to the GPD algorithm

$$\Lambda_{n+1} \quad = \quad \Lambda_n - \epsilon \nabla l(O, \Lambda), \qquad (3)$$

Λ_n is the parameter set at the nth iteration, $\nabla l(O, \Lambda)$ is the gradient of the loss function for the training sample O which belongs to the correct class c, and ϵ is a small positive learning constant.

In this paper, we report only the results obtained by sequential training. During the model training phase, we call one complete pass through the training data set as an epoch. For the case of string-by-string training, model parameters are updated several times over an epoch.

4. DATABASES

This section describes the database, SST_CD, used in this study. This database is a good challenge for speech recognizers because of its diversity. It is a compilation of databases collected during several independent data collection efforts, field trials, and live service deployments. These independent databases are denoted as DB1 through DB6. The SST_CD database contains the English digits *one* through *nine*, *zero* and *oh*. It ranges in scope from one where talkers read prepared lists of digit strings to one where the customers actually use an recognition system to access information about their credit card accounts. The data were collected over network channels using a variety of telephone handsets. Digit string lengths range from 1 to 16 digits. The SST_CD database is divided into two sets: training and testing. The training set, DB1 through DB3, includes both *read* and *spontaneous* digit input from a variety of network channels, microphones and dialect regions.

Databases	Training		Testing	
	Strings	Speakers	Strings	Speakers
DB1	2568	500	2649	500
DB2	2075	2075	1036	518
DB3	2639	2639	713	713
DB4	–	–	3063	200
DB5	–	–	4318	50
DB6	–	–	1335	1281
Total	7282	5214	13114	3262

Table 1. Regional distributions of spoken digit strings and the speaker population among the training and testing sets of the SST_CD database.

The testing set is designed to have data strings from both matched and mismatched environmental conditions and includes all six databases. All recordings in the training and testing set are valid digit strings, totaling 7282 and 13114 strings for training and testing, respectively. The data distribution of the training and testing set is shown in Table 1.

5. FEATURE EXTRACTION

Input speech is segmented into overlapping frames 30 msec long with centers 10 msec apart. Each frame is processed to give 12 LPC-derived liftered cepstral coefficients along with energy and voicing features. Since the signal has been recorded under various telephone conditions and with different transducer equipment, each cepstral feature vector is further processed using the hierarchical signal bias removal (HSBR) method [2] to reduce the effect of channel distortion. The combined feature vector is augmented with its first and second order time derivatives resulting in two different feature dimensions as explained below.

To use a well-known frame vector as a baseline system, we perform our analysis on the 38-dimensional frame vector $DDCEP^+$ consisting of the cepstrum, delta cepstrum, delta-delta cepstrum, delta energy and delta-delta energy [8]. The $DDCEP^\star$ feature set has 44 components which includes $DDCEP^+$ combined with the voicing set and the delta and delta-delta derivatives of the voicing set. The voicing set includes periodicity and jitter, computed as show in (1) and (2).

6. REVIEW OF HMM CONNECTED DIGIT RECOGNIZER

Following feature analysis, each feature vector is passed to the recognizer which models each word in the vocabulary by a set of left-to-right continuous mixture density HMM using context-dependent head-body-tail models [8]. Each word in the vocabulary is divided into a head, a body, and a tail segment. To model inter-word coarticulation, each

Feature Vector Size and Type	ML training		MSE Training	
	Wd_Er	St_Er	Wd_Er	St_Er
38 $DDCEP^{+}$	3.31%	16.61%	2.14%	10.18%
44 $DDCEP^{\star}$	3.07%	15.78%	1.28%	6.42%

Table 2. Word error rate (Wd_Er) and string error rate (St_Er) for an unknown-length grammar-based connected digit recognition task using the conventional ML and MSE training methods as a function of frame vector size and type. The 44-feature vector with voicing is substantially more accurate.

word consists of one body with multiple heads and multiple tails depending on the preceding and following contexts. In this paper, we model all possible inter-word coarticulation, resulting in a total of 276 context-dependent sub-word models. Both the head and tail models are represented with 3 states, while the body models are represented with 4 states, each having 8 mixture components. Silence is modeled with a single state model having 32 mixture components. This configuration results in a total of 276 models, 837 states and 6720 mixture components.

Training included updating all the parameters of the model, namely, means, variances and mixture gains using ML estimation followed by six epochs of MSE to further refine the estimate of the parameters. The number of competing string models was set to four and the step length was set to one during the model training phase. The HSBR codebook of size four is extracted from the mean vectors of HMMs, and each training utterance is signal conditioned by applying HSBR prior to being used in MSE training. The length of the input digit strings are assumed to be unknown during both training and testing.

7. EXPERIMENTAL RESULTS

Several sets of experiments were run to evaluate the connected digit recognizers using two types of HMMs ($DDCEP^{+}$ and $DDCEP^{\star}$) and two types of training (ML and MSE). The overall performance of the recognizers, organized as the word and string error rate as a function of the feature vector size is summarized in Table 2.

Table 2 illustrates four important results. First, under all conditions, the MSE training is superior to the ML training; the MSE-based recognizer achieves an average of 50% string and word error rate reduction, uniformly across all types of speech models (both the baseline and extended feature set HMMs), over the ML-based recognizer. Second, for the ML-based recognizer, the $DDCEP^{\star}$ based HMM is slightly superior to the baseline HMM. Thirdly, for the MSE-based recognizer, superiority of the $DDCEP^{\star}$ based HMM over the $DDCEP^{+}$ based HMM becomes significantly greater than the ML case. Finally, the reduction in both string and word error rate in going from the ML to the MSE training with use of the $DDCEP^{\star}$ based HMM (about 60%) is higher than with the baseline HMM (about 40%). This difference tends to validate our conjecture that MSE training should be used with the extended feature set because of the strong correlation between voicing and the first spectral coefficient.

8. CONCLUSIONS

In this work, features representing the periodicity and jitter of speech signals are added to a standard 38-dimensional feature vector. Connected digit recognition results comparing the traditional maximum likelihood (ML) method and the minimum string error (MSE) training methods to study the effects of including voicing features are reported. We conclude that the difference in performance with and without voicing becomes more significant when MSE training is used than when ML training is used. The best result is achieved by including voicing features and by using the MSE training algorithm, yielding a string error rate reduction of 40%, compared to the MSE-trained baseline system. This suggests that prosodic information such as periodicity and jitter is useful in speech recognition.

REFERENCES

[1] B. H. Juang, W. Chou and C. H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Transactions on Speech and Audio Processing*, Vol. 5, No.3, pp. 257-265, 1997.

[2] W. Chou, M. Rahim and E. Buhrke, "Signal conditioned minimum error rate training," *Proc. EUROSPEECH*, pp. 495-498, 1995.

[3] D. P. Prezas, J. Picone and D. L. Thomson, "Fast and accurate pitch detection using pattern recognition and adaptive time-domain analysis," *Proc. ICASSP*, pp. 109-112, 1986.

[4] B. H. Juang and L. R. Rabiner, "The segmental k-means algorithm for estimating parameters of hidden Markov models," *IEEE Transactions on Acoustics Speech and Signal Processing*, Vol. 38, No. 9, pp. 1639-1641, 1990.

[5] J. Schoentgen and R. Guchteneere, "Predictable and random components of jitter," *Speech Communication*, Vol. 21, pp. 255-272, 1997.

[6] C. H. Lee, W. Chou, B. H. Juang, L. R. Rabiner and J. G. Wilpon, "Context-dependent acoustic modeling for connected digit recognition," *Proc. ASA*, 1993.

[7] B. S. Atal L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Transactions on Acoustics Speech and Signal Processing*, Vol. 24, No. 3, 1976.

[8] E. L. Bocchieri and J. G. Wilpon, "Discriminative feature selection for speech recognition," *Computer Speech and Language*, Vol. 7, pp. 229-246, 1993.