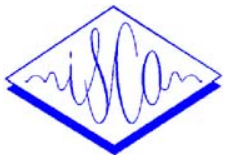

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

ROBUST ENERGY NORMALIZATION USING SPEECH/NONSPEECH DISCRIMINATOR FOR GERMAN CONNECTED DIGIT RECOGNITION

Rathinavelu Chengalvarayan

Speech Processing Group,
Lucent Speech Solutions Department
Lucent Technologies, Naperville, IL 60566, USA
Email: rathi@lucent.com

ABSTRACT

The addition of a word normalized energy contour uniformly improves performance of the HMM recognizer and makes it more robust to difference in talker populations. This kind of normalization generally requires some information on the statistics of energy features over the whole utterance, which is not a feasible solution in real-time applications due to the unnecessary long processing delay. In this paper, we propose a more efficient implementation approach for energy feature normalization where the normalization coefficients are computed using a look-a-head delay of fixed length. The experimental results on German connected digit recognition task show that a 12% string error rate reduction is obtained by using a look-a-head delay energy normalization scheme when compared to without using the energy feature. Further reduction of 10% string error rate is achieved by integrating the speech/nonspeech decision mechanism.

1. INTRODUCTION

It is generally agreed that the energy contour of an utterance contains important information about the phonetic identity of the sounds within the utterance. For example, fricatives are much lower in energy than vowels. Proper use of such energy information over time can therfore be helpful in grossly distinguishing one word utterance from another. Previous studies based on using the Itakura-Saito distortion measure, however, indicate that incorporation of energy inforamtion in pattern-comparison measures may in fact degrade the recognition performance if not properly handled. One possible modification is to normalize the gain terms by the maximum energy of a frame of speech in the utterance. Further results indicate that the energy information in the temporal pattern sequence can be useful for improving the recognition accuracy if it is properly normalized. Direct use of the absolute loudness level or gain term, however, leads to degradation in recognition accuracy [7].

Generally, the input feature to the recognizer used for recognition and modeling has been extended to include dynamic inforamtion about the first and second order derivatives of the cepstral features as well as the information about the cepstrum [4]. The combination of heterogenous parameters has also been found to be useful. Frame energy and its derivatives are often used as part of the representation for each frame. The addition of a word or sentence normalized energy contour (as an extra dimension to the feature vector) uniformly improves performance of HMM recognizer and makes it more robust to difference in talker populations and transmission conditions [3]. However, in those systems the normalization coefficients are typically computed over the whole utterance which is not a feasible solution in real-time applications due to the unnecessary long processing delay involved.

In this paper, we describe the two possible ways of normalizing the energy contour with minimal frame-delay and discuss the implementation details of the feature extraction process. We present recognition tests on two different databases, namely the German telephone digit string corpus and English cellular handsfree connected digit database. The test results show that the proposed methods for energy normalization in combination with the improved speech/nonspeech decision mechanism lead to consistent reductions in word and string error rates.

2. NORMALIZATION ALGORITHMS

In this section, we describe three different methodologies for robust energy normalization with special emphasis on speech/nonspeech discriminator that can classify speech sounds even in the presence of high noise. The short-time energy of a signal at t-th frame can be expressed as:

$$e(t) = \sum_{i=1}^{I} s_i^2(t)$$

where $s_i(t)$ refers to the amplitude of the i-th speech sample of t-th frame and e(t) is the energy of the signal over a frame duration of I samples. For computational reasons, it was found that instead of summing over $s_i^2(t)$, the summation can be done over the magnitude of the speech samples, that is

$$e(t) = \sum_{i=1}^{I} |s_i(t)|.$$

Frequently the value of e(t) is measured in db, that is 10 log_{10} $e(t)$ is used instead of e(t).

2.1. Batch Energy Normalization

The peak energy is initially determined for the whole utterance:

$$e\_max = \max\{e(t)\} \quad 1 \le t \le T$$

and each frame energy is normalized according to this peak energy:

$$ne(t) = e(t) - e\_max$$

We denote this normalization as EN_1. This scheme is not suitable for real-time application since it introduces a long delay in determining the peak energy.

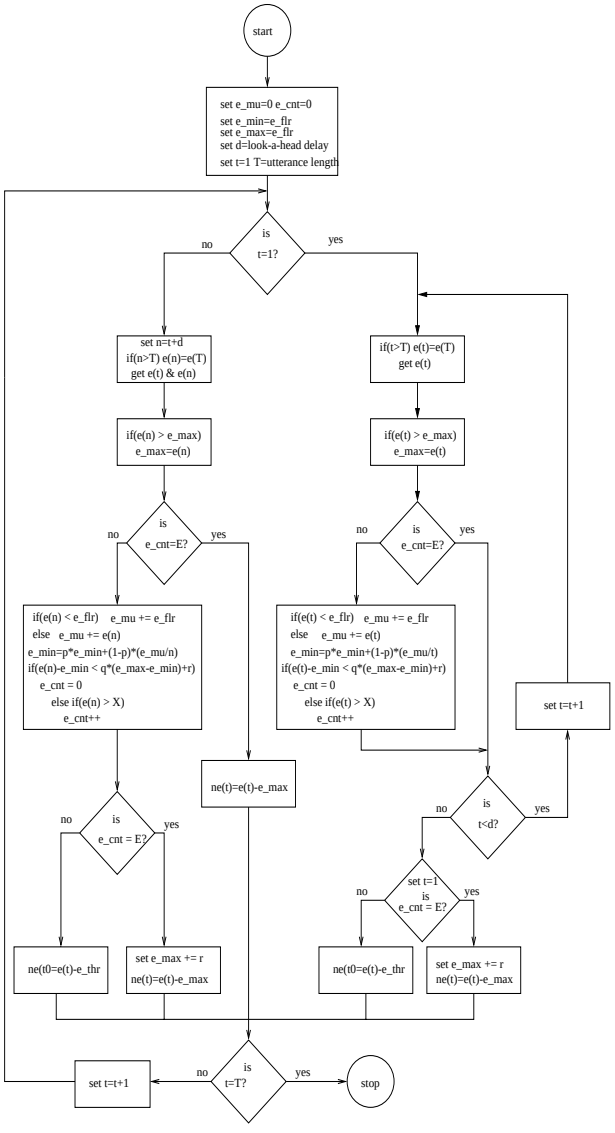

Figure 1. A flow diagram of robust energy normalization algorithm using speech/nonspeech decision mechanism.

2.2. Look-A-Head Delay Based Energy Normalization

Initially the peak-energy is set to a fixed threshold at the begining of each utterance (e_max set to e_thr). Then the peak-energy is estimated by looking at the window of look-a-head delay frames.

$$if\ e(j) \geq e\_max \quad set\ e\_max = e(j)$$
$$t \leq j \leq (t + look - a - headdelay).$$

Finally the current frame energy is normalized according to the estimated peak-energy. The above two steps are repeated until the end of an utterance is occured. We indicate this algorithm as EN_2. The main drawback of this algorithm is that the peak-energy is updated only in the upward direction (monotonically increasing trend). If a database has an uniform distribution of low and high energy speakers then it is very difficult to come up with a single threshold value that is optimal (unbiased) for most of the speakers. For example, if the threshold is set to too low then most of the background noise will tend to appear as speech. On the other hand, if the initial threshold is set to too high then most of the speech will appear as silence.

2.3. Speech/Nonspeech Discriminator Based Energy Normalization

To overcome the problems associated with EN_2, we propose a novel energy normalization algorithm based on speech/nonspeech decision mechanism. The new algorithm is simply denoted as EN_3. The idea is based on the two-level cepstral mean subtraction, where separate channel compensation is performed for segments that are classified as speech and for segments classified as background [2, 1]. In EN_3, the initial background energy is normalized with a higher peak-energy so that the background noise will appear as silence and the subsequent speech frames are normalized using the estimated peak-energy so that the noisy speech will tend to appear as speech and not as background noise.

The flow diagram of EN_3 algorith is shown in Figure 1. The constants in the flow diagram are derived empirically after extensive testing on a large amount of speech data sampled at 8kHz and the suggested analysis frame length is 10ms. Here the initial peak-energy (initially e_max is set to e_flr) can be adapted to a particular utterance and it is updated automatically as the new frame comes in. So the peak-energy has the flexibility of adapting both in downward and upward directions depending upon the current utterance energy level. First the frame energy is classified as either speech or nonspeech class. If the incoming frame stays in speech class for more than 10 frames (E is set to 10) then we pick the peak-energy in those 10 (look-a-delay is set to 10) frames and add some deviation factor (r is set to 5) to get an initial peak-energy. And this peak-energy can be revised once in the downward direction, and monotonically revised in the upward direction afterwards as exemplified in Figure 1.

To illustrate the nature of this algorithm , Figure 2 shows the actual frame energy trajectory and the corresponding parameters involved in the normalization. Top plot shows the unnormalized energy, the second plot provides the e_cnt, the third plot presents the e_max and the bottom plot illustrates the normalized energy contour for a given utterance. It is observed that the EN_3 provides better speech and nonspeech classification and further enhances the e_max estimation. The same observation is also made by using cellular car noise data as shown in Figure 3. In all the experiments

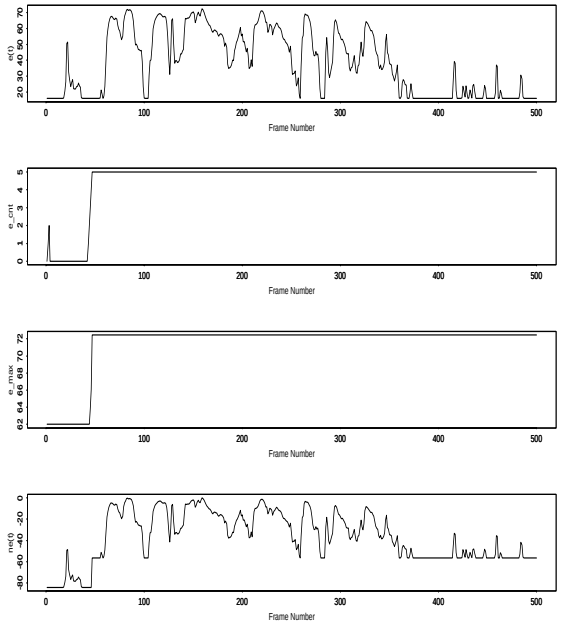

Figure 2. Typical energy measurement contours for the German utterance "0803141815" spoken by a female speaker using SieTill telephone connected digit corpus.

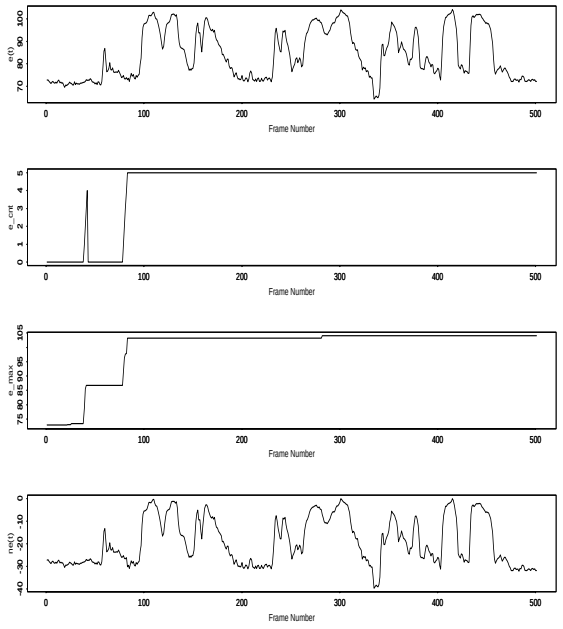

Figure 3. Typical energy measurement contours for the English utterance "4122413655" spoken by a male speaker using cellular handsfree digit string database.

Databases	Training		Testing	
	Str	Spk	Str	Spk
SieTill	14988	652	2192	69
SpeechDat	4952	1738	725	250
Cellular/HF	–	–	1000	14

Table 1. Distributions of spoken digit strings and the speaker population among the training and testing sets of the connected digit database. The top two rows represent the German connected digit corpus and the bottom row indicates the English digit string database.

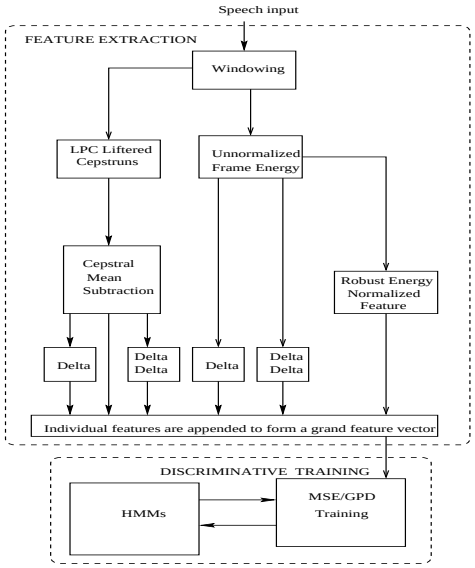

Figure 4. A block diagram of feature extraction using signal conditioned MSE training.

we tend to use the same e_thr value (e_thr is set to 100 db in our current study) for every database which is quite convenient in the sense that we don't need to tune the values for each and every data set.

3. EXPERIMENTAL RESULTS

This section describes the databases, German and English, used in this study. German corpus contains digits eins, zwei, zwo, drei, vier, funf, sechs, sieben, acht, neun and null [6, 11]. The English database contains digits one through nine, zero and oh. Digit string lengths range from 1 to 16 digits. The data distribution of the training and testing set is shown in Table 1. The English test corpus comprised speech data from 14 speakers (7 male and 7 female) collected in a car driving on the highway at 55 mph or greater. The windows were closed and the radio and fan were switched off. The subject was seated in the front passenger side and the date were recorded on CDMA cellular network via typical microphones: lapel and visor-mounted.

The basic recognizer feature set, denoted as, EN_0, consists of 38 features that includes the 12 liftered cepstral coefficients, the first and second order derivatives of cepstrums and log energies [9]. Besides the 38 features, EN_1 contains the batch normalized energy, EN_2 consists of the look-a-head based normalized energy and EN_3 has speech/nonspeech discriminator based normalized energy as

Feature Vector Size and Type	MCE training	
	Wd_Er	St_Er
38 features EN_0	7.63%	22.87%
39 features EN_1	4.98%	16.39%
39 features EN_2	6.10%	20.19%
39 features EN_3	5.52%	18.31%

Table 2. Word error rate (Wd_Er) and string error rate (St_Er) for an unknown-length grammar-based German connected digit recognition task using the MSE training methods as a function of frame vector size and type.

Feature Vector Size and Type	MCE training	
	Wd_Er	St_Er
38 features EN_0	6.50%	34.48%
39 features EN_1	4.06%	25.66%
39 features EN_2	4.63%	27.91%
39 features EN_3	4.18%	25.72%

Table 3. Word error rate (Wd_Er) and string error rate (St_Er) for a 10-digit known-length grammar-based English connected digit recognition task using the MSE training methods as a function of frame vector size and type.

explained in the previous section. Thus, each speech frame becomes represented by a vector of 39 features except for EN_0. 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. In this study, we model all possible inter-word coarticulation, resulting in a total of 276 context-dependent sub-word models. Each model is represented with 3 or 4 states, each having multiples of 4 mixture components. Silence is modeled with a single state model having 32 mixture components [1]. Training included updating all the parameters of the model, namely, means, variances and mixture gains using six epochs of MSE training [5]. Each training utterance is signal conditioned by applying cepstral mean subtraction (CMS) prior to being used in MSE training as shown in Figure 4 [8, 10].

Table 2 shows the performance comparison between various energy normalization schemes with the conventional 38 feature model EN_0 (without the energy feature). It is observed that EN_2 model performs better than 38 feature model EN_0 but inferior to batch normalized 39 feature model EN_1. Also EN_3 based model is better than EN_0 and EN_2 models. About 12% string error rate reduction is obtained by using the EN_2 scheme when compared with 38 feature model EN_0. Further reduction of 10% string error rate is achieved by using EN_3 scheme. The same set of observation is also made by running a similar kind of experiment on English cellular car noise data. The test results are shown in Table 3 for the sake of completeness. In conclusion, the EN_3 performs better than EN_0 and EN_2 models and approaches the batch normalized energy performance in noisy environment and real-time applications.

4. CONCLUSIONS

In this paper, we described the two possible ways of normalizing the energy contour for real-time applications and discussed the implementation of the recognizer. We developed a novel energy normalization technique where the initial peak-energy can be adapted to a particular utterance and it is updated automatically as the new frame comes in. An attempt has been made to use a best look-a-head delay of 20, which has been determined by an early fast experiment. The proposed algorithm uses a 20 frame look-a-head delay to get an initial peak energy and speech/nonspeech decision mechanism to initiate the peak energy adaptation. Experiments on cellular and telephone connected digit recognition task showed about 25% string error rate reduction by using the proposed energy normalization scheme when compared to without using the energy as an additional feature. Moreover, in the look-a-head case, this performance gain is obtained with the smallest implementation costs.

Acknowledgements

The author would like to thank Rafid Sukkar and Anand Setlur for helpful discussions and support in the early stages of this work.

REFERENCES

[1] R.Chengalvarayan, "A comparative study of hybrid modelling techniques for improved telephone speech recognition", *Proc. ICSLP*, 1998, pp. 313-316.

[2] S.K. Gupta, F. Soong and R. Haimi-Cohen, "High accuracy connected digit recognition for mobile applications", *Proc. ICASSP*, 1996, pp. 57-60.

[3] B.H. Juang, L.R. Rabiner, S.E. Levinson and M.M.Sondhi, "Recent developments in the application of hidden Markov models to speaker-independent isolated word recognition", *ICASSP*, 1985, pp. 131-134.

[4] J.C. Junqua, D. Fohr, J.F. Mari, T.H. Applebaum and B.A. Hanson, "Time derivatives, cepstral normalization and spectral parameter filtering for continuously spelled names over the telephone", *Proc. EUROSPEECH*, 1995, pp. 1385-1388.

[5] S. Katagiri, B.H. Juang and C.H. Lee, "Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method", *Proc. IEEE*, Vol. 86, No. 11, November 1998, pp. 2345-2373.

[6] D.Langmann, A.Fischer, F. Wuppermann, R. Haeb-Umbach and T.Eisele, "Acoustic frontends for speaker-independent digit recognition in car environments", *Proc. EUROSPEECH*, 1997, pp. 2571-2574.

[7] L. Rabiner and B.H. Juang, "Fundamentals of speech recognition", *Prentice Hall Signal Processing Series*, 1993, pp. 274-284.

[8] M. Rahim, B.H. Juang, W. Chou and E. Buhrke, "Signal conditioning techniques for robust speech recognition", *IEEE Signal Processing Letters*, Vol. 3, No. 4, 1996, pp. 107-109.

[9] D.L. Thomson and R. Chengalvarayan, "Use of periodicity and jitter as speech recognition features", *Proc. ICASSP*, 1998, pp. 21-24.

[10] O. Viikki, D. Bye, K. Laurila, "A recursive feature vector normalization approach for robust speech recognition in noise", *Proc. ICASSP*, 1998, pp. 733-736.

[11] L. Welling, S. Kanthak and H. Ney, "Improved methods for vocal tract normalization", *Proc. ICASSP*, 1999, pp. 761-764.