Energy Normalization in Automatic Speech Recognition

Nikša Jakovljević[1], Marko Janev[1], Darko Pekar[2], and Dragiša Mišković[1]

[1] University of Novi Sad, Faculty of Engineering, Trg Dositeja Obradovića 6
21000 Novi Sad, Serbia
jakovnik@uns.ns.ac.yu
[2] AlfaNum ltd., Trg Dositeja Obradovića 6
21000 Novi Sad, Serbia

Abstract. In this paper a novel method for energy normalization is presented. The objective of this method is to remove unwanted energy variations caused by different microphone gains, various loudness levels across speakers, as well as changes of single speaker loudness level over time. The solution presented here is based on principles used in automatic gain control. The use of this method results in relative improvement of the performances of an automatic speech recognition system by 26%.

1 Introduction

Acoustic variability is one of the main contributors to degradation of performances of automatic speech recognition (ASR) systems. A systematic review of what has been done in this area can be found in [1]. One of the standard features used in ASR systems is the energy of the speech signal. The energy can be helpful for phoneme discrimination (e.g. vowels vs consonants; voiced vs unvoiced).

Unwanted energy variations are caused by many factors, the dominant one being background noise. Background noise drastically changes the sound level of silent segments, and on the other hand slightly changes the sound level of loud segments. Several solutions for this problem are proposed in [3,4,5,6].

Unwanted energy variations can also be caused by: (1) different microphone gains, (2) different microphone placement, (3) variations in loudness levels across different speakers as well as (4) changes in loudness level of a single speaker over time.

A silent environment, where this ASR system will be used, results in bigger influence of other causes of energy variations. Automatic gain control which is used on sound cards does not have satisfactory effects, and for that reason a certain variation of gain control for energy normalization should be applied in the ASR front-end block.

In this paper a novel energy normalization procedure based on automatic gain control principles is presented. The paper is organized as follows. In the following section a short description of existing methods will be presented along with reasons why they could not be successfully applied in this case. The proposed algorithm will be presented in Section 3. Experimental results and a brief description of the used ASR system will be given in Section 4. Finally, in Section 5 conclusions and possible future work will be presented.

P. Sojka et al. (Eds.): TSD 2008, LNAI 5246, pp. 341–347, 2008.

2 Energy Normalization Methods in ASR Systems

In this section a brief review of existing energy normalization methods along with comments on their ability to achieve the newly established goal is presented. Energy normalizaton needs to eliminate energy variations caused by variations in loudness levels across different speakers and changes in loudness level of a single speaker over time. An additional constraint is the requirement for real-time processing.

One of the first energy normalization methods is cepstral mean normalization (CMN). Since the first cepstral coefficient is energy, CMN can be viewed as an energy normalization method, which is the usual approach in this framework. As it is well known, the basic goal of CMN is the elimination of the influence of the channel. The success of CMN depends on the duration of the averaging interval. The longer interval means that the average CMN value is less dependent on the spoken phone sequence, thus the effects of CMN are better. On the other hand, if the interval is too short, CMN may result in degradation of the ASR performance. The latter is the reason why CMN cannot be used in real-time applications. Similar conclusions hold for cepstral variance normalization (CVN). Both CMN and CVN can reduce energy dispersion caused by variations in loudness level across speakers, but cannot compensate energy variations within a single utterance. They both presume that loudness level does not change over the course of the utterance. Details about CMN and CVN can be found in [2].

Log-energy dynamic range normalization (ERN) reduces energy dispersion caused by different levels of background noise. This method is based on the fact that the same noise results in small changes in log energy on high-energy segments (e.g. stressed vowel) but in drastic changes on low-energy segments (e.g. silence, occlusions of stops and affricates). It is assumed that all utterances of all speakers have the same maximum energy as well as dynamic range. Under this assumption, the target minimum log energy value is set based on the estimated maximum energy in the utterance and the assumed dynamic range. If the minimum log energy in a single utterance is less than the calculated target minimum log energy, the energy should be rescaled to a specified target range, otherwise nothing should be done.

More details about this procedure can be found in [3,4]. It is clear that ERN is not the solution for the problem of energy normalization as defined in this paper.

3 Method Description

The input parameter for the energy normalization block is frame energy. Frames are 30 ms long and shifted by 10 ms, extracted by application of a Hamming window function. In this way central samples of the frame carry greater weight than those near the boundaries. In order to carry out energy normalization of a single utterance, it is necessary to track peak energy of successive speech segments. The standard way to achieve this is using IIR systems defined by:

$$E_p(n) = \gamma E_p(n-1) + (1-\gamma)E(n) \tag{1}$$

where $E_p(n)$ is the peak energy at the n-th frame, $E(n)$ is the energy at the n-th frame and γ is the memory coefficient. From Equation 1 it follows that the values of the

memory coefficient lie in the interval (0, 1], and that any other value has no physical meaning. Since it should be desirable to "catch" peak energy fast and to change it slowly once it is "caught", the value of the memory coefficient should be specified separately for segments with rising and falling energy. In the case of rising energy, the value of the memory coefficient should be small. On the other hand, during the segments with falling energy, the value of the memory coefficient should be relatively close to 1. In this way the resulting peak energy would change very slowly, but quickly enough to avoid omitting the next maximum if it is smaller than the previous one. In this paper the values of the memory coefficient of $\gamma_r = 0.30$ for rising and $\gamma_f = 0.99$ for falling energy were adopted. The normalization process consists of dividing the current energy value $E(n)$ with the current peak tracker value $E_p(n)$ i.e.:

$$E_n(n) = E(n)/E_p(n) \tag{2}$$

where $E_n(n)$ is the normalized value of the energy at the n-th frame. The meaning of the other variables is the same as in Equation 1. Since peak energy decreases constantly over silent segments, the normalization strategy described above results in the increase of the noise level at silent segments, as shown in Figure 1. One way to overcome this problem is the implementation of separate energy normalization procedures for silent segments and for segments with speech. This approach requires the introduction of a function for automatic detection of speech activity. A simple solution presented in [7] is applied in this paper. Two additional track curves are used, one for fast energy tracking ($E_f(n)$) and one for slow energy tracking ($E_s(n)$). If $E_f(n) > E_s(n)$, the n-th frame is a part of a speech segment, otherwise it is a part of a non-speech segment. The values of $E_f(n)$ and $E_s(n)$ are calculated by:

$$E_i(n) = \gamma E_i(n-1) + (1-\gamma)E(n) \tag{3}$$

where the index i can be either f or s. The meaning of the other parameters is the same as in Equation 1. As well as in the case of peak energy tracking, the value of the memory coefficient γ depends on whether the energy is rising or falling. Consequently there are 2 values of the memory coefficient for each of the trackers. The memory coefficients for the slow tracker in the case of rising and falling energy are marked by γ_{sr} and γ_{sf}, while those for the fast tracker are marked by γ_{fr} and γ_{ff}, with the following relations holding:

$$\begin{aligned} \gamma_{sr} &\leq \gamma_{sf} < 1 \\ \gamma_{fr} &\leq \gamma_{ff} < 1 \\ \gamma_{fr} &\leq \gamma_{sr} \\ \gamma_{ff} &\leq \gamma_{sf} \end{aligned} \tag{4}$$

Acceptable performances are achieved with the following values of memory coefficients: $\gamma_{sr} = 0.85$, $\gamma_{sf} = 0.95$, $\gamma_{fr} = 0.80$ and $\gamma_{ff} = 0.90$.

The algorithm for automatic detection of speech activity mentioned above requires that the maximum noise energy be set. A silent environment, where this ASR system will be used, provides low noise level, and thus the value of the maximum noise level can be set easily. In this specific case its value is 10^{-4}.

The rule for calculation of normalized energy is thus modified. One level of peak energy is used for silent segments (marked by $E_{ps}(n)$) and the other level of peak energy

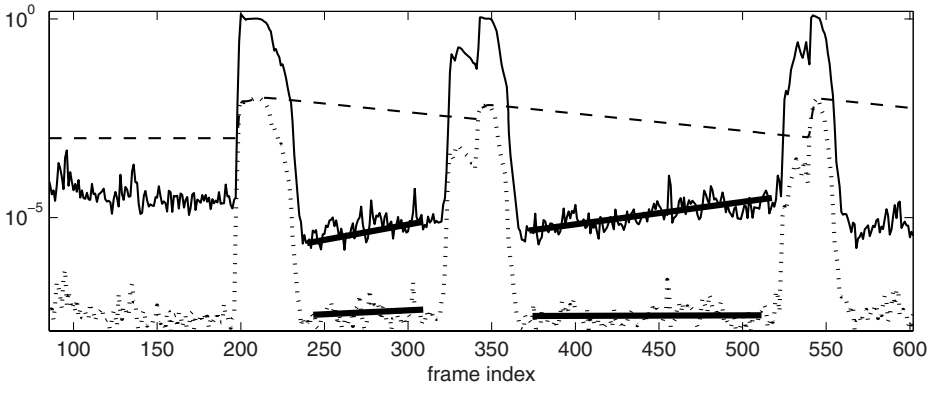

Fig. 1. Frame energy (*dotted line*), normalized energy (*solid line*) and peak energy (*dashed line*) of the speech signal. The decrease of the peak energy (normalization level) on the silent segments (*numbered from 230 to 320 and 370 to 520*) results in an increase of the normalized noise energy.

for speech segments ($E_{pv}(n)$). The value of the $E_{pv}(n)$ is the same as the value of the $E_p(n)$ defined by Equation 1. The value of the $E_{ps}(n)$ is the value of the E_{pv} in the last speech frame before the current silent segment. This new rule for energy normalization can be summarized as:

$$E_n(n) = \begin{cases} E(n)/E_{pv}(n) \text{ for speech segments} \\ E(n)/E_{ps}(n) \text{ for silent segments} \end{cases} \quad (5)$$

Another possibility is to normalize energy at silent segments with respect to the current maximum energy value within the whole utterance. Although the noise level is almost the same in the whole utterance and therefore normalization with respect to the same level should have much more sense, the uncertainty of speech activity detection at some segments with consonant-vowel transitions makes this alternative solution impossible. Disregard of this fact leads to errors such as the one shown in Figure 2.

An additional requirement that a speech segment last at least n_{min} frames is introduced in order to ensure that the value of the $E_{ps}(n)$ be the actual peak value of the last speech frame. This constraint implies that $E_{ps}(n)$ is changed only if the last n_{min} successive frames are detected as speech. In this specific case, the value of n_{min} is 3. Since the maximum signal energy rarely occurs at the beginning of the word, it is necessary to calculate the values of peak energy at several following frames as well, therefore a tolerable delay D is defined. The estimation of the maximum energy is better if the number of following frames included in the calculation is greater. On the other hand, the application demands for real-time processing set the upper bound of delay. The value of the delay D of 10 frames i.e. 100 ms was adopted.

To avoid the possibility that after a longer silence interval, energy normalization is performed with initial values related to silent frames, the minimum value of the peak energy E_0 is set. If the value of the peak energy $E_{pv}(n)$ becomes less than E_0, $E_{pv}(n)$ is set to E_0. Too great a value of E_0 at speech segments with smaller energy can result in inappropriate normalization. For this application the value of 10^{-3} was adopted.

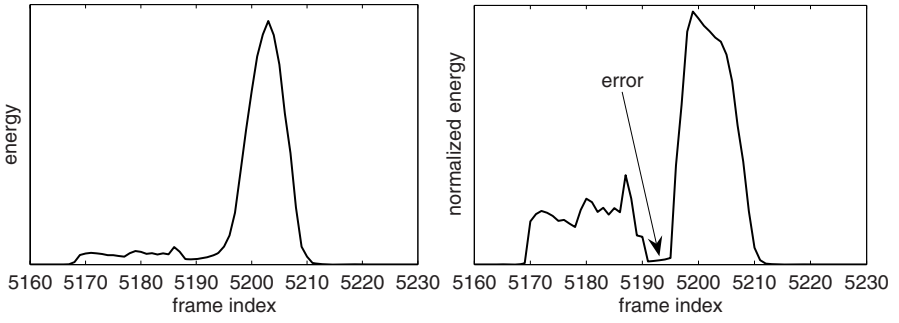

Fig. 2. Energy (*left*) and resulting normalized energy (*right*) on the same speech segment. A small part of the speech segment is detected as silence (*frames from 5191 to 5194*). That misclassification results in unacceptable variation of the normalized energy for the speech segment caused by different normalization levels for the speech and silent segments.

4 A Description of the ASR and Achieved Performances

Two systems are created to estimate performances of the proposed method for energy normalization. These two systems differ only in energy coefficients. Feature vectors in both systems, baseline system and the system which uses the proposed method for energy normalization, contain 34 coefficients, where 32 of them are common (16 Mel-frequency spectral envelope coefficients and their derivatives) for both systems. In the baseline system log energy and the derivative of energy are used instead of log of normalized energy and the derivative of normalized energy. More details about Mel-frequency spectral envelope coefficients are given in [8].

The particular software application where the ASR system will be used requires a simple grammar. The ASR input is a sequence of consonant vowel (CV) words. Between words there are silent intervals. Such a grammar structure leads to a reduced number of contexts (the left context for consonants and the right context for vowels are silence) as well as a reduction of feature dispersion caused by coarticulation effects. Although the number of possible words is small, the basic modelling unit is a context dependent phone.

The ASR system is based on hidden Markov models (HMM) and Gaussian mixture models (GMM). Instead of a diagonal covariance matrix, each Gaussian distribution is determined by a full covariance matrix. It was sufficient that an HMM state be modelled by a single Gaussian distribution, because of the reduction of the number of different contexts.

The training corpus contains 7 hours of speech. The audio files are encoded in the PCM format (22050 Hz, 16 bits per sample). As already noted, the noise level is negligible.

The results are presented in Table 1. The test corpus contains 2513 CV words i.e. 5026 phonemes. With regard to specific purposes of the ASR, phone error rate (PER)

Table 1. System performances

Is normalization used?	no. of substitutions	no. of insertions	no. of deletions	PER [%]
NO	508	56	0	11.22
YES	401	16	0	8.30

was used as a performance measure instead of word error rate. The standard way to evaluate ASR features is by measuring the relative improvement, defined by:

$$R.I. = \frac{NewAcc - BaseAcc}{100 - BaseAcc} \times 100\% \tag{6}$$

where $NewAcc$ is the accuracy of a system with new features and $BaseAcc$ is the accuracy of a baseline system. Relative improvement in terms of PER is defined by:

$$R.I. = \frac{BasePER - NewPER}{BasePER} \times 100\% \tag{7}$$

For the results presented in Table 1 the relative improvement is 26%. This improvement is caused by the reduction of energy dispersion in phone models.

5 Conclusion and Future Work

In this paper a new method for on-line energy normalization is presented. The method is based on principles used in automatic gain control. Its objective is to reduce the dispersion of energy in an HMM state, caused by variable microphone gains, loudness levels across speakers and loudness level of a single speaker over time. A relative improvement of about 26% was achieved on the test set. The test set contains high quality audio files (22050 Hz, 16 bits per sample) with low noise level.

The future research should include an applicability study on the use of this method for telephone quality speech signals with a higher noise level. A possible solution could be a combination of this method with some of the methods presented in Section 2. These steps demand the existence of an ASR system for English and an Aurora 2 test set intended for such research.

References

1. Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., Wellekens, C.: Impact of Variabilities on Speech Recognition. In: Proc. SPECOM 2006 (2006)
2. Togneri, R., Toh, A.M., Nordholm, S.: Evaluation and Modification of Cepstral Moment Normalization for Speech Recognition in Additibe Babble Ensemble. In: Proc. SST 2006, pp. 94–99 (2006)
3. Lee, Y., Ko, H.: Efective Energy Feature Compensation Using Modified Log-energy Dynamic Range Normalization for Robust Speech Recognition. IECIE Trans. Commun. Anal. E90-B(6), 1508–1511 (2007)

4. Zhu, W., O'Shaughnessy, D.: Log-energy Dynamic Range Normalization for Robust Speech Recognition. In: Proc. ICASSP 2005, vol. 1, pp. 245–248 (2005)
5. Zhu, W., O'Shaughnessy, D.: Using Noise Reduction Andspectral Emphasis Techniques to Improve ASR Performance Innoisy Conditions. In: Proc. ASRU 2003 (2003)
6. Zhu, W., O'Shaughnessy, D.: Incorporating Frequency Masking Filtering in a Standard MFCC Feature Extraction Algorithm. In: Proc. ICSP 2004, vol. 1, pp. 617–620 (2004)
7. Hänsler, E., Schmidt, G.: Acoustic Echo and Noise Control. New Jersey. Wiley, Chichester (2004)
8. Jakovljević, N., Mišković, D., Sečujski, M., Pekar, D.: Vocal Tract Normalization Based on Formant Positions. In: Proc. IS LTC 2006 (2006)