Use of syllable center detection for improved duration modeling in Chinese Mandarin connected digits recognition

Sergey Astrov, Joachim Hofer, Harald Höge

Corporate Technology, Siemens AG, D-81730 Munich, Germany
{sergey.astrov; joachim.hofer; harald.hoege}@siemens.com

10.21437/Interspeech.2007-501

Abstract

This paper describes practical approaches for improving Mandarin digit recognition accuracy, especially in cars. We consider syllable and subword unit durations as additional source of information. The explored approach was realized in two stages. First, the system performs standard speech recognition using acoustic spectral features. As a result, an n-best list of hypotheses is generated. In the second stage the hypothesis probabilities are re-estimated using duration models, thus, the hypotheses are reordered such that the correct ones are pushed to the top of the n-best list. In such a way the word error rate (WER) is reduced.

We explore state of the art approach of duration n-grams. In order to eliminate the influence of speech rate variations, the durations are normalized to a relative speech rate, a 10% relative reduction of WER was achieved. A novel approach led to 13.3% WER reduction: the durations were normalized to a syllable rate obtained from the syllable center detector.

1. Introduction

Reliable speech recognition in cars is a challenging task. Modern speech controlled systems perform well in quiet environments, but motor and street noise strongly reduce their accuracy. Noise reduction algorithms solve this problem, yet, the recognition quality does not satisfy user demands for car applications.

Higher speech recognition quality may be achieved in another way: some other information sources - such as several audio channels in microphone arrays, video stream for speech recognition combined with lip reading or implementation of additional features - may supplement commonly used audio spectral features.

Typically, suprasegmental features (durations, pitch contours, energy) with lengths from one half to several seconds are ignored in common speech recognition systems: the audio signal is divided into short term frames of 10-100 ms. The use of prosodic information can improve speech recognition quality.

This work considers the application of duration models for improved recognition of continuous digits in Chinese Mandarin. Under duration we consider the duration of suprasegmental units, such as words, syllables and whole-word HMM segments — these units have durations of more than one frame.

Laboratory speech recognizers that use duration features were known many years ago. Only in the last decade the publications describing such recognizers with practical realizations have appeared. State of the art recognizers have reduced their WER up to 30% by employing duration models [1, 2, 3, 4, 5].

Another main point of this paper is the normalization of durations to the syllable rate by means of the syllable center detector. In experiments, samples from Mandarin SPEECON speech database (car subset) were recognized using HMMs designed for embedded devices, in such a way the recognition quality in cars is evaluated.

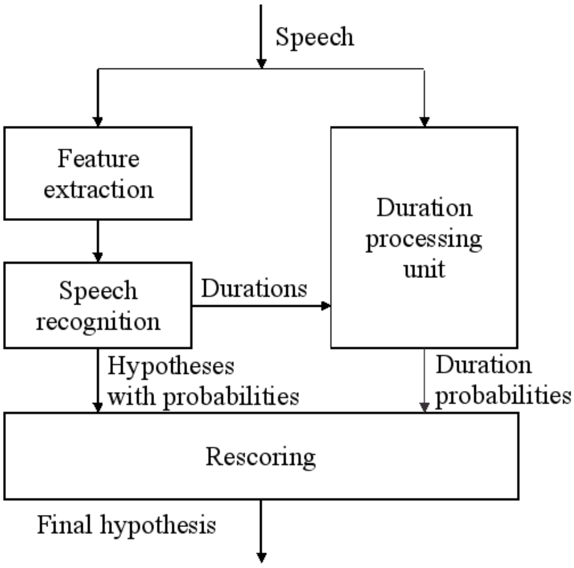

Figure 1: Recognizer structure

In Section 2 we present our speech recognition system and a general idea of the duration modeling, used speech database and recognition test sets are listed in details. Sections 3 and 4 describe theoretical and practical aspects of speech rate and duration models. Section 5 deals with the syllable center detection required for speech rate computation. In Section 6 we present obtained experimental results, and finally, in Section 7 we make conclusions and discuss our future work.

2. Recognition system

The structure of the recognition system is shown in Figure 1. The core of the system forms EAR — our HMM-based research speech recognizer. After noise reduction using spectral subtraction, the signal is converted to a sequence of features. Each feature is obtained as a result of an LDA transformation of a supervector, which consists of 13 MFCCs with deltas and delta-deltas from 3 frames. Our gender-dependent HMM set for embedded devices has totally 5285 Gaussians, the training was performed on Mandarin SPEECON database [6, 7]. The HMM training set consists of 14664 continuous digit utterances with 25708 words, the Mandarin utterances were recorded in cars in China and Taiwan.

The speech recognition result of the considered system is an n-best list of hypotheses where each entry is supplied with respective segmentation information and hypothesis probability. From the segmentation the duration of syllables[1] are obtained.

Afterwards, the duration processing unit estimates how likely the hypotheses are. The linear combination of spectral and duration probabilities is used for the reordering of the n-best list (rescoring), the most probable hypotheses are pushed to the top of the list. The n-best list is restricted by five hypotheses because further increase of the n-best list size does not improve the recognition accuracy in our case.

For the experiments with Chinese Mandarin connected digit strings we use SPEECON speech database. The duration model training set consists of 2392 continuous digit utterances with 13520 words. The experimental test set consists of 662 digit utterances with a total of 3368 digits. We selected here only Mandarin digits recorded in China. Two hands-free (middle and far distance) from four available channels were used. The recognition accuracy of the reference recognizer with the described test set is 8.3% WER.

3. Speech rate

The durations of phonetic units depend on speech rate variations: a person hardly can pronounce a phrase twice with exactly the same speech rate. In order to weaken the influence of the speech rate, the durations are normalized and the model accuracy is increased.

In the following the confusing term “speech rate” is clarified, as it has several different interpretations [8, 9]. We list some of them below and explain normalization algorithms of word durations.

3.1. Relative speech rate

The *relative speech rate* reflects how fast the utterance is pronounced in comparison to the “average” speech rate of the “average” person [9]. The equation below shows the estimation of the relative speech rate in a sentence:

$$rate = \frac{\sum \mu(w_i)}{\sum d_i}$$

where d_i is the measured duration of the i-th syllable w_i in the sentence, $\mu(w_i)$ is the mean duration of w_i obtained from training data.

During training, word durations are obtained from the automatic segmentation using forced alignment, when the phrase transcriptions are known.

During recognition the correct segmentation is not known, that is why we estimate relative speech rate for each hypothesis from its segmentation data. The normalized syllable durations are computed as follows:

$$d_{norm,i} = d_i \cdot rate$$

3.2. Syllable rate

Another speech rate measure is the number of syllables[2] per second — the *syllable rate* [10]. We perform syllable rate computation how it is described in [8]: in a window of 615 ms we count the amount of syllables and incomplete syllable parts:

$$rate = \frac{\frac{S_{l+1}-w_L}{S_{l+1}-S_l} + \frac{w_R-S_r}{S_{r+1}-S_r} + r - l + 1}{S_{l+1} - w_L + w_R - S_r + \sum_{i=l+1}^{r-1} S_{i+1} - S_i}$$

where S_i is the starting point of the i-th syllable, w_L and w_R are left and right window borders. S_l and S_r are left and right borders of the syllables that contain window borders, these syllables are weighted partially. In order to get a syllable rate as a smooth function, the Hanning window is applied, we do not show it here because of complexity, see [8] for details.

Taking into account that each word consists of exactly one syllable, the normalized word duration is computed as the reciprocal number of the average syllable rate. The syllable rate and word duration mismatch result in low duration probabilities, which would penalize the digit string hypotheses that they occur in, thus, insertion and deletion error rates can be reduced.

4. Duration models

4.1. n-gram models

A possible representation of duration models is an n-gram approach [4]. The probabilities are stored in form of histograms or modeled by the weighted sum of functions (e.g. Gaussian, gamma or log-normal distributions). For bigram models, the duration statistics are collected for two neighboring syllables. The equation below demonstrates the recognition probability computation:

$$P(d_1, d_2, \ldots, d_N | w_1, w_2, \ldots .w_N) \approx$$
$$P(d_1|w_1) \cdot \prod_{i=2}^{N} P(d_i|d_{i-1}, w_i, w_{i-1})$$

where d_i is the duration of the i-th word w_i. In order to cope with sparse data issues in training the durations are quantized into 16 levels.

During training, word durations from the forced alignment segmentation are obtained and optionally normalized to the relative speech rate (see Section 3.1), thus, the bigram models are created. In case of a single word in an utterance the duration probability is computed as $P(d_1|w_1)$.

4.2. Subword duration models

We propose another duration model described below. Each word is represented as the sequence of phoneme-like subword units. The segmentation provides information about durations $g_{k,i}$ of the k-th unit $v_{k,i}$ in the i-th word w_i and word durations d_i. The word duration probability may be estimated as:

$$P(d_i|w_i) = \prod_{k=0}^{K_i} P(g_{k,i}|v_{k,i}, d_i, w_i)$$

During the training of models $P(g_{k,i}|v_{k,i}, d_i, w_i)$, the durations are obtained from forced alignment. In the recognition stage, values of $v_{k,i}$ are estimated from the hypothesis segmentation. The word duration d_i is obtained as a reciprocal of the mean syllable rate computed by a syllable center detector over the word boundaries. In case of insertion or deletion, the mismatch of normalized word duration and subword units lead to low probability $P(d_i|w_i)$, thus, the hypotheses with insertion or deletion errors are pushed to the bottom of the n-best list. The syllable center detector is used to estimate the speech rate and thus the word lengths independent of the baseline recognition hypotheses.

[1] Mandarin digits are one-syllable words

[2] or, in general, any phonetic units (phonemes, words, etc.)

4.3. Rescoring

The rescoring procedure has to combine hypothesis probabilities from the recognition with spectral features and from the duration models. The hypothesis probability is estimated as:

$$\begin{aligned} P(W|O)_F &= \sum_D P(W,D|O) \approx \max_D P(W,D|O) \\ &= \max_D P(D|W,O) \cdot P(W|O) \\ &\approx \max_D P(D|W) \cdot P(W|O) \end{aligned}$$

where W denotes the recognized hypothesis of utterance O; D denotes the duration information; $P(W|O)$ denotes the hypothesis probability obtained from spectral features; the probability $P(D|W)$ is obtained from duration models; $P(W|O)_F$ denotes the final probability obtained from spectral and duration information.

Taking negative logarithms of the last equation we get:

$$-\ln P(W|O)_F \approx \min_D(-\ln P(D|W) - \ln P(W|O))$$

The recognition result with spectral features is not exactly equal to $-\ln P(W|O)$, that is why we have to find multiplication parameters. We have to order hypotheses in the n-best list in ascendant order of the hypothesis scores:

$$\min_D(-c_D \cdot \ln P(D|W) - c_P \cdot \ln P(W|O))$$

where $-c_P \cdot \ln P(W|O)$ is the hypothesis score from the first stage, c_D is a multiplication parameter that is estimated by performing a series of experiments with a development set of utterances.

5. Syllable center detection

A syllable and especially its center are hard to define. Nevertheless, humans often agree about the amount and position of perceived syllable centers [8]. A possible optimization criterion for an algorithm may be to maximally coincide with the human perception. The temporal distance between an automatically recognized syllable center and the nearest real center can then be used to determine error rates.

Automatically finding syllable centers is not a trivial task. Many of the proposed methods utilize subbands of the short time energy. [11] claim to have “better syllable count performance than previous methods”. They calculate the short time energy in several mel-frequency aligned sub-bands. The temporal and spectral correlation of some of those bands is calculated and syllable centers are detected with a peak picking algorithm. In a post-processing step, some of the detected syllables are deleted by using a pitch frequency estimator.

There are no labels for perceived syllable centers in the SPEECON-Mandarin speech database. In order to test the estimator on this database, the centers of formal syllables are used as a reference. As the temporal positions of the labels are not very exact, detected syllable centers are allowed to have a distance of up to 100 ms from the reference positions to be still considered correct. Of all the tested energy based methods, the one in [11] showed the best performance. Nevertheless, it still has 25% insertion and deletion errors. Those bad results may be due to the noise present in the signal that led to a large amount of insertions when the threshold of the peak picking was set too low. In order to improve the recognition performance in noisy environment, a statistical method was created.

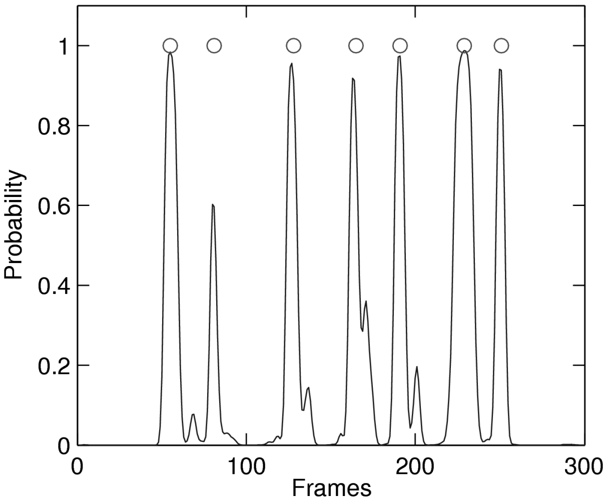

Figure 2: Estimated vowel probability and the position of syllable centers (circles)

Long TempoRal Patterns (TRAPs) [12] are a method to integrate temporal context into the features of a phoneme recognizer. Especially a modification called Split Temporal Contexts (STCs) [13] prove to be superior over traditional MFCCs with deltas and delta-deltas. Using dimensionality reduction with LDA instead of a multistage recognizer based on STCs gave similar results in our experiments while allowing much faster recognition. Thus, this method is used to calculate phoneme probabilities every 15 ms. As most syllables contain exactly one vowel, the sum of all vowel probabilities indicate a syllable center. The per frame estimated vowel probability function has high frequency components that do not contain reliable information. Therefore, a Hamming window with a length of 7 frames is convolved with the function to achieve the necessary smoothing. Figure 2 shows an example of the estimated vowel probability. Some of the local maxima do not correspond to syllable centers, but their height compared to the surrounding local minima is no more than 0.1.

A threshold based peak picking algorithm finds relevant local maxima in the syllable center probability function. It picks only those maxima where the difference in value to the surrounding local minima surpasses a certain threshold. That is, if p is the smoothed probability function and m_1 and m_2 are local maxima, then

$$\begin{aligned} p(m_1) - \min_{m_1<t<m_2}(p(t)) &> T \quad \text{and} \\ p(m_2) - \min_{m_1<t<m_2}(p(t)) &> T \end{aligned}$$

must hold for some threshold T that is optimized on a development set. The final output of the algorithm are all local maxima holding these conditions. The advantage of a relative algorithm over one with a fixed threshold is, that it does not generate insertions for relatively small local maxima in regions with an overall high probability.

The resulting syllable center detector has an insertion and deletion error rate of 9% which is less than half of the errors of the energy based method.

6. Experimental results

The connected digit recognition experiments were performed on SPEECON Mandarin database, see the detailed description of the recognition task in Section 2. The test results are shown in Table 1. In the "Duration model" column all explored models and speech rates for normalization are listed, the reference recognition system with 8.3% WER is denoted as "Baseline". The "WER" column contains absolute word error rates, and the last column shows the WER change in percents relatively to the baseline.

Duration model	WER	relative
Baseline	8.3%	
Syllable duration bigrams	7.8%	-6.3%
Syllable bigrams + relative speech rate	7.5%	-10.0%
Subword models	8.2%	-1.4%
Subword models + syllable rate	7.2%	-13.3%

Table 1: Experimental results

First, we have tested syllable length bigrams without normalization and achieved 7.8% WER, the relative reduction is 6.3%.

In the next experiment we have normalized word durations to the relative speech rate. This approach has improved recognition accuracy further: WER is now reduced by 10% (from 8.3% to 7.5%).

The use of subword duration models brings improvement: WER value decreases by 1.4% (absolute value 8.2%). The same approach with the normalization to the syllable rate leads to the best result in all experiments: WER was reduced by 13.3% (from 8.3% to 7.2%). The normalization of durations should be performed: in the database we have observed syllable durations from 0.2 seconds in a sequence of digits up to one second in case of a single word in an utterance. Our subword duration models had poor performance without normalization because of insufficient amount of training material that can cover all possible duration variations, and only with the normalization it became possible to train robust models.

7. Conclusion and future work

The example of continuous Mandarin digits recognition in cars shows that the accuracy may be improved by addition of suprasegmental information to the commonly used spectral features: WER was reduced by 13.3%.

The experiments evaluate two speech rate computation methods that use word boundary segmentation for each hypothesis and a syllable center detector. The application of both normalization methods demonstrates improvement of the recognition accuracy, besides, the second approach is advantageous over the first one because of higher WER reduction caused by employing an additional source of information. The syllable rate is suited to detect insertions and deletions, since syllable center positions in hypotheses and speech signals are indirectly compared. In case of insertions or deletions the duration models give low hypothesis probabilities.

Speech recognition in cars requires practical solutions of many further challenging problems, below is only a short list of future work.

The duration information from the recognition may be not enough accurate, it could be hard to detect syllable and subword unit boundaries: even two phoneticians may label data differently. The automatic detection of phoneme boundaries may be even less precise because of the used acoustic models: the models of transition segments are influenced by both neighboring phonemes. The use of temporal distances between phonetic unit centers instead of durations may improve recognition performance. The detection of the phoneme or syllable centers is more stable then the detection of boundaries.

The rescoring performed in an additional stage has a serious disadvantage: the hypothesis cannot be pushed to the top of the n-best list if it is not recognized. In our experiments about 50% of the sentences had no error free hypothesis in an n-best list with $n = 200$. The implementation of the duration models in the first stage together with a Viterbi search may solve this problem, because the hypotheses with improbable durations will be eliminated sooner from the n-best list, thus allowing correct hypotheses to appear.

8. References

[1] M. Ariu, T. Masuko, S. Tanaka, and A. Kawamura, "Speech recognition using syllable duration ratio model," in *Proc. ICASSP*, 2006, pp. 341–344.

[2] G. Y. Chung and S. Seneff, "A hierarchical duration model for speech recognition based on the angie framework," *Speech Communications*, vol. 27, no. 2, pp. 113–134, 1999.

[3] C. Wang and S. Seneff, "A study of tones and tempo in continuous mandarin digit strings and their application in telephone quality speech recognition," in *Proc. ICSLP*, 1998, pp. 635–638.

[4] D. Willett, F. Gerl, and R. Brueckner, "Discriminatively trained context-dependent duration-bigram models for korean digit recognition," in *Proc. ICASSP*, 2006, pp. 25–28.

[5] V. R. R. Gadde, "Modeling word durations," in *Proc. ICSLP*, 2000, pp. 601–604.

[6] R. Siemund, H. Höge, S. Kunzmann, and K. Marasek, "SPEECON — speech data for consumer devices," in *Proc. (LREC)*, 2000, pp. 883–886.

[7] SPEECON-Homepage, "*www.speechdat.org/speecon/*."

[8] H. R. Pfitzinger, "Phonetische Analyse der Sprechgeschwindigkeit (in German)," Dissertation, Institut für Phonetik und Sprachverarbeitung der Ludwig-Maximilians-Universität, 2001.

[9] C. Wang and S. Seneff, "Lexical stress modeling for improved speech recognition of spontaneous telephone speech in the JUPITER domain," in *Proc. EUROSPEECH*, 2001, pp. 2761–2764.

[10] H. R. Pfitzinger, "Two approaches to speech rate estimation," in *6th Australian Int. Conf. on Speech Science and Technology*, 1996, pp. 421–426.

[11] S. Narayanan and D. Wang, "Speech rate estimation via temporal correlation and selected sub-band correlation," in *Proc. ICASSP*, 2005, pp. 413–416.

[12] H. Hermansky and S. Sharma, "TRAPs – Classifiers of temporal patterns," in *Proc. ICSLP*, 1998, pp. 1003–1006.

[13] P. Jain and H. Hermansky, "Beyond a single critical-band in TRAP based ASR," in *Proc. EUROSPEECH*, 2003, pp. 437–440.