Use of Phonetic Unit Duration for Chinese Mandarin Digit Recognition in Cars

Sergey Astrov
Corporate Technology, Siemens AG, 81739 Munich, Germany

Abstract

State of the art speech recognizers use spectral features for the recognition. For improving of performance, especially in cars, the use of syllable durations is investigated: in experiments a 13.3% reduction of word error rate (WER) was achieved on a Chinese Mandarin continuous digits recognition in car environment. The explored approach was realized in two stages. First, the system performed standard speech recognition using acoustic spectral features. As a result, an n-best list of hypotheses was generated. In the second stage the hypothesis probabilities were re-estimated using syllable duration information, thus, the hypotheses were reordered such that the correct ones were pushed to the top of the n-best list. In such a way the word error rate (WER) was reduced. The syllable durations were normalized in order to eliminate the influence of speech rate variations, two different speech rate computation approaches were investigated.

Introduction

Reliable speech recognition in cars is a challenging task. Modern speech controlled systems perform well in quiet environment, 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 from microphone arrays, video stream used for 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 recognition systems: the speech signal is considered as a sequence of short term frames of 10-100 ms.

This paper considers the application of duration models for improved continuous digit recognition in Chinese Mandarin. The normalization of durations to the syllable rate by means of the syllable center detector is another main point. In experiments, samples from Mandarin SPEECON database were recognized using hidden Markov models (HMMs) designed for embedded devices, thus, the recognition quality in cars was evaluated.

Duration models and speech rate

The speech recognition result is an n-best list of hypotheses where each entry is supplied with respective word segmentation and hypothesis probabilities. From the segmentation the duration of syllables are obtained. Afterwards, the combination of the probabilities from the duration models and HMMs results in a new order of best hypotheses (rescoring).

Because of the different speech rates of spoken utterances the desired precision of word duration models may not be reached. Normalization of the word durations to the speech rate may increase model accuracy. But first, the confusing term "speech rate" has to be clarified, as it has several interpretations in literature. The next two sections deal with two most often observed definitions of speech rate.

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 [4]. The equation below shows the estimation of the relative speech rate in a sentence:

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

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

During the recognition the relative speech rate is estimated for each hypothesis from the segmentation data. The normalized syllable durations are estimated as:

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

Duration models are often represented as n-grams. The probabilities are stored in form of histograms or modeled by the sum of functions (e.g. Gaussian pdfs, gamma or log-normal distributions). For bigram models, the duration statistics is collected for two neighboring syllables. Equation below demonstrates the recognition probability computation:

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

Syllable rate

Another speech rate measure is the number of syllables[1] per second - the *syllable rate* [3]. For this case, another duration model is employed: each word is represented by the phoneme-like subword unit sequence. Segmentation provides information about durations $g_{k,i}$ of the k-th unit

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

$v_{k,i}$ in the i-th word w_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 word durations d_i are obtained from forced alignment. In the recognition stage the syllable rate is computed for each word by a syllable center detector. Taking in account that each word consists of exactly one syllable, word durations may be estimated as a reciprocal number of the average syllable rate. The syllable rate and word duration mismatch results in low duration probabilities, and thus, insertions or deletions can be detected.

Syllable center detection

Current state of the art syllable center detectors utilze the signal's energy [2]. For the SPEECON-Mandarin speech database these methods result in more than 25% insertion and deletion errors. Thus, a novel statistical approach for syllable center detection was created.

A high performance phoneme recognizer based on Long TempoRal Patterns (TRAPs) [1] calculates the phoneme probabilities per 15 ms frame. As most syllables contain exactly one vowel, the sum of all vowel probabilities indicate a syllable center. The per frame estimated probability function has high frequency components that do not contain reliable information. Thus, filtering with a 7 frames hamming window is applied. 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 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.

Experimental results

Experimental test set consists of 662 digit utterances with totally 3368 digits from SPEECON-Mandarin speech database (car environment, hands-free channels). The additive noise is reduced by means of spectral subtraction algorithm in preprocessing unit.

During the first stage, the speech recognition is performed using spectral features. The average WER is 8.3%, in Table 1 this result is shown as "baseline". The rescoring with word duration bigram models in the second stage leads to 7.8% WER, which is 6.3% less than the baseline WER. The normalization of word durations to the relative speech rate brings 10% improvement: WER is reduced to 7.5%. The best results yields the duration model approach with the normalization to the syllable rate: 13.3% reduction of WER (from 8.3% to 7.2%).

Algorithm	WER	relative
Baseline	8.3%	
Word duration bigrams	7.8%	-6.3%
Bigrams+relative speech rate	7.5%	-10%
Normalization to syllable rate	7.2%	-13.3%

Table 1: Experimental results

Conclusion

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 the 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 additional source of information. Syllable rate is suited to detect insertions and deletions, since syllable center positions in hypothesis and speech signal are indirectly compared. In case of insertions or deletions the duration models gives low hypothesis probabilities.

Speech recognition in cars requires practical solutions of many further challenging problems, below it is only a small list of future work: improve detection of syllable centers, realize the rescoring procedure with the search algorithm in a one stage procedure and investigate tonal features.

The author express special thanks to Joachim Hofer (Siemens AG, Corporate Technology) for the allowance to use the syllable center detection algorithm.

References

[1] Jain, P. and Hermansky, H.: Beyond a single critical-band in TRAP based ASR. In Proc. Eurospeech-2003, 437–440.

[2] Narayanan, S. and Wang, D.: Speech rate estimation via temporal correlation and selected sub-band correlation. In Proc. ICASSP-2005, 413–416.

[3] Pfitzinger, H. R.: Two approaches to speech rate estimation. In 6th Australian Int. Conf. on Speech Science and Technology-1996, 421–426.

[4] Wang, C. and Seneff, S.: Lexical stress modeling for improved speech recognition of spontaneous telephone speech in the JUPITER domain. In Proc. Eurospeech-2001, 2761–2764.