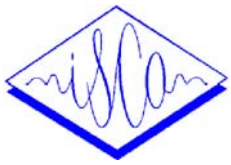

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Use of Prosodic Features for Speech Recognition

Keikichi Hirose[1] *& Nobuaki Minematsu*[2]

[1]Dept. of Frontier Informatics, School of Frontier Sciences
[2]Dept. of Inf. and Commu. Engg, School of Inf. Science and Tech.
University of Tokyo, Tokyo, Japan
{hirose, mine}@gavo.t.u-tokyo.ac.jp

Abstract

Prosody is known to play an important role in human speech perception process. Therefore, there is an increasing need to use prosodic features for the advancement of speech recognition technology. However, prosody is related to various levels of information, from linguistic, para-linguistic, to non-linguistic, and, therefore, its acoustic manifestation is rather complicated with large variations. This fact prevents prosody to be incorporated in speech recognition process. In the current paper, discussions are given on how we can utilize prosodic features, showing our research works as examples. First, an idea of including word likelihood viewed from the accent type into the recognition process is shown. Second, a scheme of using prosody to control the pruning size in the decoding process is given. Prosodic features should be modeled rather differently form segmental features. Lastly, a new language model constructed by including prosodic events is explained.

1. Introduction

In most current speech recognition systems, prosodic features are not utilized, or, we should say, are rather got rid of. In the works related to the *Verbmobil* project, prosodic boundaries were used to constrict search space [1], but they were limited to major boundaries, which might be accompanied by pauses. Also sentence final rise in fundamental frequency (F_0) counter was searched as a signal for interrogation, but contribution of this method to the recognition system was rather limited.

The most naïve idea to incorporate prosodic features in speech recognition process is to combine them with other acoustic features when constructing acoustic models. However, this does not work actually. This is because prosodic features spread to a range wider than phone or syllable, and cannot be handled in the same framework as segmental features.

Each phone has its distinctive segmental features and, from this viewpoint, segmental features can be said as the direct correlates of written language. On the other hand, prosodic features have no direct correspondence to written characters of sentence and are unique to spoken language. Therefore, prosodic features should be handled as those representing structure of human utterance in speech recognition: to construct pronunciation models using prosodic features, to calculate pronunciation likelihood, and to add it in the search process of the speech recognition. The pronunciation modeling will be in word level, such as accent type models for Japanese and tone type models for Chinese, and in phrase level (prosodic boundary modeling). The major problem is how we can construct god modeling and combine the likelihood with acoustic and language likelihood.

In the following sections, some considerations are given on how prosodic features should be processed in continuous speech recognition showing our research works as examples.

2. Prosodic Features and Speech Recognition

There may be roughly two possible ways to use prosodic features in speech recognition process. One is to control acoustic features depending on the prosodic information. Assuming that the vocal transfer functions are related to the fundamental frequencies, their relationship has been investigated. Experiments on cepstrum coefficients showed that there were clear relationship for vowels and other vowel-like sounds (such as nasals) uttered by a speaker [2]. The relation was speaker dependent and some sounds showed no relationship. These should be tackled more before applying the relationship to acoustic modeling.

The other way is to detect prosodic boundaries and to control the speech recognition process. There may be number of reports depending on when and how we should utilize prosodic information. A straightforward way is to find out prosodic boundaries prior to the speech recognition process and to use them to segment input speech. If the method properly works, it may largely increase recognition performance and reduce recognition time. Although several methods have been tried from this viewpoint, they did not work well. The major problems are low boundary detection rates and large variations of boundary positioning by each speaker at each utterance. The boundary detection rates will be improved by totally looking at various prosodic events, such as F_0 contour dips, phone duration lengthening, and so on, and/or by adopting statistical methods, but they cannot be improved very much. The results suggest that prosodic features are not enough; segmental information should also be utilized for boundary detection. Since, in most continuous speech recognizers, two-path algorithm is adopted, phoneme boundary information obtainable from the first stage can be cooperatively used to increase detection rates of prosodic boundaries. The second stage decoding process can be facilitated by the boundary information. Several methods were proposed according to this idea. The extreme case along this line will be automatic prosodic labeling, where the recognition result is known beforehand and the decoder takes alignment between input speech and phonemic transcription. This will not be the recognition, but will be a very important research topic on automatic corpus construction.

The probabilistic factor of the prosodic boundary placing (including erroneous detection) discourages us to use boundary information for speech recognition. However, we should note that major prosodic boundaries appear in most cases and are used by humans to facilitate speech perception process. A possible way will be to use prosodic boundaries

only when they are clearly found, but it throws away a large part of prosodic information. A method is necessary to represent the boundary occurrence in a statistical way and reflect the result onto the likelihood of recognition candidate. To facilitate the recognition process by constricting the search space using boundary information will be an attractive way. However, we should remind that such a constriction is sometimes harmful for the recognition result. A sophisticated answer to the problem was given as an efficient pruning during the decoding process, which is explained in section 4.

There are a number of research works, where prosodic features are used to recognize word accent types of Japanese and tone types of Chinese. Especially, in the case of Chinese, each syllable can have four different meanings depending on the tone types, and, therefore, tone recognition is an important issue in speech recognition. Although each tone type is distinguishable by its F_0 contour, correct tone recognition is not yet realized for continuous speech because of large change in F_0 contour due to tone co-articulation. In the case of accent/tone type recognition of continuous speech, segmental boundary information is mandatory. Using F_0 and power and their delta values (derivatives) as acoustic parameters of HMM's, and taking effects from preceding and following syllables into account, tone recognition rate close to 90% was realized [3]. However, most Chinese continuous speech recognizers still do not incorporate the tone recognition process. This is partly because the recognition task is not so complicated. In the case of Japanese, correct accent type recognition comes rather difficult, because each word includes several syllables and its F_0 contour varies due to various factors, such as word position in a phrase, and so on. An answer to this situation is given in section 3.

Since the current recognition scheme is highly sophisticated, it is not so easy to include the prosodic features into the recognition process. One possible answer for this situation will be the dynamic control of pruning size, as already mentioned (for details, section 4). Another possible answer will be to control the weighting of language model likelihood to acoustic model likelihood, and to construct language modeling taking prosodic boundaries into account. For this, a scheme was developed, which calculated *mora* bi-gram taking accent phrase boundaries into account [4]. Test set perplexity reduction was shown to improve the final recognition results. In this scheme, however, word dictionary was not used in the recognition process. The more realistic way is to use prosodic information in word *N*-gram calculation, which is explained in section 5.

Through the above discussions, it is clear that prosody can be (and should be) taken into account in various instances of speech recognition process in various ways. Figure 1 shows how the prosody can interact into the current 2-path speech recognizer. The prosodic features can be even used after the recognition process; to verify the recognition results from prosodic viewpoint. For this, we have developed a scheme of partial analysis-by-synthesis of F_0 contours [5]. When there are ambiguities in recognition results, each hypothesis is checked by generating F_0 contours (by a speech synthesis method) and comparing them with observed contours. The hypothesis yielding the minimum error in F_0 contours will be the correct recognition.

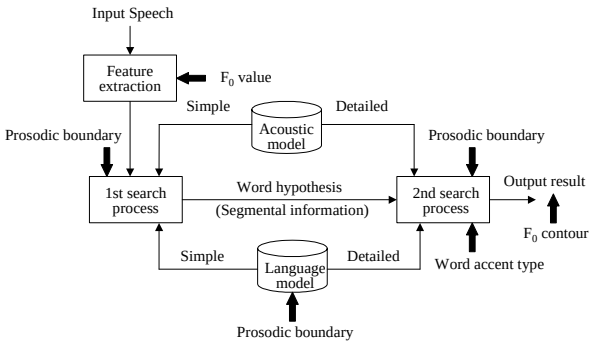

Figure 1: Prosodic information used in various stages of speech recognition.

3. Word likelihood Viewed from Accent Type

Difficulty in recognizing word accent types in continuous speech differs largely depending on the location of the word and the type of accent. Type 1 accent has a rapid downfall after the first *mora* (Japanese basic unit of utterance mostly coinciding with a syllable), and can be identified from other accent types accurately. Moreover, type 1 accent is robust; if a word of type 1 accent locates at the phrase initial, it keeps the type regardless of the following words. Also, our perceptual experiment indicated a human process that detection of type 1 accent facilitates the word search process in our mental lexicon [6]. These facts inspired us a new way to use accent type information in speech recognition.

The method first calculates *mora* F_0 values (in logarithmic scale) for the first and the second *morae* after a pause. The *mora* boundaries and pauses are those obtained in the 1st path of the recognizer *Julius*. The mora F_0 value (F_0*mora*) is defined in our former work so as to coincide with the *mora* pitch value, that humans perceive [7]. In the current experiment, the F_0*mora* is simply calculated as the power-waited F_0 average of vowel part. The F_0*mora* difference F_0*ratio* of the first and the second *morae* is calculated and for type 1 accent words and non-type-1 accent words. The F_0*ratio* distributions for both cases are assumed as *Gaussian* and, when recognizing an utterance, probability of the first word having type 1 accent is calculated and reflected to the likelihood of recognition hypothesis. The experimental results indicated that a better result was obtainable by decreasing the likelihood when the first word was non-type 1 word in the hypothesis but was type 1 from the prosodic features. They also indicated that the likelihood should be untouched, when the accent type of the hypothesis and that from prosodic features coincide [8]. Since the experiment is only preliminary, further research work is necessary before the conclusion.

4. Dynamic Control of Pruning Size Using Prosodic Boundary Information

In large vocabulary continuous speech recognition (LVCSR), reduction of amount of computation is one of important issues. In the *Viterbi* decoding, beam search has widely been applied for reducing the search space. The reduction in the computational complexity is achieved by pruning the improbable paths during recognition while keeping the

globally most likely path active. So the appropriate choice of beam width is very important to lighten computational load without serious increase of search errors during the decoding process.

When tree-structured lexicon is adopted in the *Viterbi* decoding process, the language model (LM) probabilities are factorized into tree nodes by an estimated LM probability for all possible successors. Therefore, the LM scores near the root nodes of the lexicon are usually different from actual LM scores, and, as the result, the potentially best path can be unexpectedly pruned near the root nodes. In order to avoid this situation, beam width should be set wide enough in static pruning (pruning with fixed beam width) with the penalty of increased computational load. However, as the search process comes closer to the word end, the globally most likely path (which should be selected as the decoding process) achieves a relatively high rank in search space because of linguistic and acoustic reasons. The linguistic one is that a greater degree of certainty about word identity is obtained near word ends than word starts. The acoustic one is that the acoustic certainty of a particular word is increased by matching upon more input frames. Consequently, the requirement for necessary beam width to cover the globally most likely path will be lessened at the word ends. In the case of Japanese, the word ends will also be *bunsetsu* ends, which, in many cases, accompanied by prosodic boundaries. Figure 2 shows how the normalized *Viterbi* score changes during the decoding process.

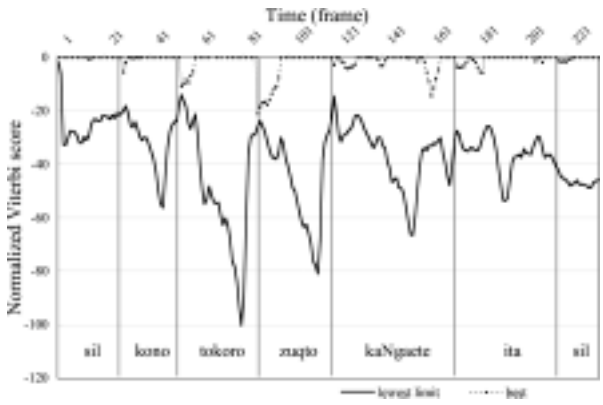

Figure 2: *Viterbi* score change during a decoding process.

Based on these considerations a scheme has developed to control the beam width wider at each prosodic boundary (accent phrase boundary) beginning and narrower toward the next boundary [9]. Beam width is controlled according to the following equations.

$$P(t) = P_{\max}(t) - \lambda(t) \quad (1)$$

$$\lambda(t) = \lambda(0) + \{\lambda_{var}(t) \times W_{LM} \times (N_{word\text{-}end}(t) + N_{phone\text{-}branch}(t))/N_{active}(t)\} \quad (2)$$

$P_{\max}(t)$ is the likelihood of the most likely path at time t and paths with likelihood lower than $P(t)$ are pruned. $\lambda_{var}(t)$ is a function with constantly decreasing its value (values between $\lambda(0)$ and 1) from an accent phrase beginning to its end. W_{LM} is the relative weight of the language model to the acoustic model. $N_{word\text{-}end}(t)$, $N_{phone\text{-}branch}(t)$, and $N_{active}(t)$ respectively denote number of paths with word end, number of paths with root node, and number of all active paths at time t. These terms are added so that the beam width is widened when ratio of LM updating nodes to active nodes is large.

Experiments were conducted by selecting 50 sentence utterances for each of 10 male speakers as the test set from Japanese Newspaper Article Corpus (JNAS). The prosodic boundaries were detected by the method formerly developed by the authors, where microscopic and macroscopic F_0 movements were inspected [10]. To realize the word accuracy rate of 86%, the necessary computational time was 30% less by the proposed dynamic pruning scheme as compared to the fixed-size pruning. The scheme has an advantage over other schemes to utilize prosodic boundary information to speech recognition in that the boundary detection errors may not cause recognition errors. The insertion errors in boundary detection process only increase the search space, and, therefore, by setting the insertion errors being dominant in the boundary detection errors, the proposed scheme will not cause a negative effect on the final recognition result.

5. *N*-Gram Language Modeling using Prosodic Boundaries

The current statistical language modeling, known as *N*-gram, is only for written texts. As outputs of human process of sound production, spoken sentences cannot be fully represented only by written language grammars. Prosodic features are considered to represent structure of speaking, and should also be counted in the language model level. This consideration led us to an idea of separately modeling the word transitions for the two cases: one crossing and the other not crossing accent phrase boundaries [11]. Since counting such transitions requires a large speech corpus, which hardly can be prepared, part-of-speech (POS) *N*-gram was first counted for a small-sized speech corpus for the two cases instead, and then the result is applied to word *N*-gram counts of a large text (newspaper) corpus to divide them accordingly (Fig. 3). Thus, two types of word *N*-gram model can be obtained.

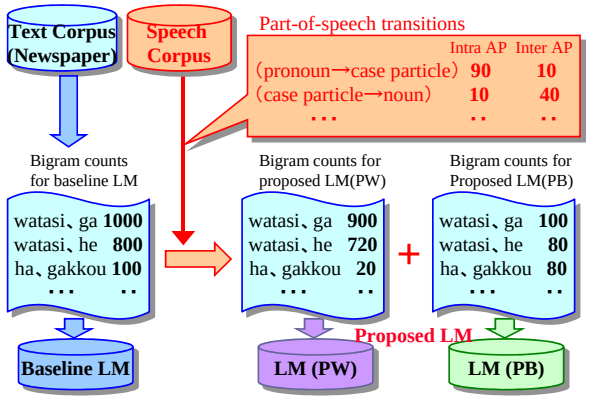

Figure 3: Construction of language models considering accent phrase boundaries.

Using ATR continuous speech corpus by two speakers, perplexity reduction from the baseline model to the proposed model was calculated for the word bi-gram. When accent

phrase boundary information of the speech corpus was used, the reduction reached 11%, and when boundaries were extracted using our formerly developed method based on mora-F_0 transition modeling [12], it still exceeded 8%. The reduction around 5% was still observed for sentences not included for the calculation of POS bi-gram and using boundaries automatically extracted from another speaker's speech. The obtained bi-gram was applied to continuous speech recognition, resulted in a two-percentage improvement of word accuracy from when the baseline model was used.

The major problem of this modeling scheme is that it requires a speech corpus. In Japanese, *bunsetsu* is defined as a basic grammatical unit, which consists of a content word followed by a particle. The content word can be a compound word consisting of two or more content words. The particle part can be null or particles. A *bunsetsu* is also defined as an utterance unit in that a sentence can be uttered by inserting a pause at each *bunsetsu* boundary, and is a unit similar to an accent phrase. This fact inspired us to use *bunsetsu* boundary instead, where, in the recognition process, the boundary was predicted from the word history equal to or longer than *N* words in the case of *N*-gram modeling [13]. The reduction in perplexity was around 8%, and its effect on speech recognition was clear when the available language corpus was rather limited (say 1 year of newspaper).

Since accent phrase boundaries mostly occur at *bunsetsu* boundaries, it will be possible to use the detected accent phrase boundaries as *bunsetsu* boundaries in the decoding process. Accent phrase boundary is detected from prosodic features, and is not predicted from word history. This implies an ability of correcting wrong hypotheses, though it should be proved through the future research.

6. Conclusion

In the preceding sections, use of prosodic features was viewed related to the current speech recognition process. Several works has also been conducted out of this scope, such as detecting utterance structures [14] and finding important words/sentences in an utterance [15]. When a large amount of data is obtainable for training acoustic and language models as in the case of recognizing speech of text-reading style on topics appearing in newspapers, a high recognition performance is obtainable without relying on prosodic features. However, when enough data come hard to be obtained, such as the case of spontaneous speech, role of prosodic features increases in speech recognition. Spontaneous speech may include number of irregularities, such as hesitations, re-statements, and so on, which may degrade speech recognition performance. Prosodic features of these parts are somewhat different from other places (of normal utterance). Detection of utterance irregularities from this viewpoint comes important for the future work.

7. References

[1] Hess, W., et al, "Prosodic modules for speech recognition and understanding in VERBMOBIL," *Computing Prosody*, Springer-Verlag, 361-382, 1997.

[2] Minematsu, N. and Nakagawa, S., "Modeling of variations in cepstral coefficients caused by F_0 changes and its application to speech processing," *Proc. ICSLP, 1063-1066*, 1998.

[3] Zhang, J. and Hirose, K., "Tone nucleus modeling for Chinese lexical tone recognition," *Speech Communication*, to be published, 2004.

[4] Hirose, K. and Iwano, K., "Detection of prosodic word boundaries by statistical modeling of *mora* transitions of fundamental frequency contours and its use for continuous speech recognition," *Proc. IEEE ICASSP, Istanbul*, 3, 1763-1766, 2000.

[5] Hirose, K., "Disambiguating recognition results by prosodic features," *Computing Prosody, edt. Y. Sagisaka, N. Campbell and N. Higuchi, Springer-Verlag*, 327-342, 1997.

[6] Hirose, K., Minematsu, N., and Ito, M., "Experimental study on the role of prosodic features in the human process of spoken word perception," *Proceedings ESCA Workshop on Prosody, Working Papers 41, Lund*, 200-203, 1993.

[7] Ishi, K. C. T., Hirose, K., and N. Minematsu, "Mora F_0 representation for accent type identification in continuous speech and considerations on its relation with perceived pitch values," *Speech Communication*, 41 (2-3), 441-453, 2003.

[8] Murakami, T., Minematsu, N., and Hirose, K., "Experimental examination about the use of prosody at the lexical level in speech recognition," *Report for the Spring Meeting, ASJ*, 1, 191-192, 2004. (in Japanese)

[9] Lee, S., Hirose, K., and Minematsu, N., "Efficient search strategy in large vocabulary continuous speech recognition using prosodic boundary information," *Proc. ICSLP, Beijing*, 4, 274-277, 2000.

[10] Hirose, K., Sakurai, A., and Konno, H., "Use of prosodic features in the recognition of continuous speech," *Proc. ICSLP, Yokohama*, 3, 1123-1126 1994.

[11] Hirose, K., Minematsu, N. and M. Terao, " Statistical language modeling with prosodic boundaries and its use for continuous speech recognition," *Proc. ICSLP, Denver*, 2, 937-940, 2002.

[12] Iwano K. and Hirose, K., "Prosodic word boundary detection using statistical modeling of moraic fundamental frequency contours and its use for continuous speech recognition," *Proc. IEEE ICASSP, Phoenix*, 1, 133-136, 1999.

[13] Chung, S., Hirose, K., and Minematsu, N., "Improvement of N-gram language modeling of Japanese using BUNSETSU boundary information and its application to large vocabulary continuous speech recognition," *Report for the Spring Meeting, ASJ*, 1, 65-66, 2004. (in Japanese)

[14] Hirose, Y., Ozeki, K., and Takagi, K., "Effectiveness of prosodic features in dependency analysis of read Japanese sentences," *Natural Language Processing*, 8 (4), 71-89, 2001.

[15] Inoue, A., Mikami, T., and Yamashita, Y., "Prediction of sentence importance for speech summarization using prosodic features," *Proc. Eurospeech, Geneva*, 1193-1196, 2003.