Rules, but what for?
- Rule description as efficient and robust abstraction of corpora and optimal fitting to applications -

Yoshinori Sagisaka†, Hirofumi Yamamoto‡, Minoru Tsuzaki† and Hiroaki Kato‡
†ATR Spoken Language Translation Research Laboratories ‡ATR International
2-2 Hikaridai Seika-cho Soraku-gun Kyoto, 619-02 Japan

Abstract

Two recent studies are introduced in speech recognition and speech synthesis to reconsider what rules should be looked for spoken language science and technology. To abstract the neighboring characteristics expressed by N-grams, multi-class composite N-grams have been proposed to model POS characteristics and inflectional forms separately. It is shown that statistical clustering can provide more compact and robust description of word neighboring characteristics than conventional N-grams. For speech synthesis, segmental duration modeling has been examined from the viewpoint of perceptual characteristics of duration changes. A series of perceptual experiments have shown the context dependency of sensitivity to duration change. These two examples respectively illustrate how current rules are interpreted to build scientifically acceptable engineering models and remind us of the deeper scientific meaning and limitation of generalization as a rule.

1. Introduction

In speech engineering, speech and language corpora have been widely used as an information source for determining concrete values for various parameters and have been effectively used for the extraction of scientific control principles buried among superficial complexities of data distribution. Corpus-based speech synthesis and statistical speech recognition have successfully provided the framework within which fundamental rules can be harmoniously embedded in a control scheme with their control parameters and where objective testing procedures can predict the performance on unseen input. From the viewpoint of corpus-based technology, rules can be regarded as a set of control models that enable efficient and robust description of data characteristics, and they can be improved through the scientific investigation of underlying mechanisms.

Despite the success of corpus-based approaches in speech technology, one may be skeptical as to whether the relationship between engineering modeling and scientific rules has been properly recognized or not. Has engineering modeling been improved as a scientifically reasonable model? Has scientific rule finding or modeling been reshaped through the feedback of inconsistencies observed in engineering models? In the following sections, we report two recent experiments to show what we have been pursuing at ATR.

First, statistical language modeling is described to show how we refine conventional N-grams engineering models based on the multiple word-classification and word set redefinition. It is shown that this refinement is not derived from simple calculation operations but from the generalization of linguistic word neighboring characteristics. Next, the results of a series of perceptual experiments are briefly introduced to show the context dependency of perceptual sensitivity to duration changes. The results show not only the need of new error measures for segmental duration rules in speech synthesis but also the need for precise definition of temporal markers associated with linguistic units for a scientific understanding of rhythm control. Finally, it is concluded that the harmonic studies of corpora, rules and measures are indispensable for both refining engineering models and pursuing scientific principles in spoken language.

2. Multi-class composite N-grams

2.1 Multi-class modeling

Word N-grams have been widely used as a statistical language model for continuous speech recognition. Though word N-grams are effective and flexible, they require a huge amount of memory compared to grammatical rule expressions. It is quite frustrating to keep redundant information embedded in N-grams and statistically obscure scores for word pairs with small occurrences. To reduce memory size, class N-grams[1] and variable length N-grams-[2] have been proposed. In these models, word-class definition is quite crucial to reduce model size without losing its accuracy.

We have introduced generalization of part of speech (POS) and inflection form (IF) classification for multiple word-categorization [3]. As shown in Figure 1, word-neighboring characteristics are more efficiently expressed by splitting word-neighboring characteristics into preceding and following characteristics independently. The preceding and following characteristics are grouped as “from-word class” and “to-word class” that are intended to generalize POS and IF respectively. Using this multi-class, conventional word class N-gram probability

\[ P(W_n|C_n)P(C_n|C_{n-1}, C_{n-2}, \ldots, C_1) \]

can be expressed by the following equation

\[ P(W_n|C_n)P(C_n|C_{n-1}^{c}, C_{n-2}^{c}, \ldots, C_1^{c}) \]

where \( W_n \), \( C_n \), \( C_n^{c} \) and \( C_1^{c} \) stand for N-th word, word class, from-word class and to-word class respectively.
This multi-class modeling not only enables the differentiation of forward/backward neighboring word attribute but also highly reduces memory size and thus provides statistically reliable probability scores.

2.2 Multi-class word clustering

Though POS and IF can be good word classifications for abstract linguistic rules, they are too rough to quantify statistical word characteristics. Finer classification can be successively obtained by directly clustering neighboring statistical characteristics as follows,

1. Assign one class per word.
2. Calculate a neighboring characteristics vector $V_f(X)$ representing following (preceding) to each class (or word) $X$:
   $$V_f(X) = \{P_f(W_1|X), P_f(W_2|X), \ldots, P_f(W_M|X)\}$$
   where $P_f(W_i|X)$ stand for backward(forward) transition probabilities from class $X$ to word $W_i$ and $M$ is a total word number.
3. Merge two classes whose dispersion weighted with the unigram probability results in the lowest rise.
4. Repeat step (2) until the number of classes is reduced to the desired number.

2.3 Multi-class composite N-gram generation by word unit set expansion

From a statistical modeling viewpoint, it is almost trivial that a linguistically given a priori word unit set would not be the best one. For sparse words, statistically salient word-class would give more reliable probability score. For frequent word sequences, longer word history or units could provide more accurate probability score than conventional word N-grams. We have proposed a generation scheme of a new unit set where word-classes, words and word sequences are employed as units according to data size and distribution [2][3].

Starting from multi-class N-grams as an initial set consisting of words and their word classes, word sequences are added to the unit set to obtain finer model multi-class composite N-grams. When frequent word sequences are added to the unit set, multi-class modeling does not create any new world-classes and no extra

2.4 Speech recognition experiments using multi-class composite N-gram

Multi-class modeling has been evaluated by the comparison of the perplexities among multi-class bi-grams, class bi-grams using same clustering method stated in 2.2 and word bi-grams. For comparison, we generated the model with the same size for from-classes and to-classes. As seen from the generation scheme, the original multi-class modeling with different each size might perform better than the one used for comparison.

As shown in the Figure 2, it turned out that the multi-class bi-grams gave smaller perplexity than the class bi-grams with equal classes every time. Furthermore, its perplexity becomes smaller than that of the word when its class number exceeds 350. These results confirm the superiority of this modeling to conventional word-class modeling and high reliability with smaller size than word N-grams. The recognition experiments also showed that the higher word accuracy with multi-class bi-grams (70.29%) than with the class bi-grams (69.78%) and the word 2-grams (69.05%) under our standard conditions [3][4].

Finally, to evaluate the performance with the multi-class composite N-grams, speech recognition experiments were carried out. As shown in Table 1, the recognition accuracy with multi-class composite N-
grams is higher than the one with the word tri-grams. In this experiment, multi-class composite N-grams were made by adding word sequences with more than twenty occurrences to multi-class N-grams consisting of 200 or 400 classes. Please note that the multi-class composite N-grams with 200 classes need less than 3% the memory of the word tri-grams. These results show the effectiveness of this modeling.

3. Auditory perceptual characteristics for segmental duration distortions

In the generation of rules to assign segmental duration for speech synthesis, rule performance has been measured by the sum of errors between the calculated duration and the observed one in the training data [5][6]. The adoption of this measure is based on the assumptions that each distortion is contextually independent and that the subjective score is monotonically degraded by the increase of the sum of individual temporal distortion. We have examined these two implicit assumptions in the light of the perceptual characteristics: (1) a single duration distortion linearly correlates with the perceived distortion regardless of the attributes of the segment in question, and (2) multiple duration distortions affect the perceived distortion independently of each other.

3.1 Perceptual weighting for temporal distortion

Previous studies reported that perceptual sensitivity to the duration change of a speech segment is affected by the segment quality; e.g., the temporal discrimination capability of vowel segments is higher than that of consonant segments [7]. On the other hand, for the influence of a segment quality, little has been known about the acceptability of a change in the segmental duration, which can be regarded as a more practical measure for the evaluation of a duration.

We have recently shown that listeners’ rating scores of acceptability against changes in segmental duration can be accurately traced by a parabolic curve as shown in Figure 3 [8][9]. It has also been shown that the absolute value of the second-order coefficient of this approximation curve, namely, the vulnerability index is generally larger for vowel segments than that for consonant segments (Figure 4, the left-hand scale). This tendency agrees with previous discrimination studies that vowel duration is more accurately discriminated than consonant duration.

Furthermore, it has been found that this variation in the vulnerability index is highly correlated with the intrinsic loudness of the segments as shown in Figure 4. A non-speech study on temporal discrimination capability, on the other hand, showed that an auditory duration with a high loudness score is more accurately discriminated than a softer duration [10]. This tendency in the temporal discrimination capability agrees with that of the acceptability measure found in Figure 4. All of these results suggest that the correlation observed between the vulnerability index (acceptability measure) and the segment loudness can be accounted for as a reflection of the general characteristics of the auditory perception. To take into account these perceptual characteristics, i.e., the dependency of duration sensitivity on the segment quality, for distortion evaluation, the loudness should be adopted as a weighting factor to approximate human subjective judgement more precisely.

3.2 Perceptual interactions among multiple distortions

A typical example of the interaction among multiple
segmental distortions may be the perceptual compensation effect in the duration of two consecutive segments. This effect is like that when the duration of two segments is modified in a compensatory manner, i.e., to lengthen one segment and to shorten the other by the same size; the total perceived distortion does not become very large in comparison with that expected from the sum of two independent modifications. The perceptual compensation effect between consecutive vowel and consonant duration has been reported for both detectability of the modification[7] and acceptability rating[5]. The compensation effect of this sort indicates that the influence of a duration distortion is not trapped within a segment but may interact beyond segmental boundaries, and also suggests that an evaluation criterion regarding each segmental distortion as independent is not perceptually valid.

Furthermore, it has been found that the amount of the perceptual compensation effect between two consecutive segments inversely correlates with the loudness difference or jump at the segmental boundary, in both detectability and acceptability tasks[12]. The amount of compensation decreased with increasing loudness jump.

A non-speech study also showed that the detectability of a compensatory temporal modification correlates with the loudness jump at the displaced boundary[12]. This suggests that the correlation observed between the perceptual compensation effect of speech and the loudness jump can be accounted for as a reflection of the general characteristics of the auditory perception.

Conventionally, while segmental distortions have been regarded as the changes of a segmental duration, all of the above notions suggest that they can also be regarded as the displacements of segmental boundaries. For describing the relationship among multiple distortions, the latter view appears to be useful.

4. Conclusion

In this paper, a new language model is proposed to reduce model size without perplexity increase using linguistically motivated clustering of neighboring word characteristics. Through the success of recognition experiment with this model, it has been shown that linguistic knowledge is quite useful in statistical modeling and that conventional linguistic notions such as “parts of speech” and “inflectional forms” are to be interpreted from the viewpoint of statistical modeling.

In the second example of human perceptual characteristics on temporal modification of speech, it is shown that the measure commonly used in current modeling does not sufficiently reflect human characteristics. Human characteristics have to be investigated systematically. From this lesson we learn that the corpus-based approach should be properly recognized with the importance of measures and that the abstract notion such as “phoneme duration” needs more profound understanding of our perceptual mechanism with the limitation of cognitive modeling.

As seen in these examples, we find that abstract rules are quite useful in modeling and that they are to be re-examined to adapt to modeling purposes. If we stick to predefined notions and do not adapt them to the modeling purpose, they might not be useful. Formal application of predefined rules and notions without modeling purposes and its constraints cannot involve real rules. Moreover, it has to be noted that rules are useful to describe the underlying principles with small information but are not capable to describe all individual sample characteristics. It is an illusion that all things can be governed by the rules. The rules we know are sometimes too small compared to the unknown whole world: as Newton said, “We are merely a child playing on the seashore before the ocean of the truth.”

References