PHONEME-BASED CONTINUOUS SPEECH RECOGNITION
WITHOUT PRE-SEGMENTATION

GONG Yifan* Jean-Paul HATON*

ABSTRACT

Phonemes, modeled by 2-5 profiles sampled from parameter vector sequences, are used as basic recognition units in a continuous speech recognition system. The segmentation is achieved during the recognition process. Starting from several reliable islands, words decomposed into syllables are hypothesized and then located in a phoneme lattice. The phoneme recognition reliability is used to guide syllable location. In an application for single-speaker spoken Chinese recognition, using a 250-rule context-free grammar with 200 terminals, a 90% sentence and 99% phoneme recognition rate is obtained.

1. INTRODUCTION

Continuous speech recognition without explicit segmentation into basic units, as in dynamic programming [Sak71] or in probabilistic models [Bah83], has achieved very low error rate. However with this scheme it is difficult to include explicit knowledge on phonetics and on phonology. Therefore training and recognition for such system require great computation effort. On the other hand, segmentation-based speech recognition efficiently reduces data rate but segmentation errors may result in fatal failures at the recognition level.

Segmenting speech signal means partitioning the signal in the time scale so that each resulting non-overlapped portion corresponds to an elementary recognition unit. Segmentation is generally realized by detecting the transition of some parameter set. Since the trajectory of continuous speech in the representation space is a non-linearly smoothed version of its elementary component units, in case few difference exists between the end of an unit and the beginning of the next one, the segment boundary cannot be correctly marked or even does not exist. We believe that a reliable segmentation cannot be achieved unless all speech units are recognized.

We present in this paper a continuous speech recognition system using phoneme as basic recognition unit, in which speech units are hypothesized and located during the parsing process. The result of the system is the parsing trees of several sentences having the best recognition scores.

* Equipe Reconnaissance des Formes et Intelligence Artificielle, Centre de Recherche en Informatique de Nancy, Inria-Lorraine/Université de Nancy 1, BP 239, 54506 Vandoeuvre CEDEX France
2. PHONEME MODEL

Speech is a physical realization of a sequence of phoneme-like units which are measurable in the speech signal. Between different realizations, within such a unit, the non-linearity is much smaller than within a word unit and this non-linearity contains information for discriminating some phonemes.

We consider parameter vectors extracted from speech as signal profiles. Each phoneme is modeled by a series of such profiles sampled from the parameter vector sequence of the speech. The number of such profiles depends on the transition of the temporal structure and the duration of the phoneme represented, ranging from 2 to 5. The time scale covered by the model is about 20ms (for some plosives) to 140ms (for some diphthongs). The minimum and maximum durations of the phoneme are also included in the model. The sampled profiles represent the time-varying structure of the phoneme. The references based on this model thus contain critical and necessary information on time variation as well as on spectral distribution of the speech signal. Compared with the use of all consecutive parameter vectors, the redundancy is reduced and the comparison of such references against unknown utterance requires therefore less computation.

In recognition, speech signal is submitted in parallel to a bank of reference filters. The filter outputs $LR_{n,i}$, expressed in likelihood ratio between the unknown utterance $S$ at time $n$, $S_n$, and reference $i$ and averaged to the number of profiles in the reference, are sorted and the best candidate phoneme labels with their likelihood are stored as phoneme symbol lattice for further processing. In summing likelihood ratio related to each profile, linear time warping is used. $LR_{n,i}$ is given by:

$$LR_{n,i} = \frac{1}{M_i} \max_{d_{\text{min}}, i \leq d \leq d_{\text{max}} i} L(S_{n-d/2}, P_i, M_i, d)$$

where $M_i$ is the number of profiles in reference $i$. $d_{\text{min}}$ and $d_{\text{max}}$ are minimum and maximum durations of the phoneme, $P_i$ is the sampled profiles of the phoneme. $l(v_i, v_j)$ is the likelihood ratio of $v_i$ to $v_j$. The distances between parameter vectors of unknown utterance and profiles of each reference are tabulated to avoid re-computation.

3. WORD SPOTTING

Our assumption is that words can be decomposed into phoneme-like units and the coarticulation effects can either be expressed by phonological rules or be treated by multiple phoneme references.

Words are decomposed into syllables by applying rules to the phonetic transcriptions and are located on the basis of syllable location. At the word level, four classes of rules are used to transform a word from its spelling form into syllables and then to a string of phonemes: the notational transform rules that convert a word spelling into a normalized form, the
word-to-syllable rules, the phonological rules that define possible variations of a syllable in a specific context, and rules that decompose triphones into a concatenation of phones or diphthongs. In addition, the word level uses a priori knowledge about phonemes and about the quality of the phoneme classification system used. Each phoneme is also described by its average and its minimum measured durations, reliability of recognition and a list of confusing phonemes together with a probability. These knowledge sources are compiled before recognition.

Instead of the traditional segmentation-based speech recognition which intends to mark the conjunction of two successive units where the coarticulation effect is most significant, we move our attention to the center of each unit. From phoneme classification results we compute the function $S_n = V_n/(V_n+C_n)$, for each profile in which $V_n$ and $C_n$ are respectively the average likelihood ratio of first $M$ vowels and $N$ consonants in the phoneme lattice. $S_n$ is then low-pass filtered to the average syllable uttering frequency. A high value of the resulting function indicates the position of the syllabic center.

A syllable is verified in the phoneme lattice by first locating its vowel and then its consonant(s). In order to locate a phoneme at predicted time $n$, we consider within an interval centered to $n$ the likelihood ratios of all occurrences of the phoneme symbol, taking into account its confusing phonemes. A symmetric triangular window centered at $n$ is then applied to the resulting list of likelihood ratios. The located position is the time position that gives the maximum value of the likelihood ratio. If two or more possibilities in syllable location exist, then they are developed in parallel.

4. RECOGNITION PROCESS

Mixed top-down and bottom-up analysis allows a better use of information obtained from a partially parsed utterance. In dealing with uncertainty in the time position and the existence of a terminal word, when locally we are not sure about the decision to be taken, we develop in parallel all alternatives. In this way we superpose the decisions and propagate properly uncertainty. If an alternative is not the correct final interpretation then, as the recognition proceeds, its quality will degrade rapidly and it will be thus removed from partial solution list.

The recognition process tries to maximize the sum of likelihood ratios in the identified zones of the input speech signal, and give the corresponding parsing trees, under the constraint of all available knowledge at different levels. Our recognizer accepts and interprets context-free semantic grammar written in the form of production rule. The parsing is based on the notion of confidence island. The recognizer maintains in parallel several partial trees, initially constructed from phoneme lattice regions of high likelihood. At each parsing step, for each active island, either top-down tree construction, or bottom-up tree construction (either to the left or right side of the island), or unification of neighboring partial trees, or verification of a hypothesized terminal word or measurement of quality and removal of islands having bad quality may be performed. Beam search [Low80] technique is used. Two functions define the searching behavior. The first one gives the maximum number of islands to be retained and the other determines the minimum acceptable quality of an island, both take the number of identi-
fied terminals in a partial tree as variable. The quality of an island is the sum of likelihood ratios of its identified terminal words divided by the occupied time interval.

5. EXPERIMENTS AND RESULTS

The system was applied to continuous spoken Chinese recognition. All words in Chinese are mono-syllabic.

Speech signal was sampled at 10kHz with 10bit precision and then pre-emphasized. Two parameter sets, LPC (12-th) and MFCC (first 12), are used for comparison. A Hamming window of 25.6ms with 10ms between windows was used. The likelihood ratio in LPC analysis was derived from Itakura distance [Ita75], modified to be gain sensitive [Noc85] while in MFCC this ratio was derived from Euclidean distance. The grammar consisted of 250 rules with 200 terminals. The maximum branch factor was about 100. For 29 vowels and 22 consonants, 260 phoneme reference were manually selected. To create these references, we simply added to the reference list the profiles of which the phoneme had been misrecognized. To train the system in a speaker-dependent mode, we processed about 10 minutes of speech.

The test corpus consisted of 100 utterances with 7 to 19 words. 90% of them were recognized with highest quality. Three utterances were rejected. The remaining seven utterances were among the five first answers. Two of the rejections were caused by insufficient phoneme training and the other by syllable location error. Less then 1% phoneme error was observed. The phoneme classification errors are due to consonant pairs such as (b,m) and (f,s) which are intrinsic in the non-descriptive acoustic analysis under moderate bit rate used for A/D conversion. These results were based on LPC analysis which, under our test condition, gave slightly better recognition rate than MFCC.

References


