THE STRUCTURAL WEIGHTED SETS METHOD FOR CONTINUOUS SPEECH AND TEXT RECOGNITION

Yuri Kosarev, Pavel Jarov, Alexander Osipov

Russian Academy of Sciences
Institute for Informatics and Automation
St. Petersburg
E-mail: kosarev@mail.iias.spb.su

Abstract

In known approaches to speech recognition based on Dynamic Programming (DP) or Hidden Markov Modelling (HMM) time sequences of elements (feature vectors, sounds, letters, etc.) as objects of evaluating or matching are used directly. Both of these approaches have the same demerit: they both can be realised only in the course of the recurrent sequential process and can't be realised in parallel. In addition, the complexity of them are relatively high.

In proposed below Structural Weighted Sets (SWS) method such sequence are reflected first into some structure as a set from relations between its elements and then a recognition is reduced to matching corresponding sets. So in this case a words matching can be realised as a finding an intersection of two sets and evaluating its relative weight. The possibility to carry out a processing in parallel is arisen. The results of simulation are represented.

1 Introduction

The survey of the publications within the speech and text processing field shows that the recognition methods are reduced there mainly to the comparison or the probability appraisal of the sequences of some elements (signal readings, feature vectors, segments, phonemes, letters, words, etc.). The HMM and DP methods were developed to make the recognition and eliminate by this the natural speech variability and the errors of preliminary signal processing too. Both of them are reduced to the sequential recurrent processing ("from left to right", "from right to left", etc.) and do not allow to organise processing in parallel. At the same time it is well known that the information processing within the people brain goes in parallel mainly, the known "simultaneous catching" mechanism prevails in the recognition processes.

HMM methods assume the speech is the Markov process, when the some sound appearance probability is depend on a definite number of previous sounds (usually 1). This assumption is not confirmed by the psycho-acoustics data and does not conform with the code speech model [1].

All said put on to idea that DP and HMM speech processing models are not perfect enough by their essence (it is confirmed by their insufficient effectiveness) and here new original and more adequate approaches are required.

The observations of the people perceiving the speech or text, shows that it is broadly used there recognition mechanisms based on separate fragments of the utterance and their combinations, allowing to appreciate likeness of, it seems, incomparable subjects; hypotheses comparison and selection mechanisms based on some inaccessible for the observation criterion.

In this paper the speech objects comparison method is offered, based not on some sequences, but on their structures comparison. In this case the speech elements sequence are transformed at first to the structural elements

set. Then two words comparison is reduced to producing two sets intersection and its relative weight estimation.

2 From sequence to structure

The essence consists in that an initial sequence of elements (IE) is transformed to the second elements (SE) set. *S*. Every SE reflects some structure fact, for example, *"a directly precedes b"* or *"a precedes b"* and so on. In general, a rule of the transformation IE $\rightarrow$ SE is chosen thus to get this set as most stable to all the possible kinds of the speech signal variability and also to distortion kinds, which can arise during speech generation, transmission and also by its initial processing.

Example: $cat \rightarrow S = \{c, a, t, ca, ct, at\}$.

It can be proved rigorously that this transformation definitively describes an original sequence. The main property of such processing is the possibility to escape the recurrent sequential processing by means of the operation on sets with operations such as **"difference of sets", "sum", "intersection", "power"**, etc.

In a common case to convert the word W represented by the phonemes sequence

$W = p_1, p_2, \ldots p_i, \ldots, p_L$ into the corresponding structure S_W it is necessary to take all the possible structural elements:

$W \rightarrow S_W = S_1 \cup S_2 \cup, \ldots, \cup S_i, \ldots, S_{L-1}$, where $S_1 = \{p_1, p_2, \ldots, p_i, \ldots, p_L\}$ is the set of phonemes containing in this word,

$S_2 = \{(p_1, p_2), (p_2, p_3), \ldots, (p_i, p_{i+1}), \ldots, (p_{L-1}, p_L)\}$

is the set of pairs of phonemes, etc.

It is reasonable to suppose that every SE has own degree of the influence on the word identification. Then for the recognition errors minimisation it is necessary to provide every SE by own weight coefficient K_S. These coefficients must be defined on the training stage. In the simplified version it is possible to assume $K_S = 1$.

3. Word matching

To get a resemblance R of two words W_1 and W_2 containing accordingly L_1 and L_2 phonemes it is necessary:
1) to fulfil transformations $W_1 \rightarrow S_{W_1}$, $W_2 \rightarrow S_{W_2}$;
2) to produce an intersection of this two sets:

$$I(S_{W_1}, S_{W_2}) = S_{W_1} \cap S_{W_2} =$$
$$= S_{1W_1} \cap S_{1W_2} \cup S_{2W_1} \cap S_{2W_2} \cup, \ldots, =$$
$$= R_1 \cup R_2 \cup, \ldots$$

3) to produce a "weighted power" of subset I:

$$R(W_1, W_2) = K_1 |R_1| + K_2 |R_2| + \ldots$$

So this weighted intersection of two structural sets can be as a resemblance measure for words recognition.

This must be normalised by use a normalising coefficient:

$$N = 1/(K_1 |S_1| + K_2 |S_2| + \ldots K_{L_m - 1} |S_{L_m - 1}|,$$

where $L_m = max(L_1, L_2)$.

So normalised resemblance measure is:

$$R^n(W_1, W_2) = R(W_1, W_2) * N$$

To include this processing into early described model for semantic interpretation [2, 3] it is required to transform resemblance measure to distinction:

$$D(W_1, W_2) = 1 - R^n(W_1, W_2)$$

In this case $D(W_1, W_2)$ will be also in normalised form.

4. Possible expansions of SWS-method

Described above SWS-method is rather universal means for speech/text processing. It can be used not only for time sequences as it is shown above, but also for spectrum conversion, for pragmatical processing etc.

4.1. Difference spectrum (DS)

It is required for speech recognition to reduce an initial description dimension, level normalisation, emphasise more important fragments, for example local extremes disposition, their relations, etc.

Let short-time spectrum of speech signal (for segments of 10-20 ms) is represented by set of n spectral components:

$$A = a_1, a_2, \ldots a_i, \ldots a_j, \ldots a_n$$

Then $DS = \{a_{ij}\} = \{sign\ (a_i - a_j)\}$, $i < j$

To reduce an influence of noise by low level signal it is required to use there the threshold δ:

$$a_{ij} = \left| \begin{array}{ll} 0 & \text{if } |a_i - a_j| < \delta, \text{ else:} \\ 1 & \text{if } a_i > a_j, \\ -1 & \text{if } a_i \le a_j \end{array} \right|$$

Such features were used in [4]. As initial representation was 10-band spectrum. From full set $\{i, j\}$, containing $C_{10}^2 = 45$ features were selected the subset of 16 more informative ones.

Lately this features were improved by weight coefficients K_{ij}:

$$a_{ij} = \left| \begin{array}{ll} 0 & \text{if } |K_{ij}\, a_i - a_j| < \delta, \text{ else:} \\ 1 & \text{if } K_{ij}\, a_i > a_j, \\ -1 & \text{if } K_{ij}\, a_i \le a_j \end{array} \right|.$$

The optimal set of K_{ij} was
{0,25; 0,5; 1,0; 2,0; 4; 0},
so full set of different features contained
$45 \times 5 = 225$ features, then 16 "best" were selected and this gave word recognition accuracy improvement from 0,94 to 0,96.

Offered DS-transformation possesses of some advantages:
- the invariance under level change;
- the invariance under spectrum deformation in framework of saving the same formant structure;
- compactness;
- simplification of spectral readings difference calculation

4.2 Pragmatical estimation

There was offered an estimation of the phrases semantic difference [3] based on manipulations on sets. Two phrases were presented in this case as sets of words and semantic difference measure was calculated by means of especial formula with use sets theory machinery. This processing was included into integral model for speech semantic interpretation [3].

4.3 Semantic-syntactical processing

There were presented an associative model for the semantic-syntactical appraisal [6]. The phrase is presented as a set of ordered pairs of words and then associative estimation for phrase is calculated by means of especial formula. This processing was included also into semantic interpretation model.

5. Applications and experiments

This SWS method was applied both for the two-level word recognition and for the handwritten text recognition.

In the first case phoneme sequence obtained by phoneme recognition level was inputted to hypothesizer. Then each hypothesis was transformed to the set of structural relation according part 2. The best hypothesis was selected according maximal normalised resemblance measure. If input phoneme error rate was about 30%, output word error rate was 10%.

In the second case [5] the sentences composed from words of limited vocabulary were inputted. About 10% mistakes like substitutions, insertions, extractions and neighbouring character transpositions were simulated. The model has corrected 98-99% of sentences, for example: *"Ho_ can I get to tthe senter?"* → *"How can I get to the center?"*. Owing to semantic-syntactic information use the model is able to correct mistakes driving inputted words to another words from the same vocabulary, for example: *table* → *able*, *table* → *tale*.

6. Conclusion

Presented SWS method for sounds and letter sequences recognition allows to carry out processing in parallel, so it is possible to join this processing with neurone machinery more closely. This approach can be disseminated also to some kinds of preliminary and high-level speech processing.

References

1. Wozencraft J.M., Reiffen B, Sequential Decoding. - Technology Press and Wiley, New York, 1961.

2. Yu.A. Kosarev. The Model of Oral Speech Semantic Interpretation: Quantitative Processing and Integration of Acoustic, Syntactic, Semantic and Pragmatic Data. Proc. German Acoustics Conference DAGA-94. (Dresden, 1994), pp. 1281-1284.

3. Yu.A. Kosarev, F.M. Kulakov. The Model of Robot Speech Operation: Integration of Signal Processing and Application Area Modelling. Proc. Int. Conf. ICARCV'94, vol. 2. Singapore. 1994, pp.1389-1381.

4. Yu.A. Kosarev, A.N. Osipov. A System for Isolated Words Recognition. Calculate Processes and Structures. A Collected multy-university book, LIAP, 1982, issue 154, pp 92-94.

5. Yuri A. Kosarev. High-Level Processing of the Handwritten Text. Proc. of the SPECOM'96. St-Petersburg. 1996. pp.113-115.

6. Yu.A. Kosarev, P.A. Jarov. Associations Help to Recognize Words. Proc. German Acoustics Conference DAGA'95. Saarbrucken, 1995, vol. 2, pp. 979-982.