MULTY-STREAM WORDS RECOGNITION BASED ON A LARGE SET OF DECISION RULES AND ACOUSTIC FEATURES

Boris Lobanov and Tatiana Levkovskaya

Institute of Engineering Cybernetics, Minsk, BELARUS
Lobanov@newman.bas-net.by

ABSTRACT

This paper investigates the Multi-Stream Automatic Speech Recognition as a part of the general task of, so-called, “collective pattern recognition”. According to this approach, a large number (“collective”) of decision rules as well as a number of different sets of the features are used. The recognition is carried out as a result of the “voting” by the “collective members” (decision rules, acoustic features) in accordance with their “competence degree” (weighting coefficients) in recognition of the certain pattern (word or sub-word).

1. Introduction

Traditionally, only one decision rule and only one set of acoustic features perform speech recognition. One of the main factors limiting traditional ASR applications is the rapid degradation of performance that occurs due to mismatch between the data encountered during recognition and the data used for training. Multi-Stream Automatic Speech Recognition (MS ASR) is a novel approach in speech recognition [1,2] that tries to solve this problem. Most of the studies in MS ASR deal with multi-band speech recognition. In this case, predetermined frequency sub-regions of the speech spectra are treated as distinct sources of information that are processed independently and then combined to perform recognition. Several resent works have suggested that MS ASR gives more accurate recognition, especially in noisy or mismatched acoustic environments that typical for the real life applications [3]. MS ASR is closely connected with the “missing feature” approach [4]. It claims that in noisy speech ASR can often be improved by simply ignoring the parts of the spectral band most effected by noise.

In this paper, unlike the mentioned above, we investigate the MS ASR as a part of the general task of, so called, “collective pattern recognition” first, we believe, investigated in [5]. According to this approach, a large number (“collective”) of decision rules as well as a number of different sets of the features are usually used. The recognition is carried out as a result of the “voting” by the “collective members” (decision rules, acoustic features) in accordance with their “competence degree” (weighting coefficients) in recognition of the certain pattern (word or sub-word in our case).

2. Method

2.1. General description

Basic speech recognition technique used in this study is Continuous Dynamic Time Warping

(CDTW) algorithm described in [6]. The main advantage of the CDTW algorithm is that it not only gives an evaluation of the word (or sub-word) presence in a continuous speech signal (a similarity value S_q for q-th word pattern) but also evaluates the time of its beginning, end and duration T_q. Also, it is robust against adding by the speaker some non-speech sounds such as breathing, lip smacks etc. Basic speech signal processing technique used here is formant analysis described in [7]. An advantage of formant analysis using is that it can potentially provide maximum speaker and channel independence thanks to estimation of articulatory based features such as formant frequencies, amplitudes and voicing degree.

The present paper deals with a task of pattern based discrete words recognition. The following set of decision rules are used for MS ASR:

1. CDTW evaluation of integral similarity for q-th word pattern – $S_q(1)$;
2. Correlation matching of the acoustic parameters of whole word to be recognized and q-th word pattern - $S_q(2)$;
3. Time correspondence way matching of the whole word - $S_q(3)$;
4. Correlation matching of the left sub-word (the left half of the word) and q-th left sub-word pattern - $S_q(4)$;
5. Correlation matching of the right sub-word and q-th right sub-word pattern - $S_q(5)$;
6. Distance matching of the left sub-word and q-th left sub-word pattern - $S_q(6)$;
7. Distance matching of the right sub-word and q-th right sub-word pattern - $S_q(7)$.

The following sub-sets of formant parameters were used for MS ASR:

A. $F_1(t)$, $F_1^{'}(t)$, $F_2(t)$, $F_2^{'}(t)$, $F_3(t)$, $F_3^{'}(t)$ - the formant frequencies and their first derivatives;

B. $A_1(t)$, $A_1^{'}(t)$, $A_2(t)$, $A_2^{'}(t)$, $A_3(t)$, $A_3^{'}(t)$ - the formant amplitudes and their first derivatives;

C. $V(t)$, $V^{'}(t)$, $E(t)$, $E^{'}(t)$ - voicing degree and frame energy of the speech signal and their first derivatives.

MS ASR procedure was organized as a collective of recognizers that includes one “erudite” recognizer and the rest 9 “competent” recognizers. The “erudite” recognizer is organized traditionally: only one decision rule (1) is used, and the full set of formant parameters (A-C) acts as acoustic features perform speech recognition. Word recognition comes through two stages. At the first stage, a certain number of word candidates (decisions) are obtained by the help of the “erudite” recognizer. The decisions are ranking in the order of decreasing of $S_q(1)$ values. The number of candidates – Q is found upon the agree:

$$S_q(1) > 0.7 * S_{\max}(1)\,. \tag{1}$$

At the second stage, these candidates are finally tested then by 9 “competent” recognizers (6 types of decision rules and 3 sets of acoustic features that were found experimentally). The patterns of these word-candidates are time-aligned at the first stage to the input signal by CDTW algorithm, so the word decisions at the second stage are made by a simple linear comparison and take a little time as compared with CDTW matching. The following formulas were used for making decisions according to rules (2)-(7):

$$S_q(2) = \frac{1}{J}\sum_{j=1}^{J} C\big(R(j), P_q(j)\big), \tag{2}$$

where J – acoustic vector parameters size. $C\big(R(j), P_q(j)\big)$ - correlation between acoustic vector parameters R(j) of the word to be recognized and acoustic vector parameters P(q) of q-th word pattern. It defined as:

$$C\big(R(j),P_q(j)\big)=\frac{\sum\limits_t \big(R(j,t)-\overline{R(j)}\big)*\big(P_q(j,t)-\overline{P_q(j)}\big)}{\sqrt{\sum\limits_t \big(R(j,t)-\overline{R(j)}\big)^2 \sum\limits_t \big(P_q(j,t)-\overline{P_q(j)}\big)^2}}\,, \tag{3}$$

where $R(j,t)$ – the value of j-th parameter in *t*-th frame of speech signal; $P_q(j,t)$ - the value of the j-th parameter in t-th frame of q-th speech pattern; $\overline{R(j)}$ and $\overline{P_q(j)}$ - average values of the j-th parameter of speech signal and pattern accordingly.

$$S_q(3) = C(W_r, W_q)\,, \qquad (4)$$

where W_r - the time correspondence way function between the r-th input speech signal and q-th pattern; W_q - the same function defined during the training for q-th pattern.

$S_q(4)$ and $S_q(5)$ are calculated according to (2), (3), where *t* is restricted from 1 to $T_q/2$ for $S_q(4)$, and from $T_q/2$ to T_q for $S_q(5)$.

$$S_q(6) = \sum_t \left(1 - \frac{1}{J} \sum_{j=1}^{J} \left| R(j,t) - P_q(j,t) \right| \right), \qquad (5)$$

where *t* is restricted from 1 to $T_q/2$.

$S_q(7)$ is calculated as $S_q(6)$ for *t* from $T_q/2$ to T_q.

For making decisions by formulas (2)-(7) the full set (A-C) of formant parameters is used. In the rest three competent recognizers the only one decision rule (2) is used, by combining with sub-sets of formant parameters: A, B or C. As a result, three additional estimations of integral similarity for q-th word pattern are obtained: $S_q(8)$, $S_q(9)$, and $S_q(10)$.

2.2. Procedures of collective decision

There are many possibilities to organize at the second stage a procedure of collective decision depending on the recognition task.

In general, the final estimation of spoken word similarity to each of q-th candidate S_q^f may be calculated as:

$$S_q^f = \sum_v b_q(v) * S_q(v)\,, \qquad (6)$$

where $b_q(1)$ - $b_q(10)$ are competence coefficients for the certain decision rule for the each of q-th word pattern. The values of these coefficients may vary from 0 to 1.

In particular, the collective decision may be determined by the first V most competent decision rules, or even by the only most competent one.

2.3. Competence coefficients estimation

The set of competence coefficients $b_q(v)$ for each of v-th recognizer concerning each of the q-th word for the given vocabulary is obtained at the training stage by using the formula:

$$b_q(v) = \left(S_q(v) - \max_{n \neq q} \; S_n(v) \right) \Big/ \left(\sigma_q(v) - \sigma_n(v) \right), \qquad (7)$$

where $S_q(v)$ - a mean similarity of each word’s samples from the training data to the q-th word pattern; $S_n(v)$ - a mean similarity of each word’s samples from the training data to the n-th ($n \neq q$) word pattern; $\sigma_q(v)$ – a dispersion of similarity of each word’s samples from the training data to the q-th word pattern; $\sigma_n(v)$ – a dispersion of similarity of each word’s samples from the training data to the n-th ($n \neq q$) word pattern.

As a result of calculation a rectangular matrix of b_q - (V*Q) is obtained.

The formula (7) suggests that the bigger the difference between similarity distributions of q-th word and the most similar to it n-th word, the bigger the competence of the v-th recognizer for the q-th word, and also the smaller their dispersions, the bigger competence.

3. Experiments and discussions

The MS ASR procedure was tested using the following database. The signals were collected over mobile phone from 20 speakers (10 male, 10 female). The text corpus used in recordings includes names of digits and letters (letter A pronounced as Anna, B as Berta, etc). From each speaker 5 calls from different locations (office, street) have been recorded. The signals have been stored in wav-format (16 bits, sampling frequency 8kHz). Half of the samples from the database was used for training, while another half - for testing. Comparative results of the testing, that are traditional (using “erudite” recognizer only), and suggested MS ASR procedure of words recognition are presented in the Table 1 for a single cluster used for each word’s pattern and in Table 2 for multi-cluster patterns (average 2.5 clusters for a word).

	Numbers	Letters
Erudite recognizer	**6.7**	**11.2**
Compitente recognizer	**5.3**	**9.1**

Table 1. Recognition error rate (%) for single-cluster patterns

	Numbers	Letters
Erudite recognizer	**1.91**	**2.85**
Compitente recognizer	**1.67**	**2.51**

Table 2. Recognition error rate (%) for multi-cluster patterns

As one can see from the tables there is a significant improvement in errors rate, especially, for single-cluster recognition. This is a preliminary result, and there are plenty rooms for future investigations concerning to choosing of different sets of decision rules, acoustic parameters, rules for competence degree estimation and rules for collective voting.

References

1. H. Bourland, S. Dupond. A New ASR Approach Based on Independent Processing and Recombination of Partial Frequency Bands. In Proc. of ICSLP, Philadelphia, USA, pp.426-429, 1994.
2. A. Janit, D. Ellis, N. Morgan. Multi-Stream Speech Recognition: Ready for Prime Time? In Proc. of the 6th European Conference on Speech Communication and Technology - EUROSPEECH ’99, Budapest, pp. 591-594, 1999.
3. S. Okawa, T. Nakajima, K. Shirai. A Recombination Strategy for Multy-Band Speech Recognition Based on Mutual Information Criterion. In Proc. of the 6th European Conference on Speech Communication and Technology - EUROSPEECH ’99, Budapest, pp. 603-606, 1999.
4. R. Lippmann, B. Carlson. Using Missing Features Theory to Actively Select Features for Robust Speech Recognition with Interruptions, Filtering and Noise. In Proc. of the 5th European Conference on Speech Communication and Technology - EUROSPEECH –97, pp. 97-40, 1997.
5. L. Rastrigin, R. Epenshtein. Method of Collective Pattern Recognition. Energy Publisher, Moscow, 1981. (in Russian)
6. B. Lobanov, T. Levkovskaya. Continuoes Speech Recognizer for Aircraft Application. In Proc. of SPECOM'97. Napoca, Romania, 1997.
7. B. Lobanov, T. Levkovskaya, I. Kheidorov. Speaker and Channal-Normalized Set of Formant Parameters for Telephone Speech Recognition. In Proc. of the 6th European Conference on Speech Communication and Technology - EUROSPEECH ’99, Budapest, pp. 331-334, 1999.