



Combining Chinese Spoken Term Detection Systems via Side-information Conditioned Linear Logistic Regression

Sha Meng, Wei-Qiang Zhang, Jia Liu

Tsinghua National Laboratory for Information Science and Technology
 Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China

mengsha@limsi.fr, {wqzhang, liuj}@tsinghua.edu.cn

Abstract

This paper examines the task of Spoken Term Detection (STD) for the Chinese language. We propose to use Linear Logistic Regression (LLR) to combine various Chinese STD systems built with different decoding units, detection units, features and phone sets. In order to solve the missing-sample problem in STD system combination, side-information reflecting the reliability of the scores for fusion is used to condition the parameters of the standard LLR model. In addition, a two-stage combination solution is proposed to overcome the data-sparse problem. The experimental results show that the proposed methods improve the overall detection performance significantly. Compared with the best single system, a relative 11.3% improvement is achieved.

Index Terms: Spoken Term Detection (STD), system combination, Linear Logistic Regression (LLR), side-information

1. Introduction

Improving accessibility of the overwhelming amounts of speech data available today necessitates the development of robust Spoken Term Detection (STD) techniques, which aim to locate user-defined terms rapidly and accurately in large audio archives. A significant amount of work have been done for STD and related topics such as Spoken Document Retrieval (SDR) and Keywords Spotting (KWS) for English and, to a lesser extent, other languages. This paper focuses on the Mandarin Chinese STD task.

The combination (or fusion) of multiple systems based on machine learning algorithm has been shown to significantly boost the performance in various speech-related tasks, for instance, speaker recognition and language recognition. In STD systems, combination has also been studied. [1] used word lattices for In-Vocabulary (INV) term detection and phonetic transcripts for Out-Of-Vocabulary (OOV) term detection separately. [2] combined a word-based system and a phonetic system with linear weights. [3] proposed to use a ROVER-like method to combine two different systems. For Chinese tasks, [4] combined the word-based system and the syllable-based system with tuned weights. [5] combined the word-based system and the character-based system in a similar way. Till now, STD system fusions have only been proposed for a small number of sub-systems, typically using a simple linear combination. In this paper, we propose to build a larger number of various sub-systems and then combine them automatically.

In studies on English STD systems, there have been numerous discussions whether words or subwords (phonemes) should

be used as indexing units. Chinese offers more choices for selecting the units, including words, characters, syllables and toneless syllables. In previous work, we studied the performances of using different units for Chinese STD task [6]. Besides the units, various features, acoustic models and language models could also be the sources of building complementary systems. The advantage of combining these systems have been proved on Speech-To-Text (STT) tasks.

Using different units or models, various STD systems can be built, each having individual strengths. We consider to combine these systems at the score level via Linear Logistic Regression (LLR) which is a successful solution for score fusion in speaker recognition and many other tasks [7]. As STD systems always have distinct result sets, missing-sample problem is serious in the combination. We assign small probabilities to the missing samples and use side-information [8] to reflect the reliability of the scores. The standard LLR parameters are then conditioned on the side-information (SI-LLR). In addition, a two-stage combination solution is proposed to solve the data-sparse problem when estimating the large set of SI-LLR parameters.

2. Chinese spoken term detection systems

2.1. Metric

The detection performance is measured in terms of an application model by assigning a value (V) to each correct output (*hit*) and a cost (C) to each incorrect output (*false alarm*). The overall Occurrence-Weighted Value (OWV) is defined in [9] with $C/V = 0.1$, as:

$$OWV(\theta) = \frac{1}{N_{true}} [N_{hit}(\theta) - \frac{C}{V} \cdot N_{fa}(\theta)] \quad (1)$$

N_{hit} , N_{fa} and N_{true} are the numbers of *hits*, *false alarms* and the real occurrences respectively. θ is the confidence threshold. The OWV with a selected threshold is Actual-OWV (AOWV), while the Maximum-OWV (MOWV) is achieved at the optimal working point¹.

2.2. Lattice-based STD system

The baseline STD system is built with a lattice-based approach. An LVCSR decoder is used to generate lattices. The original lattices are then represented as posterior lattices by calculating the posterior probabilities of nodes and arcs in the lattices [6].

¹The other metric defined in [9], Term-Weighted Value (TWV) computes term-specific values and finally averages these values over all terms in order to be less susceptible to being biased toward frequently occurring terms. There are no high-frequency terms in our task.

The first author is now working at LIMSI-CNRS, France

A posterior lattice stores 4 fields with each arc: $S[a]$, $E[a]$, $I[a]$ and $P_{arc}[a]$, and 2 fields with each node: $t[n]$ and $P_{node}[n]$. $S[a]$ and $E[a]$ are the start node and the end node for arc a , and $I[a]$ is the word identity. For a term Q , the posterior probability at time (t_s, t_e) is calculated as:

$$p(*, t_s, Q, t_e, * | lattice) = \sum_{\substack{\pi=(a_1, \dots, a_K): \\ t[S[\pi]]=t_s \wedge t[E[\pi]]=t_e \wedge I[\pi]=Q}} \frac{P_{arc}[a_1] \cdots P_{arc}[a_K]}{P_{node}[S[a_2]] \cdots P_{node}[S[a_K]]} \quad (2)$$

A hard decision whether the hypothesis is correct or not is necessary for applications. AOWV is a metric related to the chosen working point. In Equation (1), N_{hit} and N_{fa} are both functions of the confidence threshold θ . When θ is large, the detected hypotheses is more likely to be a *hit*, therefore increasing the value would cause more N_{hit} lost than $\frac{C}{V} N_{fa}$, resulting in the degradation of OWV. On the other hand, the increasing of θ in the low value region improves OWV, since when the posterior probability is small, the term $\frac{C}{V} N_{fa}$ decreases faster. The maximum OWV is expected when $\frac{C}{V} \Delta N_{fa}$ equals to ΔN_{hit} , giving the threshold at:

$$\theta = \frac{\Delta N_{hit}}{\Delta N_{hit} + \Delta N_{fa}} \geq \frac{C}{C + V} \quad (3)$$

2.3. Chinese STD with various subwords

Compared with English, Chinese has distinctive characteristics. A word is graphemically made up of characters. Each character is pronounced as a syllable. One character may map to different syllables in different contexts and many characters could share the same pronunciation. Meanwhile, Chinese is a tonal language. There are up to five tone types for each base syllable.

Characters, syllables and toneless syllables can be used to build subword-based STD systems in two ways, by directly subword-based decoding or by higher-level lattice conversion.

2.3.1. Subword-based STD systems

Subwords are used for both decoding and detection instead of words. In the decoding process, subword-based language models (LM) including character LM, syllable LM and toneless syllable LM are built for each individual system. Lattices with different units are generated respectively. In the detection process, the term is first converted to a subword sequence, then searched in the corresponding lattices.

The interpretations of the probabilities calculated by Equation (2) are different among systems. For a word-based system, the probability means how likely the word sequence composing the term exists in the segment, while for a syllable-based system, the probability corresponds to the syllable sequence. Since the STD task aims to detect terms in word-level, the word-based probability is more precise. However, since speech recognition is an error-prone process, more tolerant subword-level matching has the potential to achieve higher recall rate. For instance, syllable-matching may recall an occurrence that is misrecognized as another word with the same pronunciation. Meanwhile, the accuracies of the acoustic and language models also have impact on the posterior probability calculation.

2.3.2. Lattice conversion

Word-based decoding normally outperforms subword-based decoding due to a better LM, while subword-based detection has higher recall which may favor the overall performance. Hybrid

systems using words as decoding units and subwords as detection units are performed by lattice conversion.

Converting word lattices to subword lattices involves an arc splitting process, as a word may be split into multiple subwords. The process is trivial with posterior lattices: replacing the word arc with a sequence of connected subword arcs, with all the subword arcs having the same arc posterior as the word arc, and equally splitting the time period [6]. Converting character lattices to syllable lattices or converting syllable lattices to toneless syllable lattices is a simple label replacing operation on lattices.

When words or characters have multiple pronunciations, the best pronunciation information for each arc provided by the recognizer is used to determine the corresponding syllables.

2.4. Systems with various features and phone sets

Systems with different features and models are as well expected to bring in performance improvement by combination. In our work, MFCC and PLP are used as front-end separately, together with two distinctive phone sets: Initial-Final phone set and segmented tonal phone set [10]. With each pair of feature and phones set, systems can be built with different decoding and detection units as described in Section 2.2 and Section 2.3.

3. System combination by side-information conditioned linear logistic regression

Each system described in Section 2 generates a set of term hypotheses. All the hypotheses are used for fusion at the score level, without being pruned by a threshold. The j -th hypothesis of the i -th sub-system is represented by a tri-tuple $\{st_{ij}, et_{ij}, p_{ij}\}$, $1 \leq i \leq N$, $1 \leq j \leq M_i$. st_{ij} , et_{ij} and p_{ij} are the start time, the end time and the confidence score respectively. M_i is the number of hypotheses from the i -th sub-system and varies for different i .

The combination of the hypotheses from multiple STD systems is different from a common score-fusion task in that the hypothesis may occur at any time, thus st_{ij} and et_{ij} are quite different among systems. We first group together the hypotheses which have time overlap. Each group then generates a new result by fusing the information of all its members. The new hypothesis has boundaries as $st = \min_i \{st_i, 1 \leq i \leq N\}$ and $et = \max_i \{et_i, 1 \leq i \leq N\}$, where st_i and et_i are the start time and the end time of the hypotheses belonging to this group.

Another problem is that one sub-system does not generate hypotheses for every group. Hence there are less than N scores that can be used for fusion in each group. This is defined as the *missing-sample* problem in the STD system combination task. It is caused by the pruning in the decoding process that the potential hypothesis was pruned due to its low probability. We assign to these missing samples a small approximate probability which equals to the lowest probability of the hypotheses found from the lattices.

Scores from sub-systems, including both lattice-based posterior probabilities and the assigned approximate probabilities, are used for fusion. Scores can be combined in a linear way as

$$p = \sum_{i=1}^N \beta_i p_i, \text{ where } \beta_i \text{ are the weights for sub-systems.}$$

3.1. Linear Logistic Regression (LLR)

Given the input feature vector $\mathbf{X} = (x_1, x_2, \dots, x_N)$, where $x_i = \log[p_i/(1 - p_i)]$, Linear Logistic Regression (LLR) model uses Equation (4) to calculate the probability of the sam-

ple belonging to a class \mathbf{C} . In the STD combination task, \mathbf{C} is the *hit* class. $\{\beta_i, 1 \leq i \leq N\}$ are the weights for sub-systems, while β_0 is the intercept.

$$p(\mathbf{C}|\mathbf{X}) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^N \beta_i x_i)}} \quad (4)$$

As the missing error and false-alarming error have different costs, we optimize the object function in Equation (5) to estimate the vector $\mathbf{B} = (\beta_0, \beta_1, \dots, \beta_N)$ with the FoCol toolkit². LLR model interprets probability well, the threshold selected by Equation (3) is expected to work better than that in a system using lattice-based posterior probability as confidences.

$$O(\mathbf{B}) = V \cdot \sum_{\text{hit}} \log(p) + C \cdot \sum_{\text{false-alarm}} \log(1-p) \quad (5)$$

3.2. Side-information conditioned LLR (SI-LLR)

Standard LLR assigns a weight β_i to each sub-system for combination. The weight vector is static for all samples. It is unfair to use the same weight for the scores calculated based on lattice-based posterior probabilities and the arbitrary small score assigned to the missing samples. To make the weights for sub-systems dynamic, we consider to condition them by the side-information reflecting the reliability of the scores for fusion.

Each sample $\mathbf{X} = (x_1, x_2, \dots, x_N)$ is accompanied with a side-information vector $\mathbf{S} = (s_1, s_2, \dots, s_N)$, where s_i is assigned to the value s_{real} or s_{miss} which tells whether x_i is a real score or an arbitrary score. Let the vector $\mathbf{B} = \{\beta_i, 0 \leq i \leq N\}$ be the function of \mathbf{S} that $\mathbf{B} = (\beta_0(\mathbf{S}), \beta_1(\mathbf{S}), \dots, \beta_N(\mathbf{S}))$. Therefore, the vector \mathbf{B} becomes an $(N+1) \times N$ matrix for various \mathbf{S} .

$$\mathbf{B} = \begin{Bmatrix} \beta_{01} & \beta_{02} & \dots & \beta_{0N} \\ \beta_{11} & \beta_{12} & \dots & \beta_{1N} \\ \dots & \dots & \dots & \dots \\ \beta_{N1} & \beta_{N2} & \dots & \beta_{NN} \end{Bmatrix} \quad (6)$$

With a decided $\mathbf{S} = (s_1, s_2, \dots, s_N)$, a weight for the i -th sub-system is $\beta_i = \sum_{j=1}^N \beta_{ij} s_j, 1 \leq i \leq N$. It denotes that the weight for the i -th sub-system is not only affected by s_i which reflects the reliability of the corresponding score, it is also affected by $\{s_j, i \neq j\}$ reflecting whether the scores from other sub-systems are confident.

3.3. Two-stage combination

From a standard LLR model to a side-information conditioned LLR model, an N -dim weight vector is extended to an $N \times (N+1)$ matrix, which causes data-sparse problem in the estimation of \mathbf{B} when N is large. We consider to combine the sub-systems in two stages.

Simple linear combination with equal weights is appropriate for *similar* sub-systems that have the same probability interpretation and their performances are close. In the first stage, *similar* sub-systems built with both the same decoding units and the same detection units are linearly combined with equal weights. The hypotheses generated by each group are combined with SI-LLR method in the second stage. SI-LLR is used in the final stage in the consideration that it has advantages in interpreting probabilities.

²<http://www.dsp.sun.ac.za/fibrunner/focal/index.htm>

4. Experiments and results

4.1. Setup

We evaluate the proposed methods on a 4-hour Mandarin Chinese spontaneous set. Acoustic models are trained on 154-hour reading-style speech plus 148-hour spontaneous speech. Trigrams with different units are trained from a text corpus containing about 2.1 billion characters. A vocabulary with 68,933 words is used for LVCSR recognizer and the word breaker for both language model training and term pre-processing. There are 3979 terms used for testing. OOV terms are first segmented into INV sequences³. We build three LVCSR systems with different features and phone sets. Setup details and error rates are showed in Table 1. The CER and the GER which measures the richness of the lattices are not much different among systems⁴.

Table 1: LVCSR setups and error rates. CER: character error rate. GER: graph error rate.

id.	feature	phone-set	CER(%)	GER(%)
S	MFCC	Initial-Final	37.0	16.2
S _{st}	MFCC	Segmented Tonal	37.8	17.3
S _{plp}	PLP	Initial-Final	38.3	17.0

4.2. Various Chinese STD systems

Table 2 shows the performances of individual systems built with different units. AOWV, MOWV and the recall of lattices (REC) are reported. The recall of lattices is the percentage of term occurrences that can be found in the lattices. It is the upper-bound of the recall at a fixed working point. Word-based systems with different setups (S1, S1_{st}, S1_{plp}) have similar performances. Since the systems built in the same scheme with S, S_{st} and S_{plp} respectively perform similarly and have the same trends, we only show the results for the systems built with S.

Table 2: STD results of single systems. S1, S1_{st} and S1_{plp} are the word-based systems. Lattices in S1 are converted to lattices with different units for S1.1, S1.2, S1.3. REC: recall of lattices.

id.	index	AOWV	MOWV	REC(%)
S1	word	0.5501	0.5796	71.2
S1.1	=>character	0.5719	0.6007	73.2
S1.2	=>syllable	0.5702	0.5830	75.2
S1.3	=>toneless syl.	0.5245	0.5277	77.6
S2	character	0.4880	0.5139	70.4
S2.1	=>syllable	0.4791	0.4895	73.5
S2.2	=>toneless syl.	0.4354	0.4386	77.1
S3	syllable	0.5145	0.5278	76.0
S3.1	=>toneless syl.	0.4800	0.4812	78.4
S4	toneless syl.	0.4246	0.4306	72.9
S1 _{st}	word	0.5520	0.5883	68.2
S1 _{plp}	word	0.5358	0.5676	70.5

Among all the 10 systems built with S, the system decoding with words and detecting with characters (S1.1) gives the best AOWV at 0.5719. It is because word-based decoding generates more accurate posterior probabilities, and character-based detection does not suffer from the disagreement of the word-segmentation between the recognition results and the reference transcripts. Syllable-based systems have higher recall since they are more tolerant to recognition errors. However, the confidences are less accurate to achieve better AOWV.

³With all common characters in the vocabulary, a Chinese word can always be converted into an INV sequence.

⁴We do not intent to compare different features and phone sets here.

[6] had a different observation that S1.3 gave the best performance when systems were evaluated on Figure of Merit (FOM). It is because the recall of a lattice-based STD system is usually dominant for FOM. AOWV is more sensitive to false alarms and the accuracy of the confidence is much more important. Higher recall do not always favor AOWV⁵.

4.3. System combination

The combination methods are evaluated by a two-fold cross-validation on the test set. Table 3 shows the performance of the simple methods which averages the probabilities of all sub-systems, the LLR method and the SI-LLR method ($s_{real}=0.8$, $s_{miss}=0.2$), together with the two-stage combination.

Table 3: System combination results. REC: recall of lattices.

method	AOWV	MOWV	REC(%)
<i>3 systems: SI+SI_{st}+SI_{plp}</i>			
best single (SI _{st})	0.5520	0.5883	68.2
simple	0.5983	0.6203	
LLR	0.6020	0.6077	77.7
SI-LLR	0.6198	0.6234	
<i>10 systems: SI+SI.1+...+S4</i>			
best single (SI.1)	0.5719	0.6007	73.2
simple	0.3975	0.5550	
LLR	0.5846	0.5906	84.5
SI-LLR	0.6118	0.6120	
<i>30 systems</i>			
best single (SI.1 _{st})	0.5741	0.6058	74.7
1st stage	0.6223	0.6449	79.8
2nd stage	0.6506	0.6522	88.4

4.3.1. Systems with different features or phone sets

We first combine S1, S1_{st} and S1_{plp}, which are built with different features or phone sets via the same word-based method. The results are shown in the top block of Table 3. Compared with the best single system, the simple method, the LLR method and the SI-LLR method improve the AOWV by 8.4%, 9.1% and 12.3% respectively, together with a higher recall.

The MOWV of the simple method is very close to that of SI-LLR, however the AOMV is far behind. This is because scores are well calibrated in LLR and SI-LLR, therefore the AOMV is closer to the MOWV. On the other hand, to combine similar systems only differing in features or phone sets, the simple method works well if the working point is not taken into account. This is supportive for using the simple method in the first stage of the two-stage combination, where the threshold is not necessary.

4.3.2. Systems with different decoding units or detecting units

The middle block of Table 3 shows the results of combining the first 10 systems in Table 2, which are built with various units, diverse in probability interpretations and accuracies. The simple method hurts the overall performance due to the worse sub-systems. It was a different conclusion that the simple method benefited the FOM-evaluated KWS task reported in [6], due to the much higher recall of the combined system.

The LLR method improves the AOWV from 0.5719 to 0.5846. The decrease of MOWV is due to the more serious missing-sample problem when combining more systems. The

⁵However, too small C/V will again result in the dominant of the recall

SI-LLR method improves the AOWV further to 0.6118, 7.0% relatively from S1.1, together with a higher MOWV.

4.3.3. Two-stage combination of 30 sub-systems

The bottom block in Table 3 shows the results for the two-stage combination of all the 30 systems. The best system after the first stage gives an AOWV at 0.6223, coming from the simple combination of the three systems decoding with words and detecting with characters, using different features and phone sets. The second-stage combination by SI-LLR further improves AOWV to 0.6506, relatively 11.3% higher than the best single system.

5. Conclusion

In this paper, we built lattice-based Chinese STD systems with various decoding units, detection units, features and phone sets, then combined them via the LLR method. In order to solve the missing-sample problem, side-information reflecting the reliability of the scores for fusion was used to condition the the standard LLR parameters. In addition, a two-stage combination scheme was proposed to overcome the data-sparse problem. The experimental results showed the proposed methods improved the performance of Chinese STD system significantly.

Besides the systems and side-information used in this paper, more heterogeneous STD systems can be taken into account and more complicated side-information can be used to further benefit the combined system under the proposed scheme.

6. Acknowledgments

We would like to thank Microsoft Research Asia for providing the data and the baseline engine. This work was supported by NSFC (No. 60572083, No. 60931160443), and in part by the 863 Program (No. 2006AA010101, No. 2007AA04Z223, No. 2008AA02Z414, No. 2008AA040201).

7. References

- [1] Miller, D. R. H., et al, "Rapid and accurate spoken term detection", Proc. Interspeech'2007, Antwerp, 2007.
- [2] Yu, P. and Seide, F. "A hybrid word/phoneme-based approach for improved vocabulary-independent search in spontaneous speech", Proc. ICLSP'2004, Korean, 2004.
- [3] Tejedor, J., Wang, D., et al, "A posterior probability-based system hybridisation and combination for spoken term detection", Proc. Interspeech'2009, Brighton, 2009.
- [4] Chen, B., "Voice retrieval of Mandarin broadcast news speech", International Journal of Pattern Recognition and Artificial Intelligence. 20(1), 2006.
- [5] Lo, W.K., Meng H., Ching, P.C., "Cross-language spoken document retrieval using HMM-based retrieval model with multi-scale fusion", ACM Transactions on Asian Language Information Processing, Vol.2, iss.1, 2003.
- [6] Meng, S., Yu, P., Seide, F., Liu, J., "A study of lattice-based spoken term detection for Chinese spontaneous speech", Proc. ASRU'2007, Kyoto, 2007.
- [7] Brummer, N., Burget, L., et al, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST Speaker Recognition Evaluation 2006", IEEE Transactions on Audio, Speech, and Language Processing, 15(7), 2007.
- [8] Ferrer, L., Graciarena, M., et al, "System combination using auxiliary information for speaker verification", ICASSP, Proc. ICASSP'2008, Las Vegas, 2008.
- [9] "The spoken term detection (STD) 2006 evaluation plan": <http://www.itl.nist.gov/iad/mig/tests/std/2006/docs/std06-evalplan-v10.pdf>
- [10] Huang, C., Shi, Y., Zhou, J. L., et al, "Segmental tonal modeling for Phone Set Design in Mandarin LVCSR", Proc. ICASSP'2004, Montreal, Canada, 2004.