Low Complexity LID using Pruned Pattern Tables of LZW

S.V. Basavaraja and T.V. Sreenivas

Department of Electrical Communication Engineering,
Indian Institute of Science, Bangalore - 560 012, India.
basavaraj@ece.iisc.ernet.in, tvsree@ece.iisc.ernet.in

Abstract

We present two discriminative language modelling techniques for Lempel-Ziv-Welch (LZW) based LID system. The previous approach to LID using LZW algorithm was to directly use the LZW pattern tables for language modelling. But, since the patterns in a language pattern table are shared by other language pattern tables, confusability prevailed in the LID task. For overcoming this, we present two pruning techniques (i) Language Specific (LS-LZW)- in which patterns common to more than one pattern table are removed. (ii) Length-Frequency product based (LF-LZW)- in which patterns having their length-frequency product below a threshold are removed. These approaches reduce the classification score (Compression Ratio [LZW-CR] or the weighted discriminant score [LZW-WDS]) for non native languages and increases the LID performance considerably. Also the memory and computational requirements of these techniques are much less compared to basic LZW techniques.

Index Terms: Language modelling, PRLM, LS-LZW, LF-LZW.

1. Introduction

Usually LID is done by tokenizing the input signal first followed by building of language models for these tokens. The common tokenization for spoken language identification is that of phonemes of one or more languages [1]. In text based LID, there is an implicit assumption of common alphabets, whereas for spoken language, there is greater freedom in choosing the tokens. Among the linguistically derived tokens, phonemes can be extended to broad phonetic categories comprising of vowels, diphthongs or larger units such as syllables. Other types of tokenization include GMM tokenization [2] and fixed or variable length segment level tokenization [4]. Thus tokenization addresses the issues of resolution in the acoustic space as well as the duration of the tokens. The language models are often stochastic models over the token set, viz., unigram, bigram distributions [1][3], ergodic-HMM [4][5], duration models, etc. Various architectures of LID have been proposed [3], viz. (i) PRLM (phone recognition followed by language model), (ii) PPRLM (parallel PRLM), (iii) PPR (parallel phone recognition). In this paper, we use the PRLM architecture because of its simplicity.

Stochastic models proposed for LID are typically Markov models of small orders (unigram, bigram, trigram etc.). The language discriminability is limited by the nature of the model itself, with higher order models likely to do better, but higher order model estimates being poorer with limited training data.

The approach proposed in [6] attempts to solve the above limitation by using a deterministic model for a language. The deterministic model is automatically derived from the training data, which is assumed to be generalizable for the unseen test data as well. This is made possible by developing a pattern table of token sequences. The loss-less coding technique of LZW algorithm [8] is utilized in this respect. Once the pattern table is built, for a given test sequence, a compression ratio (LZW-CR) or weighted discriminant score (LZW-WDS) is computed for all the language pattern tables and highest scoring pattern table is reported as the language identity of the test sequence.

The LZW technique allows the basic structural unit of the language to be of variable length. So the technique captures all the advantages of higher order models but with much less training data. But the LZW technique, being a deterministic approach, builds the pattern tables regardless of the frequency of occurrence of the patterns. A pattern P occurring most in language $L1$ will also be present in the pattern table of language $L2$ even if it occurs once in the training data of language $L2$. As a result of this, more than 50% of the patterns in each pattern table do not carry any language structure information. These add confusability to the LID task, thereby inheriting the performance.

Here we propose two solutions for overcoming this limitation. We build LZW pattern tables as before but then make these pattern tables more language specific by pruning it. Two pruning techniques are discussed in Section 2. Section 3 describes the experiments, results and a comparison with bigram and basic LZW techniques. Finally we conclude in Section 4.

2. Discriminative Pattern Table Building

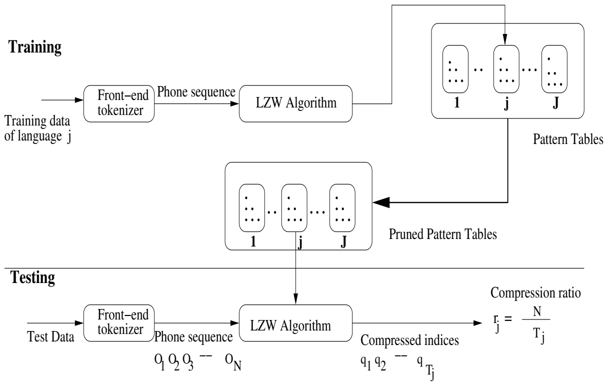

Figure 1: *Block Schematic showing the Training and Testing Phase of LS-LZW-CR based LID*

2.1. Training Stage

In training, we use LZW algorithm to build pattern tables (Fig. 1) corresponding to each language. As compared with the n-gram technique, the LZW method does not constrain the individual patterns to be of the same size. Once the LZW pattern tables are built, they are pruned by using one of following methods.

2.1.1. Language Specific (LS) Pruning

For each language pattern table, only those patterns which are unique to that pattern table are retained. i.e. a pattern P present in language $L1$ pattern table is retained only if P doesn't appear in any of the other pattern table. After this the pattern table is modified in such a way that all the patterns have their prefixes also as members of that pattern table. i.e if a language $L1$ pattern table has a pattern say $P = abcd$, then its prefixes abc, ab and a are also added to the pattern table. This is required to ensure the basic LZW pattern table property that all prefixes of a pattern in a pattern table are also patterns in that table. This leads to an increase in the pattern table size. Even with this addition, about 25% reduction in pattern table size is obtained. Thus the memory requirement as well as search complexity of LS-LZW technique is about 25% lesser than the basic LZW technique.

2.1.2. Length-Frequency (LF) product based Prunning

The frequency of occurrence of a pattern in a particular language and its length are the two main measures indicating the presence of that pattern structure in that language. So for each pattern a measure, which is the product of its frequency of occurrence in that language training data and its length, is found out and those patterns whose length-frequency product is lower than a preset threshold (for this paper all pattern tables have the same threshold 5) are retained in that pattern table. The threshold determines the number of patterns removed from the pattern table, more the threshold more will be the number of patterns removed from the table. When the threshold is set to a value 5, reduction in pattern table size is more than 50%. This considerably reduces the computation required on the test data, yet at the same time preserves the patterns essential for performing LID. The prefix property of the LZW pattern tables is not violated by this pruning. For easiness in writing, in future we refer to the pruned pattern tables as pattern tables.

2.2. Testing

Using the pattern tables from the discriminative training, we obtain a classification score for any test sequence. Two such scores have been identified [6].

2.2.1. Compression Ratio (CR)

For the test sequence, each newly found pattern is coded by its index in the pattern table. Since, the test sequence of patterns is represented only by a sequence of indices, the algorithm achieves compression. The test sequence is separately compressed by the pattern table of each language. For the given test sequence, if a pattern table is representative of its language and if the test sequence contains patterns unique to that language, the compression ratio will be high. Conversely, if the phoneme sequence does not correspond to the language of the pattern table, the phonemes get coded individually resulting in a low compression ratio. We define the compression ratio as the ratio of the number of phonemes in the test sequence to the number of indices obtained after LZW compression.

2.2.2. Weighted Discriminant Score (WDS)

Here each pattern is assigned a weighting factor. For a pattern p_i of length s, the weight factor $L_j(p_i)$ for language j is calculated as:

$$L_j(p_i) = \frac{N_{p_i}}{N_s} \tag{1}$$

where N_{pi} denotes the number of times the pattern p_i occurred in the training data and N_s denotes the number of patterns of length s in the training data. The weight factors are normalized by dividing each weight by the sum of weight factors of all the patterns in the pattern table. This sort of weighting is done for all the language pattern tables.

For a test sequence o its discriminant score for a language j, $L_j(o)$ is calculated as follows. The test sequence o is converted into a sequence of patterns by using pattern table of language j. Let q_{ji}, $i = 1, ..., T_j$ denote these patterns.

Now the discriminant score of o for language j is defined as the product of the weight factors of the individual patterns.

$$L_j(o) = \prod_{i=1}^{T_j} L_j(q_i) \tag{2}$$

The assumption here is that, the individual patterns are independent. If the patterns are not independent they would not have occurred separately in the pattern table. i.e. if a pattern q_k and q_l, $k, l = 1..T_j$ are not independent of each other, then q_k and q_l would not be separately present but the concatenated pattern $q_k q_l$ would be present in the pattern table. This assumption will hold good when the training data for building the pattern tables contain most of the valid patterns occurring in the language.

From the set of languages J, the language index for the test sequence is j^*

$$j^* = \arg\max_{j \in J}(L_j(o)) \tag{3}$$

Thus the training stage corresponds to building of all the language pattern tables followed by pruning of the pattern tables and the testing stage involves only the classification score calculation.

3. Experiments and Results

The experiments for the LID task are performed on the 6 language OGI-TS data base, which contains manually labeled phonetic transcriptions. The 6 languages are : English, German, Hindi, Japanese, Mandarin and Spanish. The OGI-TS database uses transcription based on the multi-language motivated Worldbet [7]. The transcriptions of the *story-bt* sentences of the OGI-TS database uses 923 symbols in all, from the 6 languages. The phonetic detail is made explicit by use of diacritics. The diacritics are merged into the base labels leaving us with approximately 150 symbols. By grouping together similar sounding phonemes, this is further reduced to 50 language-independent phonetic units. The resulting 50 units, which include several silence and non-speech units, are shown in Table 1.

Now, we present the results of spoken language identification on the 6 languages of the OGI-TS database. Each *story-bt* utterance is at least 45 sec long and is spoken by a unique speaker. We divide the utterances of each language into two parts, training speakers and testing speakers (mutually exclusive). The *story-bt* being extempore and free, makes the LID task text independent and speaker independent. To simulate the real tokenization we introduce controlled amount of token errors to manually assigned phonetic labels. Noisy tokenization is realized by first generating one random variable for each token. This random variable takes on values 1 and 0 with probabilities p and $1 - p$ respectively, where p is the induced artificial error rate. For each token, if the value of the corresponding random variable is 1, then that token is replaced by any one of the other tokens (all with equal probability). On the other hand, if the value of this random variable is 0, then that token is left unmodified. Thus, after this process we get a noisy tokenization of the speech utterance with an error rate of p. Noise is added to the training tokens as well as to the test phonemes. We have generated token sequences with 30% error (corresponding to typical phoneme error rates of an automated front end) to test the LID task.

Table 1: *50 size Phone Inventory including non-speech symbols.*

vowel(14)	i, 3r, I, u, E, >, @, &, o, a, 8, e, 2, ax
semivowels(4)	w, l, r, j
diphthongs(8)	ai, ei, ou, au, iu, Eax, oi, uax
nasals(3)	m, n, N
fricative(9)	f, s, sh, v, z, h, D, G, T
affricate(3)	dZ, ts, cC
stops(6)	b, d, g, k, p, t
non speech(3)	pause, line,breath,smacking noise, other noises

Training data for each language consists of 20000 phonemes. Test utterances have lengths varying from 20 to 300 phonemes. Using the mentioned techniques namely Bigram, basic LZW with compression ratio (LZW-CR) and with weighted discriminant score (LZW-WDS), Language Specific LZW with compression ratio (LS-LZW-CR) and with weighted discriminant score (LS-LZW-WDS) and Length Frequency product based LZW with compression ratio (LF-LZW-CR) and with weighted discriminant score (LF-LZW-WDS), the LID task has been performed and the results averaged over the 6 languages for $p = 0.3$ is reported in Table 2. The graphical illustration of the average LID performance for p = 0.3 is also shown in Fig. 2 (WDS as the measure) and Fig. 3 (CR as the measure).

These experiments justify our claim that the LS-LZW and LF-LZW techniques give discriminative language models by removing the confusable patterns. For a test sequence, these methods reduce its classification score for non native languages, thereby provide good LID performance. Also the reduction in memory and test data score computation complexity is around 25% for the LS-LZW techniques whereas its around 50% for the LF-LZW techniques.

4. Conclusions

We propose two discriminative language modelling techniques using LZW based LID namely Language Specific-LZW, in which we remove patterns from the pattern ta-

Table 2: *Average LID accuracy for $p = 0.3$ (30% tokenization noise).*

Test seq size	Bigram	LZW-CR	LZW-WDS	LS-LZW-CR	LS-LZW-WDS	LF-LZW-CR	LF-LZW-WDS
20	66.05	38.5	76.40	49.15	77.32	46.73	78.84
40	66.04	60.78	89.34	70.27	90.39	68.53	90.78
60	66.49	71.41	94.5	80.73	95.42	79.64	95.01
80	71.89	80.19	97.00	87.60	97.67	86.09	97.56
100	73.78	86.73	98.60	91.26	98.23	90.42	98.62
150	78.48	92.00	99.53	97.06	99.53	95.91	99.39
200	81.20	95.59	99.71	98.09	99.60	97.44	99.71
250	82.30	97.19	99.64	99.28	99.86	98.89	100
300	86.35	98.48	100	99.26	100	99.26	100

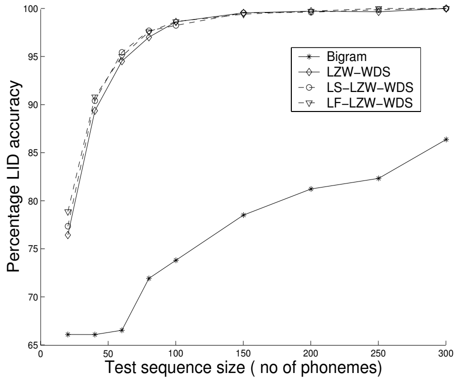

Figure 2: *Average LID accuracy for $p = 0.3$ using Weighted Discriminant Score (WDS) as a measure.*

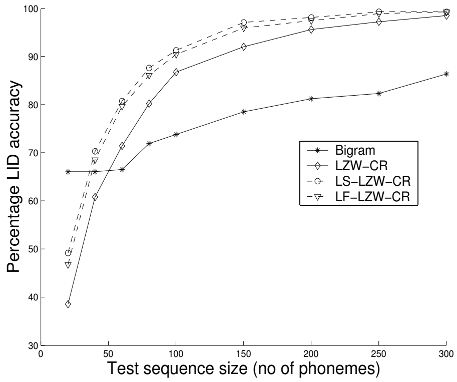

Figure 3: *Average LID accuracy for $p = 0.3$ using Compression Ratio (CR) as a measure.*

ble common to more than one language and LF-LZW, in which we remove patterns from the pattern table whose length-frequency product falls below a preset threshold. We thus maintain the good language modelling capability of the LZW technique and yet at the same time increase language discriminability to the pattern tables. This claim is justified by the LID experiments using these methods, where we show a great deal of improvement over the basic LZW technique and bigram technique. All this is achieved at a much reduced computation and memory as well.

5. References

[1] M.A. Zissman, "Language identification using phoneme recognition and phonotactic language modelling", Proc. ICASSP, Apr 1995, pp-3503-3506.

[2] P.A. Torres-Carrasquillo, D.A. Reynolds and J.R. Deller, "Language identification using Gaussian mixture model tokenization", Proc. ICASSP,Apr 2002, Vol.1, pp-I-757-I-760.

[3] M.A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech", IEEE Trans Speech and Audio Processing, Vol. 4, No. 1, Jan 1996, pp-31-44.

[4] V. Ramasubramanian, A.K.V. SaiJayaram and T.V. Sreenivas, "Language identification using parallel sub-word recognition - an ergodic HMM equivalence", Proc. Eurospeech, Geneva, Sep 2003, pp-1357-1360.

[5] S.A.Santosh Kumar and V. Ramasubramanian, "Automatic language identification using ergodic-HMM", Proc. ICASSP, Apr 2005, pp-I-609-I-612.

[6] S.V. Basavaraja and T.V. Sreenivas, "LZW Based Distance Measures for Spoken Language Identification", Proc. IEEE Odyssey 2006: The Speaker and Language Recognition Workshop, Puerto Rico, Jun 2006.

[7] T. Lander, "The CSLU labeling guide", Center For Spoken Language Understanding, Oregon Graduate Institute, May 1997.

[8] Mark Nelson, "LZW Data Compression", Dr Dobbs Journal, Oct 1989