TUNING PHONE DECODERS FOR LANGUAGE IDENTIFICATION

C.P. Santhosh Kumar[1], Haizhou Li[2], Rong Tong[2,3], Pavel Matějka[4], Lukáš Burget[4], Jan Černocký[4]

[1]ECE Department, Amrita Vishwa Vidyapeetham, Ettimadai, India
[2] Institute for Infocomm Research, A*Star, 1 Fusionopolis Way, Singapore
[3] School of Computer Engineering, Nanyang Technological University, Singapore
[4] Speech@FIT, Brno University of Technology, Czech Republic

cs_kumar@ettimadai.amrita.edu, {hli, tongrong}@i2r.a-star.edu.sg,
{matejkap, burget, cernocky}@fit.vutbr.cz

ABSTRACT

Phonotactic approach, phone recognition to be followed by language modeling, is one of the most popular approaches to language identification (LID). In this work, we explore how language identification accuracy of a phone decoder can be enhanced by varying acoustic resolution of the phone decoder, and subsequently how multiresolution versions of the same decoder can be integrated to improve the LID accuracy. We use mutual information to select the optimum set of phones for a specific acoustic resolution. Further, we propose strategies for building multilingual systems suitable for LID applications, and subsequently fine tune these systems to enhance the overall accuracy.

Index Terms— Phonotactic language identification, hidden Markov models, neural networks, mutual information, multilingual

1. INTRODUCTION

In most of the speech databases, phone definitions are done according to IPA or SAMPA or a similar definition to optimize the accuracy for speech recognition. In language identification (LID) using phonotactic approach [1, 2, 9], words across languages are pronounced using the phones of the phone decoder and the difference in these pronunciations are captured using n-gram language models to identify the language of the spoken utterance.

It is well known that the performance of the language identification systems can be enhanced by using decoders specific to target languages. In this case, we may choose language specific decoders generated from the labeled data or derive language specific phones from a multilingual phone inventory as in [3]. However, this approach is not practical when the number of target languages in the LID system becomes large. Therefore, it would be interesting to optimize the systems for the overall performance across all the target languages. There has also been effort to improve the performance of the LID systems using discriminative approaches [5], or by combining the output of several decoders [6, 7]. While the work in [3] was interesting for target oriented phone selection, it was focused on choosing the set of phones from a multilingual phone inventory, specific to the target languages. Extending [3] to select the optimum set of phones across all target languages (TOPT extended), the performance deteriorated from the baseline system, when applied to decoders not specific to the target languages.

It may be noted that phone recognition accuracy and LID accuracy are highly correlated. Further, when a phone is removed from the decoder, the examples which otherwise are recognized as this phone would now be recognized as other phones. The effect of these two factors were not considered in the selection of the phones in [3]. As a result, it could not be extended to non-target language decoders. We, in this work, formulate the phone selection process as an optimization of mutual information [4, 10, 11] of the new decoder, taking into consideration the above two factors. We further check the effectiveness of the approach using a hybrid hidden Markov model - neural network (HMM-NN) implementation [8, 9] of the Hungarian [1] decoder.

Further, we propose strategies for developing multilingual decoders suitable for LID applications. We also fine tune the resolution of the multilingual decoder and integrate with the monolingual decoders to enhance the LID accuracy. Hungarian and Czech[1] decoders were used for building the multilingual decoders.

We present the theoretical formulation of the new phone selection algorithm, using bigram counts. We use mutual information to select the optimum set of phones and later verify if the phones not selected should be removed from the database or should they be replaced with a phone that has similar acoustic characteristics based on the mutual information of the new phone decoder resulting from the decision. Thus, in our work, the phone selection problem is reformulated as a phone elimination problem, and we refer to this approach as **P**hone **S**election by **E**limination (PSE).

2. PHONE SELECTION BY ELIMINATION (PSE)

2.1. Mutual information

An intuitively plausible measure of the average amount of information provided by the random event T about the random event L is the average difference between the number of bits it takes to specify the outcome of L when the outcome of T is not known and the outcome of T is known. Mutual information is a powerful tool to measure the dependency between random variables[4]. Suppose a discrete token variable (unigram, bigram or a trigram) and language class variable

This work was partly supported by European project AMIDA (FP6-033812), by Grant Agency of Czech Republic under project No. 102/08/0707 and by Czech Ministry of Education under project No. MSM0021630528. Santhosh Kumar's stay at BUT was sponsored by AMIDA training programme. Authors would like to thank Petr Schwarz, BUT, for the many fruitful discussions during the course of this work.

[1]http://catalog.elra.info/product_info.php?products_id=1045

are T and L, respectively, the mutual information between the two variables is defined as:

$$I(T,L) = H(L) - H(L|T) = \sum_T \sum_L p(t_i, l_j) \log \left(\frac{p(t_i, l_j)}{p(t_i)p(l_j)} \right) \quad (1)$$

where $H(\cdot)$ is the entropy.

Ideally, from the string of tokens, the language identity should be obtained. This means that $I(T,L)$, the average mutual information between the random variable t_i and the random variable l_j should be maximized, or rather the decoder should have phones with high mutual information.

Further, we consider an LID task for N languages as N two class problems, where each of the target language needs to be distinguished from the rest of the languages. Thus, we can formulate the language recognition as a series of two class separation problems. For each target language, a one-versus-rest classifier can be built, which is indeed the case, we need to negate the non-target languages and accept the target language for the specific sample. We consider the identity of a language l as a random variable that can take one of many values depending on the number of languages to be recognized. Given the task of classifying a target language l against the competing languages, we define another variable $\Gamma = \{l^+, l^-\}$, which has two values, where l^+ denotes that it is the target language, l^- denotes that it is not the target language. We use $bg(t_i, t_j)$ to denote the bigram with phone t_i to be followed by t_j. For a given language l, the presence of the bigram $bg(t_i, t_j)$ is another random variable that takes two possible values in $\beta = \{bg(t_i, t_j)_{l^+}, bg(t_i, t_j)_{l^-}\}$ where $bg(t_i, t_j)_{l^+}$ denotes that the bigram $bg(t_i, t_j)$ is present in language l and $bg(t_i, t_j)_{l^-}$ denotes that the bigram $bg(t_i, t_j)$ is present in language l^-. Thus, Γ is the target versus non-target classes, and β is the set of phones that are present in the target language or non-target languages or both.

Mutual information of a phone can then be expressed in terms of the presence bigram β and the language category Γ summed over all languages that can be estimated as:

$$I(\beta;\Gamma) = \sum_{l \in \Gamma} \sum_{bg(t_i,t_j) \in \beta} a(t_i) p(bg(t_i, t_j), l) \log \frac{p(bg(t_i, t_j), l)}{p(bg(t_i, t_j))p(l)} \quad (2)$$

where $p(bg(t_i, t_j), l)$ is the probability that the bigram $bg(t_i, t_j)$ appears in language l, $p(bg(t_i, t_j))$ and $p(l)$ are the probability of bigram $bg(t_i, t_j)$ and language l respectively, and $a(t)$ is the recognition accuracy of the phone t without considering insertion and deletion errors.

2.2. Entropy leak

Let us consider a system with three phones ah, ih and uh, and their recognition accuracy without considering insertion and deletion errors be 90, 80, and 40 per cent respectively. It is clear that more examples of the phone uh are being recognized as ah or ih resulting in inconsistency in the definitions of ah and ih when they are recognized using the phone decoder. This in turn increases the randomness (entropy) in the recognition of phones ah and ih and reduces the mutual information of these phones, phones with high recognition accuracy. Thus, the mutual information of a phone calculated from the output sequence is after the blurring effect of the phone recognition errors. To penalise phones with negative contribution towards the LID performance of other phones, we chose to multiply the mutual information of every phone with the phone recognition accuracy, without considering the insertion errors.

merge	i	i:	merge	t	t1:
merge	s	s:	merge	ts	ts_
merge	r	r:	merge	S	Z
merge	z	z:	merge	m	m:
merge	k	k:	merge	d_	d_:
merge	d_	t1	merge	j	j:
merge	S	S:	merge	z	dz
remove x					

Table 1. *Decision whether to merge or remove the phones not selected for the decoder*

From eqns. (1) and (2), it seems that selecting phones with large values of mutual information is the way to choose phones for LID systems, it is now clear that we also need to consider the negative contribution towards the performance of other phones, which is influenced by the phone recognition accuracy.

2.3. Selection of phones

Rather than choosing the phone with the highest mutual information, our strategy is to eliminate the ones that cause minimum loss in the mutual information of the phone decoder after its deletion or substitution with another phone. In other words, we modified the phone selection problem to a phone elimination/deletion problem.

If a phone is selected for deletion, it could either be removed from the decoder, or edit the phone sequence at the output of the decoder and replace it with the acoustically closest phone or merge the examples of the two phones before building the acoustic models. In the case of deletion, all examples which would otherwise have been recognized by this phone will now be recognized as other phones. Confusion matrix of the decoder was used to estimate what percentage of this phone examples are distributed towards the counts of other phones. Similarly, if it is chosen for replacement with another phone, then all the phones which were recognized correctly as this phone would get replaced by the new target phone and the rest would get distributed towards the counts of the other phone bigrams. This distribution also can be estimated from the confusion matrix of the phone decoder.

We calculate the mutual information of the system without the phone under consideration and check if the mutual information of the system would be better with a removal of the phone from the decoder or with a substitution with the acoustically closest phone and make a decision as regards to every phone chosen for deletion. Phones were classified into broad phonetic categories and the closest phone in the same group only was considered for a subsitution. If a phone x is chosen for subsitution with phone y and vice versa, then they are not selected in this phase for deletion/substitution, there is a chance that they could be parallel models of the same phone. In this experiment, however, we did not encounter such case.

Table 1 lists the results of identifying the phones with minimum effect on the mutual information of the phone decoder. $merge\ y\ y$: means the phone y: is merged with y, either in the model or during the post-editing of the labels. $remove\ x$ means phone x can be deleted from the decoder. In our experiments, merging of the phones in the model gave slightly better results than post-editing, and there-

at start, 45 phones			7.15e-02
merge	pau_	int_	7.49e-02
merge	pau_	spk_	7.51e-02

Table 2. *Merging of the final set of phones for better mutual information*

fore all the results reported in this work use merging of phones in the model.

2.4. Measuring the acoustic similarity of phones

In our work, we use continuous density HMMs where the probability distributions were modeled using neural networks[8, 9], and the probability was not measured in the likelihood sense across all possible state sequences, but for the best state sequence S_{opt}. Examples were therefore force aligned with the correct transcription to the state level and the distance of phone λ_1 and λ_2 can be expressed using KL-divergence [12] as in [13]:

$$\begin{aligned} \mathrm{D}(\lambda_1, \lambda_2) &= \frac{1}{N}\sum_{k=1}^{N}\frac{1}{T_k}\Big\{\sum_{t_k=1}^{T_k}\log(p(\mathbf{o}(t_k)|\lambda_1, s(t_k))) \\ &- \log(p(\mathbf{o}(t_k)|\lambda_2, s(t_k)))\Big\} \end{aligned} \quad (3)$$

$\mathbf{o}(t)$ is the feature vector at time t, and $s(t)$ is the state at time t, and $p(o(t)|\lambda, s(t)))$ is the probability of $o(t)$ for model λ, at state $s(t)$ obtained after force aligning the example with their correct phone state transcription , T_k is the number of frames in the k^{th} example and N is the total number of examples of the model λ_1.

In our case, however, a merge resulted in a better phone recognition and LID accuracy and therefore the results reported in this work use merging of phones before training the acoustic models.

2.5. Merging phones with high mutual information

By now, we have decided the number of phone models to be used in the decoder. There may be phones in the decoder, that are parallel models of the same phone, or phones having diverse acoustic characteristics, but representing the same linguistic event.

For this, at a time, we chose a phone to be merged with every other phone in the selected phone inventory, select the best merge in terms of the mutual information if such a merge enhanced the mutual information of the decoder. After deciding on this, it starts over with the next phone merge and goes until such a merge does not lead to an increase in the mutual information of the decoder. In this case, it may be noted that the models were maintained in the decoder, only replacing of the labels was done if it enhances the mutual information of the decoder. Table 2 shows how the mutual information of the decoder increases for certain phone pair merges.

3. EXPERIMENTS AND RESULTS

3.1. Fine tuning a monolingual decoder

For benchmarking the results, we used 30 secs. segments of NIST 2005 dataset[1], while CallFriend database[2] for the respective languages was used for training the language models. Bigram counts

[1]http://www.nist.gov/speech/test/lre
[2]http://www.ldc.upenn.edu

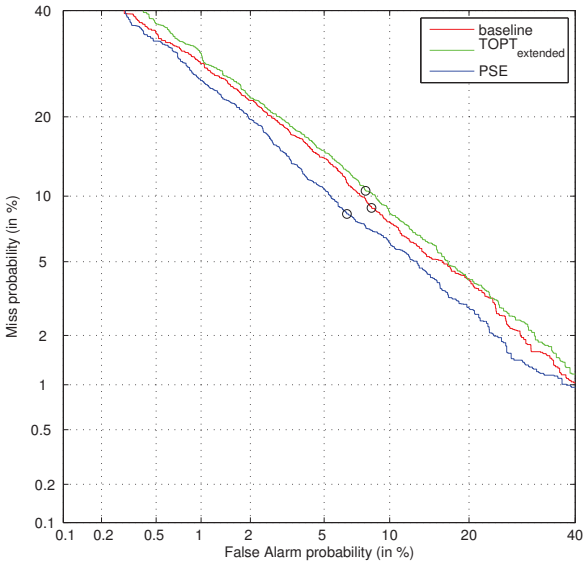

Fig. 1. *DET curve - Comparison of the LID results of the baseline system, TOPT extended[3], and PSE. min DCF points are marked using o.*

	Baseline	TOPT extended	PSE
min. DET	8.62	9.19	7.41
EER	8.69	9.26	7.58

Table 3. *Comparison of the LID results of TOPT extended and PSE for a decoder with 43 phones derived from the baseline system*

required for estimating the mutual information were calculated from the data prepared by us for the calibration of the LID systems. All phone decoders were fine tuned for the optimum phone insertion penalty, and was found that it is different from the optimum phone insertion penalty value for the best phone recognition results. Our baseline monolingual Hungarian and Czech decoders have 61 and 45 phones respectively. In all the experiments reported in this work, we used 500 neurons to model the probability distributions in the HMM-NN decoder. The performance of the baseline system could have been enhanced by using a bigger neural network to model the probability distributions, but for compuational considerations, we chose a moderate size of 500 neurons.

Table 3 and Figure 1 compare the results the LID performance for the 30 secs. segments of the NIST data for the baseline system with 61 phones, and the system with 43 selected phones using the TOPT extended and PSE.

Further, we varied the acoustic resolution of the system across different number of phones. Reducing the number of phones was

No. of phones	35 hu_{35}	43 hu_{43}	48 hu_{48}	57 hu_{57}	61 hu_{61}	fused system
min. DET	8.11	**7.41**	7.70	8.20	8.62	**5.92**
EER	8.28	**7.58**	7.80	8.34	8.69	**6.11**

Table 4. *Effect of varying the acoustic resolution on the LID performance*

	one to one mapping				many to one mapping			
No. phones	70	73	76	82	64	69	72	81
min. DET	8.68	8.37	8.47	8.62	8.92	8.14	**7.77**	8.49
EER	8.74	8.45	8.79	8.66	8.94	8.42	**7.96**	8.52

Table 5. *LID - Multilingual approach*

Decoder	HU	CZ	multilingual	fused (HU+CZ)	fused (ML)
min. DET	8.62	9.87	7.44	6.69	**5.01**
EER	8.69	9.98	7.80	6.78	**5.11**

Table 6. *Final system - fused using linear backend*

found to be meaningful for the Hungarian decoder, and we note that there is always an acoustic resolution that is optimum for the LID task, which sometimes could be different from the resolution arising out of the phone definitions of the database. Table 4 shows the LID performance of systems with different acoustic resolutions derived from the same baseline system and also the final system when the output of systems with 43, 48 and 57 phones were fused into a single output using a linear backend [7]. This combination outperformed other combinations.

3.2. Multilingual phone recognition for LID

For building the multilingual(ML) decoder, we considered the Hungarian(HU) and Czech(CZ) decoders. First, we grouped phones in each of the languages to phonetically motivated clusters, and let each phone in CZ (source language) to be mapped to a phone in HU (target language) if the distance (eqn. (3)) between them is less than a chosen threshold. This threshold effectively decides the number of phones in the multilingual system. Mapping is allowed between members of the same phonetic cluster only, and details of the algorithm can be found in [13]. Now, there are several strategies possible:

1. One to one mapping (**o2o**) - Only one phone in the source language is allowed to be mapped to a target language phone. If more than one phone satisfies the distance criteria, the closest is selected for mapping.
2. Many to one mapping(**m2o**) - All phones in the source language satisfying the distance criteria are mapped to the target language phone.
3. Use SAMPA/IPA mapping

Table 5 lists the LID performance of the **o2o** and **m2o** mapping of phones across CZ and HU for different phone inventory sizes. Incidentally, for the multilingual system using SAMPA mapping with 75 phones, we got 8.44 and 8.46 per cents respectively for min. DET and EER respectively. This is very close to the LID results for the best **o2o** mapping with 73 phones, for which phone mappings were similar, but not the same as SAMPA. In the **m2o** mapping with 72 phones, 34 phones from the CZ were mapped to 28 phones in HU and 11 phones in CZ remained without any mapping.

Subsequently, the multilingual decoders were fine tuned for the acoustic resolution using PSE, and a system with 47 phones gave 7.44 and 7.80 per cents min. DET and EER respectively as its performance measure. This decoder was derived from the **m2o** mapped decoder with 72 phones and has 33 phones from the CZ decoder mapped to 27 HU phones, and 11 phones not shared between CZ and HU.

Further, we integrated CZ, HU using a linear backend[7], and also the PSE modified HU decoder hu_{43}, CZ and the multilingual decoder with 47 phones derived from the **m2o** mapped decoder with 72 phones. Table 6 lists the results of each of the systems.

4. CONCLUSION

We presented a method to vary the acoustic resolution of phone decoders to enhance the language identification performance using phone selection by elimination approach. Also, it was shown that such multiresolution systems developed for the same language could be integrated together for better overall LID performance.

We then proposed strategies for the development of multilingual decoders and then fine tuned these decoders to optimize the LID performance. Further, the multilingual decoder and monolingual decoders were integrated to enhance the overall system performance. It was seen that the phone mapping using SAMPA/IPA suitable for speech recognition applications is not the optimum mapping for LID applications.

5. REFERENCES

[1] M.A.Zissman, "Comparison of four approaches to automatic language identification of telephone speech", IEEE Trans. on Speech and Audio Processing, Vol. 4, No. 1, pp 31.34, 1996.

[2] Muthusamy, Y.K., Bernard, E., Cole, R.A, "Automatic Language Identification: A Review/Tutorial". IEEE Signal Processing Magazine, 1994 vol. 11, no. 4, pp. 33-41.

[3] R. Tong, et al., "A target oriented phonotactic frontend for spoken language recognition", IEEE Trans. on Audio, Speech and Language Processing, Vol. 17, No. 7, 2009, pp. 1335-1347.

[4] Z. Huang, A. Acero, and H. W. Hon, Spoken Language Processing - A guide to theory, algorithm, and system development, Prentice Hall Inc., 2001

[5] P. Matějka. P. Schwarz, L. Burget, J. Črnocký, "Use of anti-models to further improve state-of-the-art PRLM language recognition system", Proc. ICASSP, 2006

[6] N.Brummer, J.D.Preez, "Application dependent evaluation of speaker detection, Computer Speech and Language", Vol. 20, pp. 230-275, 2006

[7] N. Brummer, "FoCal Multi-class toolkit for evaluation, fusion, and calibration of multi-class recognition scores", June 2007.

[8] P. Schwarz, P. Matějka, and J. Černocký, "Towards lower error rates in phoneme recognition," Proceedings of 7th International Conference Text, Speech and Dialogue 2004, 2004.

[9] P. Matějka, P. Schwarz , J. Černocký, P. Chytil, "Phonotactic Language Identification using High Quality Phoneme Recognition", Proc. Eurospeech2005, Sept. 2005.

[10] R. Rosenfeld, "Adaptive statistical language modeling: a maximum entropy approach", Ph.D Thesis, Cernegie Mellon University, 1992.

[11] K.W. Church, "Word association norms, mutual information and lexicography", Computational Linguistics, Vol. 16, No. 1, March 1990, pp. 22-29.

[12] Kullback, S., 1958. Information Theory and Statistics. Wiley, New York.

[13] C. S. Kumar, et al, "Training acoustic models for languages with insufficient training data", Oriental COCOSDA 2008, Nov. 25-27, Kyoto, Japan.