

EFFICIENT ADAPTATION TEXT DESIGN BASED ON THE KULLBACK-LEIBLER MEASURE

Xiaodong Cui and Abeer Alwan

Department of Electrical Engineering
University of California, Los Angeles, CA 90095
Email: xdcui@icsl.ucla.edu, alwan@icsl.ucla.edu

ABSTRACT

This paper proposes an efficient algorithm for the automatic selection of sentences given a desired phoneme distribution. The algorithm is based on the Kullback-Leibler measure under the criterion of minimum cross-entropy. One application of this algorithm is the design of adaptation text for automatic speech recognition with a particular phoneme distribution. The algorithm is efficient and flexible, especially in the case of limited text size. Experimental results verify the advantage of this approach.

1. INTRODUCTION

The automatic selection of sentences is important in several applications such as designing speaker adaptation text for ASR and speech synthesis. In speaker adaptation, the adaptation text is used to modify the speaker independent model to a specific user. Obviously, the less sentences the talker has to read the better. In some cases, unit (phoneme or sub-word unit) models are not adapted properly due to limited adaptation text. For example, some phonemes may occur more frequently than others. This unbalanced phoneme distribution can be problematic for system adaptation. Therefore, for supervised speaker adaptation, it is important to efficiently design adaptation text with phonemes (or other units) assuming a predefined (typically balanced) distribution while keeping the text size small. There are several methods to tackle this problem such as those reported in [1], [2] and [3]. Almost all the proposed methods are add-on procedures. Namely, given a score to each sentence to be selected, choose the sentence with the best score and add it to the list. Repeat this procedure until a certain criterion is met. Different approaches define different scores which reflect the satisfaction degree to one's problem. In [1], the score is defined as the frequency difference between the entire corpus and the current sentence set while in [2], a normalized inner product is used. In considering the sentence selection problem, phoneme frequency is certainly the most useful information which is exploited in the above two deterministic

methods. A more complicated model-based strategy in [3] makes strong assumptions on the model structure and ties the model to a greedy algorithm. In this case, the rank of the design matrix and its cardinality constitute the score for selection.

In this paper, we propose a novel approach to design adaptation text efficiently and flexibly. The algorithm enables the designer to predefine the desired phoneme distribution as well as the size of the text. During the design process, the algorithm selects from a large text pool the best sentences in a heuristic way according to the criterion of minimum Kullback-Leibler (KL) measure [4]. Hence, we treat the sentence selection problem in a statistical manner.

In Section 2, we introduce the KL measure and its properties. Then, in Section 3, a heuristic algorithm for text design under the minimum cross-entropy principle is described. Experimental results are shown in Section 4. We conclude with a summary in Section 5.

2. KULLBACK-LEIBLER MEASURE

The KL measure is a widely-used measure in statistics which depicts similarity or "closeness" between two probability distributions [5].

In the discrete case, it is defined as

$$I(\mathbf{p} \parallel \mathbf{p}^0) = \sum_{k=1}^n p_k \log \frac{p_k}{p_k^0} \quad (1)$$

where \mathbf{p}^0 and \mathbf{p} are two probability distributions. From the statistical point of view, the KL measure is the expected logarithm of the likelihood ratio. (By convention, we let $0 \log \frac{0}{p} = 0$.) \mathbf{p}^0 can be considered the true distribution while \mathbf{p} is the one used to approximate it. In Eq. 1, n is the number of element events in the discrete probability space which, in our application, is the number of phonemes.

It is not difficult to prove that the KL measure has the following properties:

$$1) I(\mathbf{p} \parallel \mathbf{p}^0) \geq 0$$

$$2) I(\mathbf{p} \parallel \mathbf{p}^0) = 0 \iff \mathbf{p} = \mathbf{p}^0$$

The measure we use here is a one-way deviation measure which is also known as the Kullback-Leibler divergence or measure of cross-entropy. In optimization problems, the objective function $I(\mathbf{p} \parallel \mathbf{p}^0)$ is minimized with respect to \mathbf{p} in order to get the best approximation to the true distribution. In special cases, when \mathbf{p}^0 is uniformly distributed, the minimization of cross-entropy leads to maximum entropy.

In this paper, \mathbf{p}^0 is a predefined ideal phoneme probability distribution which usually assumes uniform or other task-specific distribution. \mathbf{p} is the practical phoneme distribution one obtains during the design process. The algorithm is originally motivated by the idea of choosing a phoneme probability distribution that is "closest" to the prior one in the space of discrete probability distribution according to a proper measure which in this case is the KL measure.

3. ADAPTATION TEXT DESIGN ALGORITHM

In order to design satisfactory text for speaker adaptation, we want to choose sentences from a large corpus. The goal is to use the text to cover as many phonemes as possible and to cover them equally or in a predefined way. The first goal is not difficult to achieve since in most cases the entire phoneme set can be covered by a reasonable number of sentences. The second goal is more challenging. For some adaptation text, phonemes are distributed unevenly which results in unbalanced data for different phonemes. The algorithm described here solves the following problem: given the size of the text, we choose the sentences that cover all the phonemes in the phoneme set while maintaining a desired phoneme distribution.

Let $S = \{s_1, s_2, \dots, s_M\}$ denote the text corpus from which we select adaptation sentences; where s_i stands for the i th sentence and there are a total of M sentences. Suppose we need to select N sentences from the corpus as adaptation text which is denoted by $A_N = \{s'_1, s'_2, \dots, s'_N\}$, where $s'_i \in S$ and $N \leq M$.

We want to select those N sentences whose phoneme distribution is the "closest" to the prior one. That is,

$$\begin{aligned} A_{N,opt} &= \operatorname{argmin}_{A_N \subseteq S} I(\mathbf{p}(A_N) \parallel \mathbf{p}^0) \\ &= \operatorname{argmin}_{A_N \subseteq S} \sum_{k=1}^n p_k(A_N) \log \frac{p_k(A_N)}{p_k^0} \end{aligned} \quad (2)$$

where $\mathbf{p}(A_N)$ denotes that the phoneme distribution is a function of the N sentences selected from the text pool.

The most straightforward way to solve the above optimization problem is to traverse all the N -sentence selection cases in the whole M -sentence text pool. But this process is computationally expensive. If we define the calculation of phoneme statistics as one computation unit, the above

- 1) Set ideal prior phoneme distribution to \mathbf{p}^0
- 2) $A_0 = \phi$
- 3) For $i = 0$ to $N-1$
 - a) For every sentence $s_j \in S \setminus A_i$
 - b) Let $B_{ij} = A_i \cup \{s_j\}$
 - c) Calculate the phoneme probability distribution $\mathbf{p}(B_{ij})$ defined by frequency of occurrence
 - d) Calculate $I(\mathbf{p}(B_{ij}) \parallel \mathbf{p}^0)$, the Kullback-Leibler measure between $\mathbf{p}(B_{ij})$ and \mathbf{p}^0
 - e) Select the sentence $s_{j,opt}$ so that

$$B_{ij,opt} = \operatorname{argmin}_{B_{ij}} \sum_{k=1}^n p_k(B_{ij}) \log \frac{p_k(B_{ij})}{p_k^0}$$
 - f) $A_{i+1} = A_i \cup \{s_{j,opt}\}$
 - g) End

Fig. 1. Heuristic text selection algorithm based on the minimum KL measure

process requires $\binom{M}{N}$ such computation units. Finally, we have to make a comparison and get the best choice. Even for relatively small numbers of N and M , the above computation requirement is very high (e.g. when $N=20$, $M=500$, $\binom{M}{N} \approx 2.7 \times 10^{35}$).

Here we propose a quasi-optimal heuristic design method which gives good results while keeping the computational load reasonable.

Fig. 1 provides the details of the algorithm. Starting from the empty set, we add one sentence at a time until the N -sentence requirement is reached. Each time, we select a sentence so that the newly formed sentence set has the minimum KL measure between its phoneme distribution and the prior one.

The computational load for the algorithm is in the order of $O(NM)$ (for $N=20$, $M=500$, $O(NM) = O(10^4)$) with respect to the computation unit mentioned above. This is affordable even for relatively large N and M . On a Pentium IV 933MHz processor with 256M of memory, it took about 11 minutes for the selection of 40 sentences from the 450-sentence TIMIT corpus and about 25 minutes for the 100 sentences case. Furthermore, it is easy to observe that the

resulting adaptation text sets have the following property:

$$A_0 \subset A_1 \subset A_2 \subset \dots \subset A_N \quad (3)$$

This property implies a convenient way to get a new set of text from an existing one. In fact, for $L \leq N$, A_L can be constructed by choosing the first L sentences from A_N in which sentences are arranged in the sequential generation order described by the algorithm. But note that it is not necessarily advantageous to increase N beyond a certain number.

The initialization of prior phoneme distribution \mathbf{p}^0 is also flexible. One can let \mathbf{p}^0 assume some other reasonable form besides the uniform one and generate adaptation text with phoneme distribution approximating it.

4. EXPERIMENTAL RESULTS

We choose the phoneme set in the BEEP dictionary provided by HTK. There are 45 phonemes in this set (including a short pause symbol). In the design experiment, we only use the first 44 phonemes, as shown in Table. 1, for computing the statistics and ignore the short pause symbol which is not meaningful in this design process.

EY	AA	AE	OW	IY	OH	EH	AO	AY
AH	AW	IA	UW	ER	UH	OY	UA	EA
AX	IH	Z	K	N	D	V	B	S
NG	M	T	SH	R	L	HH	G	JH
ZH	P	Y	TH	CH	F	W	DH	

Table 1. 44 phonemes used in this paper

Our text pool comes from TIMIT’s 450 SA sentences which includes phonetically-compact sentences designed to provide a good coverage of phonemes and phonetic contexts.

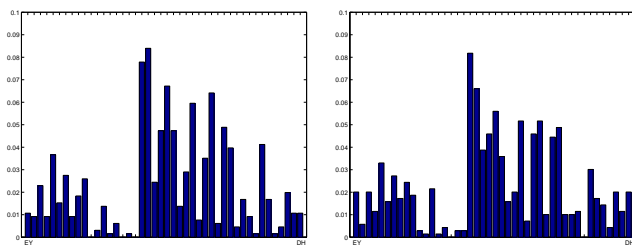


Fig. 2. Two instances of phoneme distributions of 20 randomly selected sentences from TIMIT.

Figs. 2 and 3 illustrate the phoneme distributions if we randomly select 20 or 40 sentences from the corpus. The horizontal axis indices from left to right are phonemes in

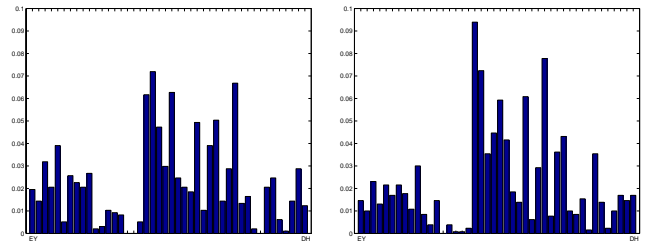


Fig. 3. Two instances of phoneme distributions of 40 randomly selected sentences from TIMIT.

the row order of Table. 1 (same order for following figures). Generally speaking, phoneme distributions from these sentences are highly unbalanced. Some of the phonemes appear much more frequently than others.

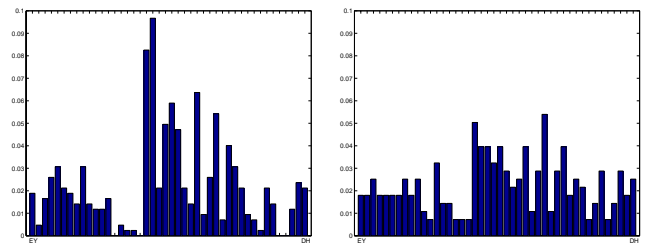


Fig. 4. Comparison of the phoneme distributions of 10 sentences chosen randomly(left) and using the minimum KL measure(right) from TIMIT.

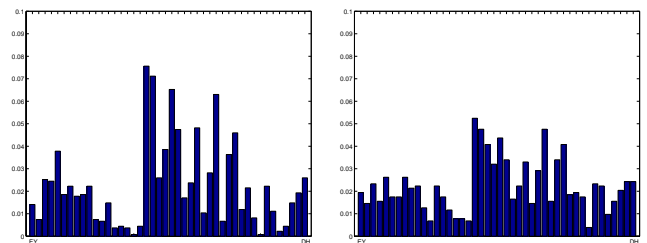


Fig. 5. Comparison of the phoneme distributions of 40 sentences chosen randomly(left) and using the minimum KL measure(right) from TIMIT.

Figs. 4 and 5 show the experimental results for the 44 phoneme distributions of 10 and 40 sentences selected from the 450 TIMIT SA sentences by minimizing the KL measure with the prior phoneme distribution set to a uniform one. As mentioned before, under this condition, the effect is the same as that under the maximum entropy principle. The figures show a relatively flatter distribution in those sentences chosen by the minimum KL measure. As shown, some phonemes have higher probability than others because

they appear more frequently in words. There are certain embedded statistical distributions inside words for vowels and consonants and may be considered as constraints in this optimization problem.

As the number of sentences grows, the advantage of using this algorithm is not as pronounced. This is illustrated in the 100-sentence case in Fig. 6.

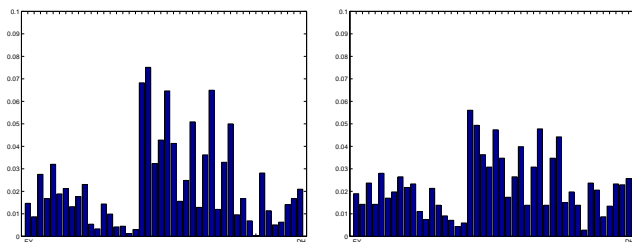


Fig. 6. Comparison of the phoneme distributions of 100 sentences chosen randomly(left) and using the minimum KL measure(right) from TIMIT.

In the above experiments, the prior phoneme probability distributions are assumed to be uniform. However, the flexibility of this approach lies in the fact that one can tune p^0 to any distribution. For example, from the former figures some phonemes appear more often than others (e.g. /AX/, /IH/) which may not be beneficial for system adaptation. In this case, we can tune p^0 .

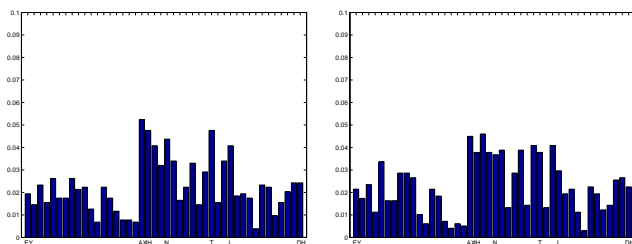


Fig. 7. Comparison of the phoneme distributions of 40 sentences chosen with uniform prior(left) and non-uniform prior(right) from TIMIT.

In Fig. 7 we set the prior phoneme distribution p^0 to the form that the five phonemes (/AX/, /IH/, /N/, /T/ and /L/) with the highest probabilities in the uniform case have one third the probabilities of the rest. From the result, it can be observed that the distribution is somewhat flatter. If the size of text pool is larger, the result might be better since we have more sentences to choose from. In fact, the choice of p^0 can allow for the control of each phoneme distribution by giving it a proper probability or weight.

In a specific speech recognition task, the text corpus defining the task also has certain specific phoneme distribution as discussed in [2]. In order to design matched adapta-

tion text, we have to select those sentences whose phoneme distribution resembles the one of the task. Our algorithm has clear advantages in this case. What is only required would be to obtain the phoneme statistical property p^0 from the task and set it as the prior phoneme distribution in our algorithm.

5. SUMMARY

In this paper, we propose an efficient algorithm based on minimizing the KL measure to automatically select sentences with a desired phoneme distribution. The heuristic method leads to adaptation text with good statistical property and good approximation to the predefined prior phoneme distribution. Theoretically, we can handle each phoneme weight in the phoneme set by a deliberate choice of the prior phoneme distribution. Experimental results show satisfactory performance of this algorithm and especially in the case of limited text size.

6. ACKNOWLEDGEMENTS

This work was supported in part by NSF.¹

7. REFERENCES

- [1] Vlasta Radova and Petr Vopalka, "Methods of sentences selection for read-speech corpus design," *Proceedings of second International Workshop on Text, Speech and Dialogue*, pp. 165–170, 1999.
- [2] Jia lin Shen, Hsin min Wang, Ren yuan Lyu, and Lin shan Lee, "Automatic selection of phonetically distributed sentence sets for speaker adaptation with application to large vocabulary mandarin speech recognition," *Computer Speech and Language*, vol. 13, pp. 79–97, 1999.
- [3] Jan P.H. van Santen and Adam L. Buchsbaum, "Methods for optimal text selection," in *Proc. of EuroSpeech 1997*, pp. 553–556.
- [4] Shu-Cherng Fang, Jay R. Rajasekera, and H.-S. Jacob Tsao, *Entropy optimization and mathematical programming*, Kluwer Academic Publishers, 1997.
- [5] Solomon Kullback, *Information theory and statistics*, John Wiley & Sons, 1959.

¹This material is based upon work supported by the National Science Foundation under Grant No. ANI-0085773. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).