DISCRIMINATION POWER WEIGHTED SUBWORD-BASED SPEAKER VERIFICATION

Siu-Man Chan, Man-Hung Siu

Dept. of EEE, Hong Kong University of Science and Technology
Clear WaterBay, Kowloon, Hong Kong
eeman@ust.hk, eemsiu@ust.hk

ABSTRACT

Since the discriminative powers of different phones vary, weighting techniques can be applied to bias the log-likelihood ratio of different phone segments to emphasize the more discriminating phones. First, we derive a Kullback-Leibler distance-based phone log-likelihood weighting scheme that incorporate the duration information. Second, we propose a posterior probability transformation method such that contributions from different phones are weighted implicitly according to their discrimination power. Using the two proposed approaches reduces speaker verification equal error rate by more than 10% on the YOHO database.

1. INTRODUCTION

Security is a key issue in this information age. Authenticating a user's identity is needed before one can access sensitive information or location. Speaker verification, which uses voice to authenticate a user's identity, is convenient that no password is needed and can be used over remote networks, such as over the telephone.

Speaker verification can be characterized into two types: text-independent or text-dependent. Gaussian Mixture Model (GMM) is commonly used in a text-independent system. In text-dependent systems, because the transcription of the spoken words are known, subword-based models are typically used. In either approaches, a log-likelihood ratio (LLR), comparing the test likelihood from the target distribution to that from the imposter distribution is computed as a score for verification decision. Using different modeling units in the speaker verification task, the LLR can be computed per HMM state, phone or word and then combined.

In subword-based verification system, the LLR distributions can differ from phone to phone and from speaker to speaker. Techniques such as Zero-normalization, are introduced to stabilize verification score alleviate the speaker variation such that speaker independent thresholding [1] can be effective. To handle variation across different phonetic models, other researchers have shown that certain sounds, such as high energy nasals and vowels have more discriminative power in speaker verification task than some other phones.

Because the general hypotheses test framework commonly used in speaker verification does not account for these variations in discriminative power, an explicit weighting or normalization scheme is needed. For example, [2] suggesting that LLR can be weighted according to phone energy level. In this paper we examine two ways to weight the verification score to take advantage of the variation of phone discriminative power. First, we derive a Kullback-Leibler distance-based phone LLR weighting scheme. We showed that this turned out to be similar to the proposed weighting in [3] but can incorporate the duration information directly.

Second, we propose a posterior probability transformation method such that contributions from different phones are weighted implicitly according to their discrimination power. Under this formulation and assuming each speech frame to be independent, each phone score is weighted by the "distance" between the target-specific distribution and imposter distribution of the LLR score and the phone duration. However, differ from other weighting scheme, this weighted phone score is offset by a phone-dependent bias before being combined and transformed by a sigmoid function to a valid posterior probability between 0 and 1. To account for the potential correlation between frames within a phone, an extra phone-dependent exponentiation factors are introduced. We then show that the posterior probability transformation with a phone dependent correlation-compensation factor is a special case of generalized linear model. The posterior probability approach is compared with a Kullback-Leibler distance based weighting schemes, both in terms of formulation and performance and the KL distance schemes.

The rest of this paper is organized as follows. In the next section, the general framework of verification via LLR test is discussed, a mathematical basis of developing weight schemes are described. The proposed posterior probability approach on subword model weighting method is described in Section 3. Experimental results are reported in Section 4 and the conclusions are presented in Section 5.

2. SPEAKER VERIFICATION WITH SUBWORD WEIGHTING

2.1. General Verification Framework

For a given test utterance $O = \{O_1, O_2, \ldots, O_N\}$, speaker verification decides whether O is produced by the claimed speaker.

This verification task can be treated as a hypothesis testing problem deciding whether the utterance is generated by the claimed speaker model λ (null hypothesis) against the alternative that it is produced by an imposter $\overline{\lambda}$ (alternative hypothesis). The LLR of the two hypotheses is given by,

$$r(O, \lambda, \overline{\lambda}) = \log P(O|\lambda) - \log P(O|\overline{\lambda}). \quad (1)$$

The decision process is typically based on the following rule:

$$\frac{1}{N}\sum_{n=1}^{N} r(O_n, \lambda_n, \overline{\lambda}_n) \begin{cases} > \tau & \text{accept} \\ \leq \tau & \text{reject,} \end{cases} \quad (2)$$

where N is the length of the test utterance. In this simple approach, the LLR of the claimed speaker model and the imposter model is compared with the threshold τ. However, both τ and the likelihood may depend on the speaker or on the sequence of words spoken. Thus, score normalization may be needed.

2.2. Subword-based Speaker Verification

In subword based text-dependent speaker verification, the input utterance is segmented into P phones, typically by using the Viterbi algorithm. The observation sequence can then be expressed as $O = \{O_1^{t_1}, O_{t_1+1}^{t_2}, \ldots, O_{t_{P-1}+1}^{t_P}\}$, where frame $t_{p-1}+1$ to frame t_p are associated with the p^{th} phone. We simplify the expression of phone duration, $t_p - t_{p-1} + 1$ by denoting it as ℓ_p in the rest of the paper. We denote the LLR of an i-th observation of phone p as $s_{p,i}$. $s_{p,i}$ can be computed by

$$\begin{aligned} s_{p,i} &= r(O_{t_{p-1}+i}, \lambda_p, \overline{\lambda}_p) \\ &= \log p(O_{t_{p-1}+i}|\lambda_p) - \log p(O_{t_{p-1}+i}|\overline{\lambda}_p) \end{aligned} \quad (3)$$

The verification score in Equation 2, denoted as $VER^{(1)}$, can be expressed as

$$VER^{(1)} = \frac{1}{t_P}\sum_{p=1}^{P}\sum_{i=1}^{\ell_p} s_{p,i} = \frac{1}{t_P}\sum_{p=1}^{P}\overline{s}_p\ell_p, \quad (4)$$

where $\overline{s}_p = \frac{1}{l_p}\sum_{i=1}^{l_p} s_{p,i}$ is the average log-likelihood ratio of the p^{th} phone in the test utterance. One can view $\overline{s}_p$ as the contribution of phone p to the verification score weighted by duration ℓ_p. If some subwords have more discriminative power than others, one approach to improve verification performance is to weight these subword scores $\overline{s}_p$ based on their contributions instead of their duration.

2.3. Kullback-Leibler Distance Phoneme Weighting

Importance of phone segments can be increased by weighting with the discrimination power of different phones. How to measure the discrimination power of the phone? One approach is to examine the conditional distributions of $\overline{s}_p$ given the target model and imposter model. This is similar to what was suggested in [4] in confidence estimation. While $\overline{s}_p$ actually is the average log-likelihood ratio, it is now considered a random variable and its class-conditional distributions can be estimated. If the two class-conditional distributions are "far away" from each other, one believes that $\overline{s}_p$ has good discrimination power. Kullback-Leibler (KL) distance is one kind of distances to measure how far two distributions are.

Denote the target distribution as $p_T(\overline{s}_p) = N(\bar{\mu}_p^T, \bar{\sigma}_p^2)$ and the imposter distribution as $p_I(\overline{s}_p) = N(\bar{\mu}_p^I, \bar{\sigma}_p^2)$. The bar above the mean and variance is used to denote that the mean and variance are over the per-frame score $\overline{s}_p$ instead of the individual likelihood ratio $s_{p,i}$. Because of the limited target data, note that both the imposter and target share the same variance $\overline{\sigma}_p^2$. A similar approach is taken by other researchers in normalizing against speaker variations, such as in Z-normalization [1]. The KL distance can be obtained by

$$KL_p = -(E_T \log \frac{p_T(\overline{s}_p)}{p_I(\overline{s}_p)} + E_I \log \frac{p_I(\overline{s}_p)}{p_T(\overline{s}_p)})$$

and

$$\begin{aligned} E_T \log \frac{p_t(\overline{s}_p)}{p_i(\overline{s}_p)} &= \int_s N(s_p; \bar{\mu}_p^T, \bar{\sigma}_p^2) \log\left[\frac{N(s_p; \bar{\mu}_p^T, \bar{\sigma}_p^2)}{N(s_p; \bar{\mu}_p^I, \bar{\sigma}_p^2)}\right] ds \\ &= \frac{\bar{\mu}_p^T\bar{\mu}_p^I - \bar{\mu}_p^{T\,2}}{\bar{\sigma}_p^2} + \frac{\bar{\mu}_p^{T\,2} - \bar{\mu}_p^{I\,2}}{2\bar{\sigma}_p^2} \end{aligned} \quad (5)$$

Thus,

$$KL_p = (\frac{\bar{\mu}_p^T - \bar{\mu}_p^I}{\overline{\sigma}_p})^2. \quad (6)$$

In [3], a phone weighting scheme that normalizes the difference between the target mean and imposter mean with imposter variance was proposed with an added exponent γ to adjust the dominance of each phone. γ is typically found empirically. With such a weighting factor, the verification score of a utterance, denoted as $VER^{(2)}$ is given by,

$$VER^{(2)} = \frac{1}{P}\sum_{p=1}^{N}\overline{s}_p w_p = \frac{1}{P}\sum_{p=1}^{N}\overline{s}_p\left(\frac{\bar{\mu}_p^T - \bar{\mu}_p^I}{\overline{\sigma}_p}\right)^{\gamma}. \quad (7)$$

When γ is set to 2, it is equivalent to a weight with the KL distance. Because we are normalizing against phone variations, these weights can be estimated either speaker dependently or speaker independently. Instead of letting the average phone score $\overline{s}_p$ as a random variable, we can model the per-frame score $s_{p,i}$ to be Gaussian distributed whose mean and variance are phone dependent. Then, a similar KL distance can be derived, denoted as $VER^{(3)}$, given by,

$$VER^{(3)} = \frac{1}{\sum_{p=1}^{N} w_p\ell_p}\sum_{n=1}^{N}\sum_{i}^{\ell_p} s_{p,i}w_p, \quad (8)$$

where $w_p = \left(\frac{\mu_p^T - \mu_p^I}{\sigma_p}\right)^{\gamma}$.

In effect, the duration becomes an explicit factor scaling the weighting factor.

3. POSTERIOR PROBABILITY TRANSFORMATION

3.1. Posterior Probability Score

One difficulty with weighting the phone discrimination power and the duration is that very often, they are correlated. For example, vowels often are more discriminative but also longer. Trying to factor out the individual differences systematically is not easy. Following by the assumption that the likelihood ratios are themselves Gaussian distributed random variables, we can derive the posterior probability of whether a test utterance is produced by the target speaker T, denoted as $P(T|s_1, s_2, \ldots, s_N)$ where s_i is the individual likelihood ratio as defined in Equation 3. To take into consideration the duration information, we assume that the frame-by-frame likelihood ratio, $s_{p,i}$ is a random variable. The conditional distribution of $s_{p,i}$ under the target and imposter conditions are denoted as,

$$\begin{aligned} p(s_{p,i}|T) &= N(s_{p,i}; \mu_p^T, \sigma_p^2), \\ p(s_{p,i}|I) &= N(s_{p,i}; \mu_p^I, \sigma_p^2). \end{aligned}$$

The posterior probability of observing $P(T|s_1, s_2, \ldots, s_N)$ with phone alignment $t_1, \ldots, t_P$ can be expressed as

$$\begin{aligned} &P(T|s_1, s_2, \ldots, s_N) \\ &= \frac{P(s_1, s_2, \ldots, s_N|T)P(T)}{P(s_1, s_2, \ldots, s_N|T)P(T) + P(s_1, s_2 \ldots, s_N|I)P(I)}. \end{aligned}$$

Here s_n is the log-likelihood ratio $r(O_n, \lambda, \overline{\lambda})$ of the n^{th} frame of the test utterance, and there are totally N frames in the utterance.

Assuming the observations within each subword are independent, we have

$$
\begin{aligned}
&P(T|s_1, s_2, \ldots, s_N) \\
&= \frac{\prod_{p=1}^{P} P(s_{t_{p-1}+1}^{t_p}|T)}{\prod_{p=1}^{P} P(s_{t_{p-1}+1}^{t_p}|T) + \prod_{p=1}^{P} P(s_{t_{p-1}+1}^{t_p}|I)} \\
&= \frac{1}{1 + e^{-\frac{1}{2}\left\{\sum_{p=1}^{P}\sum_{i=1}^{\ell_p}\left[(\frac{s_{t_{p-1}+i}-\mu_p^I}{\sigma_p})^2 - (\frac{s_{t_{p-1}+i}-\mu_p^T}{\sigma_p})^2\right]\right\}}}
\end{aligned} \tag{9}
$$

The sigmoid function, denoted as $sgm(\alpha(x-\tau))$ is given by

$$
sgm(\alpha(x-\tau)) = \frac{1}{1+e^{-\alpha(x-\tau)}} \tag{10}
$$

By denoting $s_{p,i} = s_{t_{p-1}+i}$ and using the sigmoid function in Equation 10, we can rewrite Equation 9 as

$$
\begin{aligned}
&P(T|s_1, s_2, \ldots, s_N) \\
&= sgm\left(\sum_{p=1}^{P}\sum_{i=1}^{\ell_p}\left[\frac{s_{p,i}(\mu_p^T - \mu_p^I)}{{\sigma_p}^2} - \frac{(\mu_p^T - \mu_p^I)(\mu_p^T + \mu_p^I)}{2\sigma_p^2}\right]\right) \\
&= sgm\left(\sum_{p=1}^{P}\sum_{i=1}^{\ell_p}[W_p s_{p,i} - W_p B_p]\right) \\
&= sgm\left(\sum_{p=1}^{P}\ell_p W_p(\overline{s}_p - B_p)\right)
\end{aligned} \tag{11}
$$

Here, we substitute $W_p = \frac{\mu_p^T - \mu_p^I}{\sigma_p^2}$ and $B_p = \frac{\mu_p^T + \mu_p^I}{2}$.

Since sigmoid is a monotonic function, it transforms the verification score to posterior probabilities. However, it does not change the verification result in terms of equal error rate. In the transformation of the LLR score $\overline{s}_p$ in Equation 11, the decision of each subword depends on $\overline{s}_p - B_p$, and the contribution of the decision is weighted with the segment length ℓ_p and W_p. Again, W_p is the difference between the target and imposter means normalized by variance. However, the score is not weighted directly. Instead, it is first shifted by B_p, which is mid-point between the two conditional means. In effect, a positive contribution occurred only if $s_{p,i}$ is closer to the target mean than the imposter mean.

Comparing the posterior probability transformation with $VER^{(3)}$, there are two main differences. First, the weighting terms $\ell_p \frac{\mu_p^T - \mu_p^I}{\sigma_p}$, and $\ell_p \frac{\mu_p^T - \mu_p^I}{\sigma_p^2}$ have a minor difference that is the later is normalized by σ_p^2. Second, the transformation scheme contains a shift of phones verification score. Notice that these shifts only depend on the phone labels and phone durations but are independent of the value of $\overline{s}_p$.

3.2. Generalized Linear Model (GLM) Based Posterior Probability

In the last section, we derived Equation 11 with the assumption that LLR are independent. It is well-known that speech observations are correlated. If we consider the joint random variables, x and y, $p(x,y) = p(x)p(x)$ only if they are independent. In the extreme case where $x = y$, then, $p(x,y) = p(x) = (p(x)p(y))^{1/2}$. Adding exponents to the joint probabilities is a way to compensate for the wrongly-estimated probability. Thus, a scaling factor γ can be added in Equation 11.

$$
\begin{aligned}
&P(T|s_1, \ldots, s_N) = \\
&= \frac{\prod_{p=1}^{P} P(s_{t_{p-1}+1}^{t_p}|T)^{\gamma_p}}{\prod_{p=1}^{P} P(s_{t_{p-1}+1}^{t_p}|T)^{\gamma_p} + \prod_{p=1}^{P} P(s_{t_{p-1}+1}^{t_p}|I)^{\gamma_p}} \\
&= sgm\left(\sum_{p=1}^{P}\gamma_p(\ell_p W_p \overline{s}_p - \ell_p W_p B_p)\right)
\end{aligned} \tag{12}
$$

Interestingly, the γ_p becomes a scaling factor on the per phone contribution. How we can estimate the correlation factor γ_p for each phones? Equation 12 expresses the class posterior probability as a sigmoid function of a linearly weighted combination of per phone scores. This is a special case of the generalized linear model or logistic regression. We rewrite Equation 12 in GLM notations as follows.

$$
GLM(x_p, \gamma_p) = sgm(\gamma_0 + \sum_{p=1}^{P}\gamma_p x_p), \tag{13}
$$

where $x_p = \ell_p(W_p\overline{s}_p - W_p B_p)$ are the features to the GLM. Well-known solutions that maximize either likelihood or square errors [5] are available.

4. EXPERIMENTS

Subword-based text dependent speaker verification experiments are conducted to evaluate the proposed algorithms. We use the duration weighted $VER^{(1)}$ as the baseline and compare it with the two KL-based approaches, i) $VER^{(2)}$ in phone based that does not take the phone duration into account and ii) $VER^{(3)}$ that is in frame based that contains duration information. Then, we evaluate the performance of posterior probability transformation with and without the correlation compensation.

4.1. Data

The corpus used in these experiments is the YOHO speaker verification corpus which is widely used for speaker verification experiments [6, 7, 8, 9]. For each speaker in the corpus, there are 4 enrollment sessions, each with 24 utterances, and 10 verification sessions, each with 4 utterances. Each utterance consists of three sets of two-digit numbers, such as (e.g. thirty-four sixty-one seventy-six). All sessions were recorded in an office environment using high quality telephone handset and sampled at 8kHz. Since the LLR distributions of female and male speakers are quite different, only male speakers were used in our experiments. 50 speakers were held out as a development set. So, the remaining 52 speakers were selected as the target speakers and the imposters for other claimed speakers.

4.2. Experimental Setup

The universal models $\bar{\lambda}$ were trained using the digits subset of the Macrophone corpus [10], with no training data from YOHO. Features used were 12 MFCC (Mel-Frequency Cepstral Coefficient) plus energy and their first and second order derivatives. Two sets of models were tested, i) 20 3-state left-to-right monophone HMM models and ii) 48 3-state triphone HMM models with up to 14 and 10 Gaussian mixtures per state. In order to estimate the per phone parameters of the conditional distributions of the LLR, 5-fold cross

	Baseline	Weighted		
	EER(%) $VER^{(1)}$	EER(%) $VER^{(2)}$	EER(%) $VER^{(3)}$	EER(%) $Post.P$
triphone	0.808	0.846(2.5)	0.731(2)	0.731
monophone	1.192	1.423(0.5)	1.115(0.5)	1.038

Table 1. Comparison between the EER of different weighting schemes with the baseline for monophone and triphone model

EER(%) Without γ	EER(%) SD γ
0.731	0.692

Table 2. Comparison between the EER of posterior transformation with or without γ for triphone model

validations were applied. The 4 enrollment data sets per target speaker were partitioned into 5 subsets, 4 of which were used to estimate the target speaker model λ via adaptation of the universal model [11] and one was used to estimate the target parameters $\bar{\mu}_p^T$ and μ_p^T. The parameters for the imposter, $\bar{\mu}_p^I$, μ_p^I, $\bar{\sigma}_p^2$ and σ_p^2 were estimated using the held-out development set, in which 20 utterances from each of the 50 speakers were selected. Similarly, the γ used in GLM-based posterior probability model were determined using cross-validation and the held-out development set.

After the parameters for the conditional distributions were estimated, a new model λ for the target speakers using all 4 enrollment sessions in YOHO training was created, again by adapting from the universal model $\bar{\lambda}$. The performance of the speaker verification with different weighting methods were evaluated using equal error rate (EER) across all speakers.

4.3. Experimental Results and Discussions

The experiment results using SI weights are tabulated in Table 1.

The first columns is the baseline result $VER^{(1)}$. While $VER^{(2)}$ was proposed with speaker dependent weights in [3], we found that SI weights perform as well when evaluated on the EER across all speakers, although SD weights improve per speaker EER in some cases.

The result of $VER^{(2)}$ is much worse than the baseline while $VER^{(3)}$ gave significant improvement. This suggests that the duration information is important to use in conjunction with phone weighting. In both $VER^{(2)}$ and $VER^{(3)}$, γ have to be estimated. In Table 1, the best empirically estimated γ are 2.5 and 2 for triphone models and 0.5 and 1 for monophone models for $VER^{(2)}$ and $VER^{(3)}$ respectively. While the weights equal to KL distance when γ is set to 2 which was reported in [3] as the best γ,we found that a smaller γ gave better performance on monophone model. This may be due to the correlation between adjacent frames compound the effect that make a smaller γ more preferable, or it is due to some model dependent difference.differences.

The last column of Table 1 shows the results of the posterior probability transformation. We can see that the SI posterior probability transform gave an additional reduction in EER on monophone model as compared to the baseline.

As suggested in Section 3.2, exponents can be added to account for frame correlation. Table 2 shows the results of using extra exponents on the triphone task. It turns out that it is useful when it was estimated in a speaker dependent fashion.

5. CONCLUSIONS

In this paper we proposed two weighting schemes to emphasize subword units that are more discriminative in speaker verification. By assuming the per-frame loge likelihood ratios between the target and imposter models to be a random variable, we proposed to use the Kullback-Leibler distance between the subword-dependent class conditionals as a subword weight. This is shown to reduce EER by 10%. Instead of computing the weighted combination of the subword scores, we also proposed a posterior probability transformation of that weights the subword units implicitly. It turns out that under the posterior probability framework, the subword scores not only are weighted but also shifted by a subword-dependent bias before they are combined. In addition, the posterior probability framework can be view as a special case of GLM. Experimental results shown that the posterior probability is slightly better than then KL weighting approach.

6. ACKNOWLEDGMENT

This work is partially supported by RGC grant no. CA02/03.EG05.

7. REFERENCES

[1] R. Auckenthaler, M. Carey, and H. Lloyd-thomas, "Score normalization for text-independent speaker verification systems," *IEEE on Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[2] S. Parthasarathy and A. E. Rosenberg, "General phrase speaker verification using sub-word background models and likelihood-ratio scoring," in *Proceedings on ICSLP*, 1996.

[3] S. Ahn, S. Kang, and H. Ko, "On effective speaker verification based on subword model," in *Proceedings on ICSLP*, 2002, pp. 1361–1364.

[4] P. Jeanrenaud, M. Siu, and H. Gish, "Large vocabulary word scoring as a basis for transcription generation," in *Proceedings on EUROSPEECH*, 1995, pp. 2149–2152.

[5] P. McCullagh and J. A. Nelder, *Generalized linear models*, Chapman and Hall, London, 1989.

[6] D. A. Reynolds and B. A. Carlson, "Text-dependent speaker verification using decoupled and integrated speaker and speech recognizers," in *Proceedings on EUROSPEECH*, 1995, pp. 647–650.

[7] J. Jr. Campbell, "Testing with the yoho cd-rom voice verification corpus," in *Proceedings on ICASSP*, 1995, pp. 341–345.

[8] N. B. Yoma, "Speaker verification in noise using a stochastic version of the weighted viterbi algorithm," *IEEE Transactions on Speech and Audio Processing*, vol. 10, pp. 158–166, Mar 2002.

[9] J. K. Koolwaaij and L. Boves, "A new procedure for classifying speakers in speaker verification systems," in *Proceeding EUROSPEECH*, 1993, pp. 2355–2358.

[10] J. Bernstein, K. Taussig, and J. Godfrey, "Macrophone: An american english telephone speech corpus for the polyphone project," in *Proceedings on ICASSP*, 1994, pp. 19–22.

[11] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," in *Digital Signal Processing*, 2000, vol. 10, pp. 19–41.