SETIT 2005
3rd International Conference: **S**ciences of **E**lectronic, **T**echnologies of **I**nformation and **T**elecommunications
March 27-31, 2005 – TUNISIA

Fusion of Cohort-Word and Speech Background Model Based Confidence Scores for Improved Keyword Confidence Scoring and Verification

K. Thambiratnam and S. Sridharan

Speech and Audio Research Laboratory, Queensland University of Technology,
GPO Box 2434, Brisbane, Australia 4001

k.thambiratnam@qut.edu.au

s.sridharan@qut.edu.au

Abstract: This paper presents results for Keyword Verification (KV) experiments using Cohort-Word Verification (CWV) and neural-network fusion of CWV and Speech Background Model based Verification (SBMV). Baseline experiments found that CWV excelled in short-word KV while SBMV was more robust for long-word KV. Fusion of CWV and SBMV yielded dramatic improvements in both short-to-medium length KV using the fusion of two CWV systems, and medium-to-long length KV using fusion of SBMV and CWV. Overall, through target word phone length dependent fusion of verifiers, it was possible to at least halve the false rejection rate of the baseline SBMV verifier for all evaluated keyword length classes.
Key words: Confidence Scoring, Keyword Spotting, Keyword Verification, Verifier Fusion.

1. Introduction

Keyword verification (KV) algorithms are an effective method of improving the robustness of speech recognition applications. Such algorithms have played an important role in many state of the art speech recognition systems, such as speech transcription systems, dialog systems and keyword spotting systems.

This paper focuses on the application of KV to keyword spotting systems (eg. audio database search engines, real-time keyword monitoring and command control systems). Keyword spotting systems rely heavily on a robust KV system to reduce false alarm rates. In applications that require high keyword spotting speeds, computationally intensive KV methods such as confusion networks (Mangu & al., 2000) are unsuitable. Instead, KV methods that are faster and only require acoustic analysis of the KV candidate are used.

Two such KV methods are Speech Background Model based Verification (SBMV) (Wilpon & al., 1990) and Cohort-Word Verification (CWV) (Thambiratnam & al., 2003). Both methods were found in preliminary experiments to have acceptable KV performance for medium and long keywords, but poorer performance for short-word KV. In particular, SBMV performance was significantly degraded for short target keyword lengths, having up to 5 times the false rejection obtained of long-word SBMV at the same false acceptance rate. Reduced performance was also observed for short-word CWV, though the extent of degradation was not as much.

To address the issue of poor short-word KV performance, this paper proposes the use of neural network fusion of CWV and SBMV. It is anticipated that combining verifiers would yield improvements in KV through the use of orthogonal information. Additionally there would be little impact on overall execution speed since both KV methods are fast.

This paper reports on experiments for neural network fusion of CWV and SBMV. Results are presented for a variety of target word phone lengths as well as for a number of fused system architectures. Additionally the performance of baseline unfused CWV and SBMV are provided.

2. Background

2.1 Speech background model log-likelihood ratio

A keyword confidence scoring metric typically takes the form of the Log-Likelihood Ratio (LLR):

$$C = \log(p(X \mid \lambda_{keyword})) - \log(p(X \mid \lambda_{nonkeyword})) \quad (1)$$

where X is the sequence of observations corresponding to the word candidate to be verified, $\lambda_{keyword}$ is the acoustic model for the target keyword (eg. concatenated phones or triphones) and $\lambda_{nonkeyword}$ is the acoustic model for the non-keyword against which the target word is scored.

The choice of the non-keyword model plays an important role in determining the quality of the confidence score. Numerous non-keyword models have been proposed in literature, including anti-syllable models (Xin & al., 2001), a uniform distribution (Silaghi & al., 2000) and a speech background model (Wilpon & al., 1990).

In speech background model KV, the non-keyword model is represented by a high-order Gaussian Mixture Model (GMM) trained on a very large speech corpus. This GMM is referred to as a speech background model (SBM). The confidence score log likelihood ratio is then given by:

$$C = \log(p(X \mid \lambda_{keyword})) - \log(p(X \mid \lambda_{SBM})) \quad (2)$$

2.2 Cohort-word keyword verification

A difficulty encountered with KV log-likelihood ratio confidence scores is how to accurately model the non-keywords. A good non-keyword model should model all keywords that are not the target keyword. This is difficult when the vocabulary of the system is unbounded or very large. The cohort-word confidence metric uses a non-keyword set that contains words that have a similar pronunciation to the target keyword. The metric attempts to measure how much better the observation sequence is modelled by the target keyword model than the other words in the target language with similar pronunciations. This is an extension of the method of anti-syllables proposed in (Xin et al., 2001), however selection is performed at the linguistically more natural word level rather than the sub-word level.

Selecting non-keywords at the word level instead of the sub-word level ensures that the non-keyword set only contains words that exist in the target language. For example, given a target keyword DINER, the non-keyword set may contain words such as DONOR and LINER when selecting non-keywords at a word level. However, when selecting at a sub-word level (eg. such as anti-syllables proposed in (Xin et al., 2001) the word DINER (syllable sequence DI-NER) may result in non-existent non-keywords such as GI-NER and RI-BER.

In addition, CWV should provide better discrimination in the case where the candidate to be verified is an actual occurrence of one of the words in the cohort-word set. This may well be a very common case, particularly in keyword spotting systems where the false alarms that are to be rejected are usually instances of words that are acoustically very similar to the target word.

2.2.1 Cohort-word verification confidence score

Given a target keyword w, let $R(w) = \{r_1, r_2, ..., r_N\}$ be defined as the cohort-word set (set of words in the target language that have a similar pronunciation) of word w. Note that the cohort-word set $R(w)$ does not include the target word w. The corresponding models of the cohort-word set are then used as the non-keyword model term in equation 1. Hence the cohort-word confidence score is given by:

$$C_1(w) = \log(p(X \mid \lambda_w)) - \log(\sum_{i=1}^{N} p(X \mid \lambda_{r_i})) \quad (3)$$

This formulation of the cohort-word confidence score requires calculation of likelihoods for all words in the cohort-word set. This can be computationally very expensive since the cohort-word set may be very large depending on the cohort-word selection criteria. To address this issue, a simplified approximation that is computationally less expensive is used. Let $S(w,K) = \{s_1, s_2, ..., s_K\}$ be the subset of the K top scoring words from the cohort-word set $R(w)$. Then the simplified cohort-word confidence score is defined as:

$$C_2(w,K) = \log(p(X \mid \lambda_w)) - \log(\sum_{i=1}^{K} p(X \mid \lambda_{s_i})) \quad (4)$$

This simplified cohort-word confidence score only requires the likelihoods of the K best scoring cohort words, which can be quickly obtained from a K-best recognition pass.

2.2.2 Cohort word selection

The cohort-word set is constructed by selecting words from a large word-to-phone dictionary. A cohort word is defined as a word with a pronunciation similar to the target word, where similarity is measured using the Minimum Edit Distance (MED) string alignment algorithm (Jurafsky, 2000).

Let $V = \{v_1, v_2, ..., v_M\}$ be the set of words in the target language and $\phi(w)$ be a mapping from the word w to the phonetic transcription of w. Then the distance between the target word, w, and the word v_i taken from the word set V is given by:

$$d(w, v_i) = MED(\phi(w), \phi(v_i)) \quad (5)$$

where $MED(a,b)$ is the Minimum Edit Distance between sequence a and sequence b.

Let the range $[d_{min}, d_{max}]$ be defined as the cohort-word selection range. The cohort-word selection range is used to restrict the cohort-word set to words that are within a certain MED distance from the target word (d_{min} was required as preliminary experiments

found that using cohort words that were very close in MED distance to the target word actually gave very poor performance).

Given the cohort-word distance formulation and the cohort-word selection range, the cohort-word set of target word *w* is then defined as:

$$R(w) = \{v \in V \mid d_{\min} \le d(w,v) \le d_{\max}\} \qquad (6)$$

To reduce the computational time required for cohort-word scoring, random sampling of the cohort-word set is used to reduce its size. This gives the reduced cohort-word set:

$$R(w,N) = top(N, shuffle(R(w))) \qquad (7)$$

where the *shuffle* function randomly shuffles a set, and the *top* function returns the first *N* elements of a set.

The MED algorithm allows weights to be specified for the insertion, deletion, substitution and match operations that are used during string alignment. In cohort-word selection, these weights can be used to favour certain types of cohort words over others. For example, using a high insertion penalty will result in more cohort-words that have fewer phones in their pronunciation than the target word. Let $\{\psi_d, \psi_i, \psi_s, \psi_m\}$ be defined as the cohort-word selection deletion, insertion, substitution and match penalties respectively. Collectively the set of parameters, $\{\psi_d, \psi_i, \psi_s, \psi_m, d_{min}, d_{max}\}$, is referred to as the cohort-word selection parameters.

3. Experiment set up

All experiments were performed on a subset of the clean microphone speech Wall Street Journal 1 (WSJ1) test dataset. Speech was parameterised using Mel Frequency Cepstral Coefficient (12 MFCCs + deltas + accelerations) with RASTA post-filtering. HMMs (12-mixture triphones) trained on WSJ1 training data were used for acoustic scoring. A 256-component GMM trained on the WSJ1 training data set was used as the speech background model for SBMV.

3.1 Evaluation word candidate sets

Experiments were performed on 3 different word length classes: 4-phone words, 6-phone words and 8-phone words (where a n-phone word is a word with n phones in its phone transcription). For each phone-length class, 3 word candidate sets were built from the WSJ1 test data set. The first set contained 600 word candidates and was used for fused word-verifier training (the TRAIN set) while the second and third sets contained 300 word candidates each and were used for word-verifier evaluation (the EVAL1 and EVAL2 sets).

Each set consisted of approximately 50% true word candidates and 50% false word candidates. True word candidates were taken from a forced-aligned word transcription of the test data. False word candidates were obtained by taking the false alarm outputs of a keyword spotter. The false alarm outputs were used instead of randomly selected locations in the test audio to increase the difficulty of KV for the false candidates. Since a keyword spotter's false alarms are likely to be at areas that are acoustically similar to the target keyword being spotted, KV of these false alarms should be markedly more difficult than KV of randomly chosen word occurrences. This hypothesis was confirmed in experiments reported in (Thambiratnam, 2003).

3.2 Performance Metrics

Although there are a variety of measures for measuring KV performance (eg. equal error rate and figure of merit), the false rejection (FR) rate at 10% false acceptance (FA) rate was chosen for evaluation. This metric favours keyword spotting applications where it is more important not to incorrectly react to a false word occurrence (eg. command control systems) rather than to accidentally miss a true word occurrence. This operating point is also well suited for large database search applications where one is interested in returning a smaller result set to a user with confident results rather than a large result set with a lot of false occurrences. Additionally Detection Error Trade-off (DET) plots are provided for the final optimal systems to provide performance trends across operating points.

3.3 Base Verification Procedure

Baseline SBMV results were obtained using the confidence score given by equation 2. The SBMV confidence score was calculated for each word candidate in the EVAL1 and EVAL2 sets. Thresholding was then used to calculate FR at 10% FA.

CWV performance was evaluated for a variety of cohort-word selection parameters. To reduce the scope of the experiments, the restrictions $\psi_i = 1$, $\psi_s = 0$, $\psi_m \in \{1,2\}$, $\psi_s \in \{1,2\}$, and $1 \le d_{min} \le d_{max} \le 4$ were used.

For each word candidate in the EVAL1 and EVAL2 sets, the reduced cohort-word set $R'(w,N)$ of size $N=200$ was found using the cohort-word selection procedure. The simplified cohort-word confidence score $C_2(w,K)$ was then calculated using the reduced cohort-word set $R'(w,N)$ and with $K=1$ (this restriction on K was used to reduce computation time). FR was finally calculated at 10% FA by confidence score thresholding.

3.4 Fused Verification Procedure

Word-verifier fusion was performed using a multi-layer perceptron neural network. Confidence scores from the verifiers to be fused were used as the input values. A 25 node hidden layer was used in the intermediary layer and 2 nodes were used in the output layer, one for true occurrences and one for false occurrences. A separate neural network was trained on the TRAIN set for each phone-length using 4-fold cross validation. Evaluation was then performed on the EVAL1 and EVAL2 sets for each phone length. The output value of the true occurrence output node was thresholded to obtain FR at 10% FA.

4. Experiments and Results

Baseline SBMV and CWV performances were measured to provide a benchmark for fused verifier experiments. Results for the baseline systems are given in tables 1 and 2. Although CWV was evaluated for a variety of cohort-word selection parameters (see section 3.3), only the best performing CWV methods are shown to conserve space. Cohort-word selection parameters are specified in the format $\{\psi_d, \psi_i, \psi_s, \psi_m, d_{min}, d_{max}\}$.

Phone Len	FR EVAL1	FR EVAL2	FR Avg.
4	25.3	30.2	27.8
6	7.8	11.8	9.8
8	4.2	4.6	4.4

Table 1: Best SBMV baseline verifier performance

The results show that the absolute gain in FR of CWV over SBMV dropped off as target keyword phone-length increased. While there were positive gains for 4-phone and 6-phone keywords (8.8% absolute for 4-phone, 4.6% absolute for 6-phone), there was a loss in verifier performance over SBMV for 8-phone keywords. This observation motivated SBMV-CWV fusion experiments. It was anticipated that a fused SBMV-CWV verifier could combine the short keyword performance of CWV with the long keyword performance of SBMV. Table 3 shows the results of experiments to evaluate the performance of fused SBMV-CWV keyword verification.

Phone Len	CWV Best Method	FR EVAL1	FR EVAL2	FR Avg.
4	EVAL1 {2,1,3,3}	16.2	28.9	22.6
4	EVAL2 {2,1,2,4}	31.8	17.5	24.7
4	All {1,2,3,3}	17.9	20.1	19.0
6	EVAL1 {2,1,3,4}	1.2	9.2	5.2
6	EVAL2 {1,2,3,4}	10.0	7.5	8.8
6	All {2,1,3,4}	1.2	9.2	5.2
8	EVAL1 {1,1,4,4}	7.2	8.5	7.8
8	EVAL2 {2,1,4,4}	15.6	6.7	11.2
8	All {1,1,4,4}	7.2	8.5	7.8

Table 2: Best CWV baseline verifier performance

The results of these experiments demonstrated that a fused SBMV-CWV verifier was able to provide consistent gains in FR over the baseline SBMV systems as well as the best performing individual CWV systems. For all phone lengths the best overall performing SBMV-CWV system was able to at least halve the false rejection rate of the baseline SBMV system. The figures also showed that the relative gain of SBMV-CWV over the best overall performing CWV baseline verifier increased with phone length (25.3% 4-phone, 30.8% 6-phone, 71.8% 8-phone). This gain with longer phone lengths was consistent with the trends seen in SBMV performance across phone length.

Phone Len	SMBV-CWV Best Method	FR EVAL1	FR EVAL2	FR Avg.
4	EVAL1 {2,1,1,1}	12.2	21.5	16.8
4	EVAL2 {2,2,2,3}	20.9	14.5	17.7
4	All {2,2,1,3}	13.2	15.3	14.2
6	EVAL1 {2,1,3,3}	1.4	7.2	4.3
6	EVAL2 {2,1,2,3}	2.4	4.7	3.6
6	All {2,1,2,3}	2.4	4.7	3.6
8	EVAL1 {1,2,1,4}	1.8	2.6	2.2
8	EVAL2 {1,1,2,3}	8.1	2.4	5.2
8	All {1,2,1,4}	1.8	2.6	2.2

Table 3: Best fused SBMV-CWV verifier performance

Since SBMV performance was poor for short keywords, experiments were performed to evaluate fusion of multiple CWVs. Although improvements in performance were observed for SBMV-CWV over CWV for short keywords, it was hoped that fusing multiple well performing cohort-word verifiers would yield even greater improvements, since the individual CWV systems performed better than SBMV for short-word KV. To reduce the scope of the experiments, fusion of only 2 verifiers at a time was considered. Table 4 shows the best performing CWV-CWV fused verifiers.

Phone Len	CWV-CWV Best Method	FR EVAL1	FR EVAL2	FR Avg.
4	EVAL1 {1,2,3,3},{2,1,4,4}	11.5	18.8	15.1
4	EVAL2 {1,2,3,3},{2,2,2,3}	20.2	8.7	14.5
4	All {1,2,3,3},{2,1,1,3}	13.9	10.1	12.0
6	EVAL1 {2,1,3,4},{1,1,3,4}	1.1	7.2	4.1
6	EVAL2 {2,1,3,4},{1,2,3,4}	2.0	4.3	3.2
6	All {2,1,3,4},{1,2,3,4}	2.0	4.3	3.2
8	EVAL1 {1,1,4,4,},{2,2,1,3}	4.5	7.2	5.9
8	EVAL2 {1,1,4,4},{2,2,4,4}	5.7	5.5	5.6
8	All {1,1,4,4},{2,2,4,4}	5.7	5.5	5.6

Table 4: Best fused CWV-CWV verifier performance

The CWV-CWV architecture seemed to be

particularly well suited for short KV, resulting in a 15.5%, 36.8% and 56.8% relative gain in overall FR over the best SBMV-CWV, CWV and SBMV verifiers respectively. Improvements were also seen for the 6-phone keyword experiments, although the gain over the SBMV-CWV system was much more marginal. CWV-CWV fusion did not yield improved performance for the 8-phone keyword set, though this was expected considering that individual CWV did not outperform SBMV for the 8-phone keyword sets.

A consistent trend that was noted across all experiments was the dependence of CWV performance on cohort-word selection parameters. Although not shown here, performance of CWV varied dramatically with cohort-word selection parameters. For example for the 4-phone EVAL1 set, the best performing CWV configuration had an FR of 16.2% while the average FR across all evaluated CWV configurations was 29.3%. The use of CWV therefore requires careful tuning of the cohort-word selection parameters. However it appears that the relative performance between various CWV configurations remains reasonably consistent between fused and unfused systems. This means that the optimum cohort-word selection parameters for CWV are most likely to give close to optimum performance when used in a fused CWV-CWV or SBMV-CWV system. For example, the best EVAL1 CWV system was only 0.4% poorer when used in a fused CWV-CWV system compared to the best EVAL1 CWV-CWV system. Unfortunately there does not appear to be a single set of cohort-word selection parameters that is optimal for all target word phone lengths.

As previously discussed, short-word KV is a particular difficult problem leading to significantly higher false rejection rates compared to KV for longer target word lengths. Hence the absolute gains observed using the CWV and CWV-CWV for short-word KV are particularly pleasing. The DET plot in figure 4 further demonstrates the benefits of CWV and CWV-CWV. Both methods consistently outperformed SBMV for short-word KV at the majority of operating points.

Overall, the experimental results suggest that an optimum KV system would be keyword phone-length dependent. A CWV-CWV verifier would be used for short length keywords, a SBMV-CWV or CWV-CWV verifier for medium length keywords and a SBMV-CWV verifier for long keywords. Using this approach, a 5.8% overall FR could be obtained for all keyword lengths on the EVAL1 and EVAL2 word candidate sets. This is a significant gain over the 14.0% overall FR using SBMV alone and 10.7\% overall FR using the optimal CWV verifier for each keyword length class.

5. Conclusion

The experiments demonstrated that fused SBMV-CWV and CWV-CWV verification yielded dramatic gains in KV performance over unfused SBMV and CWV. The CWV-CWV system was more suited to short-to-medium length KV while the SBMV-CWV system performed best for medium-to-long length KV. Overall a 5.8% false rejection rate at 10% false acceptance was achieved using a combination of CWV-CWV and SBMV-CWV. This was a considerable gain over the 14.0% false rejection rate achieved by the baseline SBMV system and the 10.7% false rejection rate obtained using the best performing baseline CWV systems.

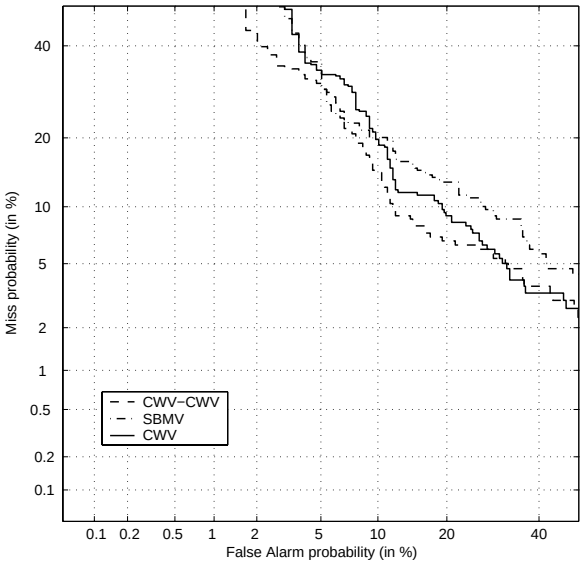

Figure 1: DET plot for best verifiers on combined 4-phone EVAL1 and EVAL2 sets

6. References

[1] L. Mangu, E. Brill, and A. Stolcke, “Finding consensus in speech recognition: Word error minimization and other applications of confusion networks,” *Computer, Speech and Language*, 2000.

[2] K. Thambiratnam and S. Sridharan, “Isolated word verification using cohort word-level verification,” in *Proceedings of Eurospeech 2003*, Geneva, Switzerland, 2003.

[3] L. Xin and B. Wang, “Utterance verification for spontaneous mandarin speech keyword spotting,” in *Proceedings ICII 2001*, Beijing, 2001.

[4] M. Silaghi and H. Bourlard, “A new keyword spotting approach based on iterative dynamic programming,” in *IEEE International Conference on Acoustics, Speech and Signal Processing 2000*, 2000.

[5] J. G. Wilpon, L. R. Rabiner, C. H. Lee, and E. R. Goldman, “Automatic recognition of keywords in unconstrained speech using hidden markov models,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 38, pp. 1870–1878, 1990.

[6] M. J. H. Jurafsky, *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River: Prentice Hall, 2000, ch. Minimum Edit Distance.