AN EMPIRICAL STUDY OF CONFUSION MODELING IN KEYWORD SEARCH FOR LOW RESOURCE LANGUAGES

Murat Saraclar [1,2], *Abhinav Sethy* [1], *Bhuvana Ramabhadran* [1], *Lidia Mangu* [1], *Jia Cui* [1], *Xiaodong Cui* [1], *Brian Kingsbury* [1], *Jonathan Mamou* [3]

[1] IBM T.J. Watson Research Center
Yorktown Heights, N.Y. 10568, USA

[2] Bogazici University, Electrical and Electronic Eng. Dept.
Bebek 34342, Istanbul, Turkey

[3] IBM Haifa Research Labs,
Haifa 31905, Israel

ABSTRACT

Keyword search, in the context of low resource languages, has emerged as a key area of research. The dominant approach in keyword search is to use Automatic Speech Recognition (ASR) as a front end to produce a representation of audio that can be indexed. The biggest drawback of this approach lies in its the inability to deal with out-of-vocabulary words and query terms that are not in the ASR system output. In this paper we present an empirical study evaluating various approaches based on using confusion models as query expansion techniques to address this problem. We present results across four languages using a range of confusion models which lead to significant improvements in keyword search performance as measured by the Maximum Term Weighted Value (MTWV) metric.

1. INTRODUCTION

The goal of keyword search (KWS) is to find all occurrences of a keyword or consecutive sequence of keywords (a *query*), in a large audio corpus. Unlike in keyword spotting [1], in pre-indexed keyword search the corpus to be searched is indexed with no prior knowledge of the query terms. This implies, no knowledge of the query terms is used during audio processing, allowing the systems to be more general. The state-of-the-art speech retrieval systems [2], use an Automatic Speech Recognition (ASR) system as the front-end. The Spoken Term Detection (STD) 2006 evaluation [3], a pilot competition run by the U.S. National Institute of Standards and Technology (NIST) in 2006, introduced the Actual Term-Weighted Value (ATWV) and MTWV metrics for evaluating keyword search systems, in addition to the various metrics to measure misses and false alarms commonly used in the literature [4]. Recently, IARPA introduced the Babel program [5], with the main goal of reducing the performance gap of spoken-term detection systems between high-resource, well- studied languages and low-resource, lightly studied languages. In the limited language pack track of the Babel program, only ten hours of transcribed data is used for building systems. In this paper, we focus on four of the languages used under this program, namely, Pashto, Tagalog, Turkish and Vietnamese. To put our results in perspective, the target ATWV for the Babel program is 0.3 for the full language pack track where systems are built with a factor of 10 more training data. This highlights the challenging nature of this task. The KWS systems in this paper were one of the best performing systems of the program, in the 2013 Babel program evaluation.

Keyword search systems must handle two types of queries, terms that are in the vocabulary of the speech recognition systems (in-vocabulary or IV) and terms that are out-of-vocabulary (OOV). OOV query terms become more important in low-resource tasks, where a limited amount of training data is available. Many approaches to model token (word or sub-word unit) confusability [6, 7] for OOV search have been proposed in the literature. In this work, we derive from our prior research [8, 9, 10, 11] and present a study of the impact of modeling confusability at various stages of the spoken term detection system for these low-resource languages, namely indexing system, query generation and search.

After reviewing the related work in Section 2, we begin with a description of the KWS system in Section 3. The confusion modeling approaches studied in detail are presented in Section 4. The data and ASR systems used in our KWS system are presented in Section 5. Section 6 presents the impact of the confusion models on the different languages. Section 7 concludes with a summary of these results.

2. RELATED WORK

In KWS or STD, OOV queries can contain one or more OOV terms. Several approaches to improve performance on OOV query terms have been proposed in the literature. In this section, we will provide a brief overview of these, focussing specifically on prior work that use ASR systems as the front end to STD systems, and incorporate confusability at the acoustic and linguistic levels. One of the early attempts to use N-best lists and phonetic confusion matrices to compensate for errors made by an ASR system was proposed in [12, 13]. Approaches that combine word and sub-word indexes (phonetic or syllabic) have also been proposed to tackle the OOV problem [14]. Information Retrieval based approaches such as query expansion and stemming [15] have also provided gains in searching for OOV terms. [6] introduced the use of acoustic confusability for handling OOV query terms. Once the query is expanded to its phonetic representations, confusable in-vocabulary phrases were generated using the recognizer's dictionary, a language model and a confusion matrix

This work was done while Murat Saraclar was visiting IBM.

that provided scores for the confusions between phonemes as well as the likelihood of inserting and deleting each phoneme. A vocabulary independent, hybrid LVCSR approach to audio indexing and search showed that using phonetic confusions derived from posterior probabilities estimated by a neural network in the retrieval of OOV queries can help in reducing misses [8]. These confusions can also be HMM based phone confusion estimates [7]. In [16, 10], the authors introduced the WFST based approach to STD. Since then, the standard approach to detecting OOV terms employs this approach, wherein the query terms are converted to a sub-word sequence (usually phonemes) by grapheme-to-phoneme (g2p) rules, a confusion matrix is used to allow for recognizer errors, and this sequence is then searched for in previously-generated subword lattices. Increased false alarms from g2p generation are controlled with various constraints on the g2p and the confusion matrices [10, 9]. In morphologically rich languages such as Turkish, using morphs or morphemes as the sub-word unit has been shown to be effective [17, 18].

In [19], the use of discriminative training to construct a phoneme confusion model is proposed. The criterion used is based on the Figure of Merit (FOM), which is directly related to the KWS performance. The value of pronunciation lexicons for keyword search in low-resource languages was studied in [20] for Tagalog, where the lexicon of the LVCSR system used in key-word search is augmented to cover all the OOV query terms by stitching together fragments of matching entries in a reference lexicon, using prefix and postfix transductions and grapheme to phoneme prediction. While it is not so interesting to expand the lexicon and process the audio to be indexed multiple times, an efficient KWS that maps the OOV term to an IV proxy that is closest to the phonetic representaton of the OOV term was utilized. Similar approach was presented in earlier work for query-by-example search on OOVs [9]. Both these approaches presented significant gains in KWS performance on OOV terms by augmenting lexicons. MediaEval2012's Spoken Web Search task is similar in flavor to the Babel task. The development data available is approximately four hours [21]. However, no ASR system was used in that work and issues with false alarms and misses arose from acoustic confusability.

3. KWS SYSTEM ARCHITECTURE

The KWS system considered in this paper is based on Weighted Finite State Transducers (WFSTs). In the WFST approach, the output of the ASR system (typically lattices) is represented as a WFST where input arc labels are words and the arc weights are ASR scores. Timing information is represented as output arc labels consisting of begin and end time pairs. An index of all possible substrings contained in the input WFSTs is built using the algorithm proposed in [16] and extended to include timing information in [22]. The resulting index itself is a WFST mapping substrings to utterance numbers and occurrence times together with the posterior probability of occurrence. The queries are also represented as WFSTs. Retrieval consists of composing each query WFST with the index.

The KWS system makes use of a word (or token) index and a phonetic index. The word (or token) index is obtained directly from the output of the ASR system whereas the phone index is obtained by first mapping the ASR output to the phone level. The use of a phone index is akin to document expansion used in information retrieval.

IV queries are directly searched using the word level index. For OOV queries we first convert the query to phones (using a grapheme-to-phoneme converter) and then search the phonetic index. Alternatively the phonetic representation of OOV queries is mapped back to a word representation using the inverse of the ASR pronunciation lexicon transducer, and the new query is matched against the word level index.

Although the use of lattices makes the KWS system more robust to ASR errors, it is not always possible to find exact matches to the queries. To allow inexact matches we use the idea of query expansion to create alternate representation of the query. In the WFST framework, query expansion is achieved by composing the query with a confusion model (which is also represented as a WFST) before composing it with the index. In the next section we describe the confusion models used in this study.

For each query, the output of the KWS system is a list of hits ranked by the posterior probability of occurrence. Since it is desirable to set a global detection score threshold in order to limit the number of hits for each query, the final scores are normalized for each query. For the ATWV metric, the optimal term specific threshold (TST) can be determined if the number of occurrences of the query in the test corpus is known apriori. It was shown that the sum of posterior probabilities over the test corpus is a reasonable approximation to the number of occurrences for the purpose of TST [23]. As an alternative to normalizing the posterior probabilities by the TST, we normalize the posterior probabilities of occurrence by their sum over the test corpus. This Sum-To-One (STO) normalization performs similar to the TST normalization.

4. CONFUSION MODELS

A key component of our KWS system is the confusion model which allows us to do a fuzzy or inexact search. This is especially important in the context of this paper, where we focus on systems built with limited resources. For such systems, the ASR performance tends to be poor (high WER) and the system vocabulary is limited. The small vocabulary of the ASR system implies that a large fraction of query terms are OOV. The poor quality of ASR output implies that we will miss occurences of many query terms. By creating alternate representation of query terms in word and phonetic forms, a confusion model allows us to recover misses and handle OOV terms, albeit at the risk of increased false alarms.

We consider the following confusion models.

- Phone to Phone transducer (**P2P**): This confusion model is used in conjunction with phonetic indices. To create this model, Viterbi alignments of the training data are obtained from the transcriptions using any acoustic model [1]. The same acoustic model is used to decode the training data with a unigram word LM. State-level confusability is computed by comparing the two sets of alignments from the ground truth and decoding hypotheses, respectively, which is then converted to phone-level confusability. The training set has approximately 10 hours of audio data.
- Token to Token transducer (**T2T**): Word level alignments of the decoded hypothesis and training text generated for the P2P model, are used to build a token level transducer (T2T) as well. The T2T transducer models ASR substitution errors and can be applied to word indices directly.
- Phone To Phone to Token transducer (**P2P2T**): The P2P model can be used directly on word indices by composing it with the (inverted) decoder dictionary so that it maps phonetic strings to ASR tokens (P2P2T), i.e. finds IV proxies.

[1] We used a speaker independent (SI) deep neural network (DNN).

Language	Training Data			Development Data
	(hours)	(tokens)	(types)	(hours)
Pashto	11	117 K	7 K	10
Tagalog	11.5	73 K	6 K	10
Turkish	12	79 K	12 K	10
Vietnamese	11	122 K	3 K	10

Table 1: Data statistics for each language. *types* refers to number of distinct tokens

Language	WER	LatDensity Deep	LatDensity Shallow
Pashto	63.7	33953	1761
Turkish	65.7	23273	798
Tagalog	64.5	12865	497
Vietnamese	70.2	58513	1665

Table 2: Average number of arcs per second (**LatDensity**) and WER for deep and shallow lattices

These confusion models are used in our keyword search architecture by way of query expansion. We take the transducer representation of the query term and compose it with the confusion model. We derive n-best paths from the composed transducers. The n-best paths can be considered as alternate plausible representations of the query term. For the P2P2T model, the n-best operation can be done after the composition with the inverted lexicon or before. If the n-best is done as the final step after the composition with lexicon we refer to the model as P2P2T and if it is done right after the P2P model we refer to it as **P2P2TN**.

5. DATA AND ASR SYSTEM DESCRIPTIONS

In this study, data from the Babel limited language pack track for four languages, namely Pashto, Turkish, Tagalog, and Vietnamese are used. The information about the data for these languages is given in Table 1.

The acoustic model used in these experiments for all languages is the IBM Speaker-Adapted DNN (SA DNN) system which uses a deep neural network (DNN) acoustic model with the standard front-end pipeline [24]. The DNN takes 9 frames of 40-dimensional speaker adapted discriminative features as input, contains 5 hidden layers with 1,024 logistic units per layer, and has a final softmax output with 1,500 targets. Training occurs in three phases: (1) layer-wise discriminative pre-training using the cross-entropy criterion, (2) stochastic gradient descent training using back-propagation and the cross-entropy criterion, and (3) distributed Hessian-free training using the state-level minimum Bayes risk criterion [25]. The lexicon is the same as that provided with the corpus (contains words in the training data only). The language model (LM) is a trigram LM with modified Kneser-Ney smoothing that was trained on the acoustic transcripts. Deep and shallow lattices were generated using a static decoder by adjusting the beam parameters. Silences and hesitations were treated as transparent words to ensure more unique words appeared in the lattice. The lattice sizes and the corresponding one-best Word Error Rate (WER) are presented in Table 2.

Language	Total	No exact match	OOV
Pashto	967	243	215
Turkish	307	94	83
Tagalog	353	89	80
Vietnamese	200	87	28

Table 3: Number of queries with no hits without query expansion, number of oov queries and total number of queries for each language

6. RESULTS

In this section, we present our findings on using confusion models for keyword search. We first look at the following dimensions of confusion as they are used in our KWS system:

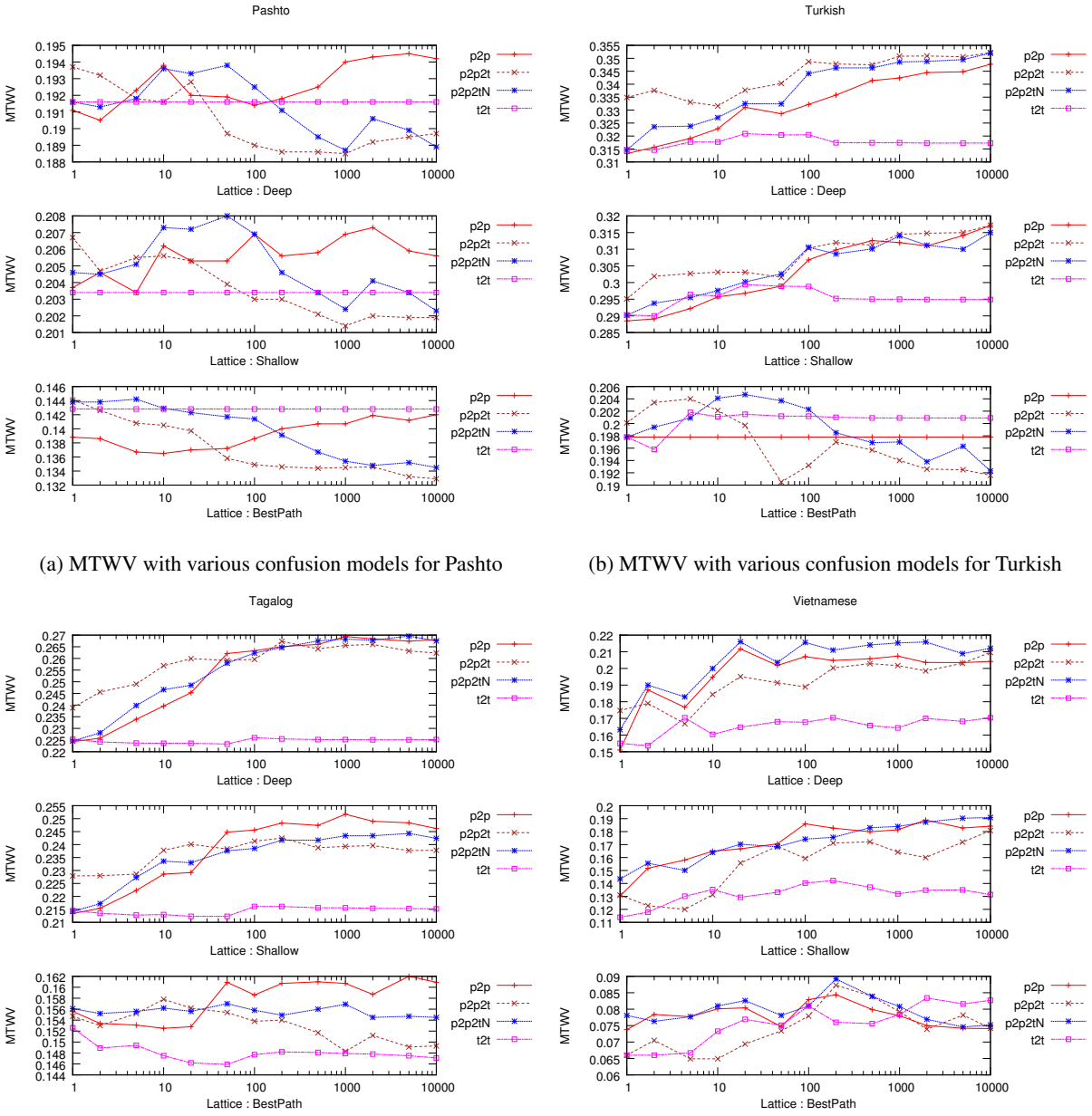
- Type of confusion model: Phonetic confusion (P2P), Token confusion (T2T), and Phonetic confusion converted to token confusion (P2P2T).
- Depth or strength of the confusion model: We consider different depths of confusion (n-best) ranging from 1 to 10000. For the P2P2T confusion model we consider the two variants in Section 4.
- Depth of lattices: We consider deep and shallow lattices along with best-path output generated from deep lattices. See Section 5 for details on how the lattices were generated.

Our results for Pashto (Figure 1a), Turkish (Figure 1b), Tagalog (Figure 1c) and Vietnamese (Figure 1d) show that good improvements in MTWV can be expected with confusion models. Each subplot in these figures provides MTWV performance when using the one-best hypothesis, shallow lattices and deep lattices for keyword search in combination with different confusion models. Here are the observations that generalize across these languages:

- Phone to Phone to Token (P2P2T) models work as effectively as pure phonetic (P2P) models. Unlike the P2P model which requires a phone level index, the P2P2T model operates from a word level index. Since the two perform roughly the same we can significantly simplify the computational architecture and resource requirements by keeping just one word level index and using P2P2T.
- Deeper lattices are a better way to improve performance compared to query expansion. However the additional gains from confusion modeling appear to hold over both shallow and deep lattices indicating that confusion models and lattice depth are complimentary.
- While choosing the optimal confusion model depth is tricky, it is better to err on the side of deeper models.

Table 3 lists the total number of query terms, OOV terms and IV terms for which no exact match was found in word search. We use the search cascade described in Section 3. Table 4 shows the best performing models and their MTWVs for all the four languages. We consider MTWV for the full set of queries as well as subsets containing just the in-vocabulary (IV MTWV) and out-of-vocabulary (OOV MTWV) terms. Note that the values in the MTWV columns are averaged over the respective subsets (Full, IV, OOV). Thus the MTWV's of just IV terms or OOV terms can be higher then the MTWV for all queries.

Further analysis on the impact of confusion models are presented on one language, Tagalog. When searching with confusion



(a) MTWV with various confusion models for Pashto

(b) MTWV with various confusion models for Turkish

(c) MTWV with various confusion models for Tagalog

(d) MTWV with various confusion models for Vietnamese

Fig. 1: Variation of MTWV with n-best for various confusion models

models, it is desirable that the lattices are phonetically closer to the audio so that we can find matches for out of vocabulary and infrequent terms. By increasing the acoustic weight we can improve the phonetic match but this tends to degrade performance for in-vocabulary terms. We use two sets of lattices generated with different acoustic weights. The first set of lattices, which are generated with an acoustic weight that minimizes the WER, are used for IV terms and for OOV queries we use lattices generated with higher acoustic weight. For Tagalog, the optimal acoustic weight was 0.13 for WER. By using a higher acoustic weight for OOV terms we could improve the MTWV from 0.2696 to 0.2786, as shown in Table 5.

Figure 2 shows the DET curves with optimal n for the various confusion models. As can be seen from the plots, the P2P model performs better at high false alarm regions whereas the P2P2T and P2P2TN confusion models are better at low false alarm regions. For MTWV as defined in the Babel program, they perform equally well.

We also found that score normalization plays a critical role in improving MTWV scores with confusability models. Figure 3 shows the False Alarm(FA) vs Miss(PMiss) tradeoff for P2P search with different n-best depths with sum-to-one normalization (STO) which is the default for our system. The DET curves show that with higher N its possible to reduce misses at cost of some FA and achieve an

Language	Model	MTWV	IV MTWV	OOV MTWV
Pashto	identity	0.1916	**0.2127**	0
	p2p	**0.1945**	0.2123	**0.0330**
	t2t	0.1916	**0.2127**	0.0000
	p2p2t	0.1937	**0.2127**	0.0209
	p2p2tN	0.1938	0.2126	0.0222
Turkish	identity	0.3147	0.4287	0
	p2p	0.3477	**0.4457**	0.0792
	t2t	0.3209	0.4371	0.0000
	p2p2t	**0.3522**	0.4349	**0.1147**
	p2p2tN	0.3520	0.4363	0.1115
Tagalog	identity	0.2253	0.2907	0
	p2p	0.2694	0.2906	0.1887
	t2t	0.2260	0.2897	0.0000
	p2p2t	0.2673	**0.2912**	0.1777
	p2p2tN	**0.2696**	0.2908	**0.1904**
Vietnamese	identity	0.1550	0.1802	0
	p2p	0.2117	**0.2067**	0.2421
	t2t	0.1704	0.1981	0.0000
	p2p2t	0.2096	0.1964	**0.2728**
	p2p2tN	**0.2160**	0.2049	0.2490

Table 4: Best MTWV for all query terms, IV terms and OOV terms with different confusion models. *identity* refers to exact search in word lattices

Ac wt	MTWV
0.13	0.2696
0.14	0.2762
0.15	0.2786
0.16	0.2768

Table 5: MTWV vs acoustic weight for Tagalog

overall improvement in MTWV. However when we look at the DET curves without STO (Figure 4) we find that the FA/Miss tradeoff with increasing N does not allow for improving MTWV.

7. CONCLUSION

We summarize our key findings as follows:

- For limited resource KWS systems, the performance can be improved by using a confusion model. Given the limited amount of data a phone level confusion model is preferable.
- Among the alternatives for applying the confusion model, there does not seem to be a clear winner. However, mapping the queries into a phonetic representation, applying the confusion model and mapping back to words has the advantage of requiring only a word level index.
- The size of the lattices plays an important role, especially in the limited resource case where error rates are high.
- Score normalization plays a important role in getting gains from confusion models.
- Finally, increasing the confusability by optimizing the acoustic weight to maximize MTWV is a promising approach.

Acknowledgement

This study uses the IARPA Babel Program base period language collection releases babel104b-v0.4bY, babel105b-v0.4, babel106b-v0.2g, and babel107b-v0.7, supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ARL) contract number W911NF-12-C-0012. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

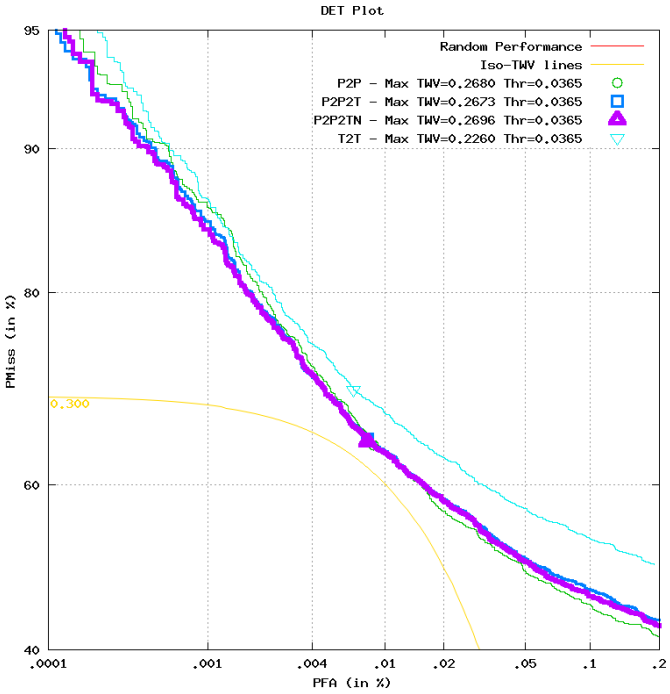


Fig. 2: DET curves with various confusion models for Tagalog. The n-best depth for each model was optimized for MTWV.

8. REFERENCES

[1] R. C. Rose and D. B. Paul, "A Hidden Markov Model based keyword recognition system," in *Proc. ICASSP*, 1990, pp. 129–132.

[2] C. Chelba, T. J. Hazen, and M. Saraçlar, "Retrieval and browsing of spoken content," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 39–49, 2008.

[3] "Spoken term detection evaluation," http://www.nist.gov/speech/tests/std/.

[4] J. G. Fiscus, J. G. Ajot, J. Garofalo, and G. Doddington, "Results of the 2006 spoken term detection evaluation," in *Proc. SIGIR Workshop on Searching Spontaneous Conversational Speech*, 2007, pp. 51–57.

[5] "IARPA broad agency announcement IARPA-BAA-11-02," 2011.

[6] B. Logan and J. M. Van Thong, "Confusion-based query expansion for OOV words in spoken document retrieval," in *Proc. ICSLP*, 2002, pp. 1997–2000.

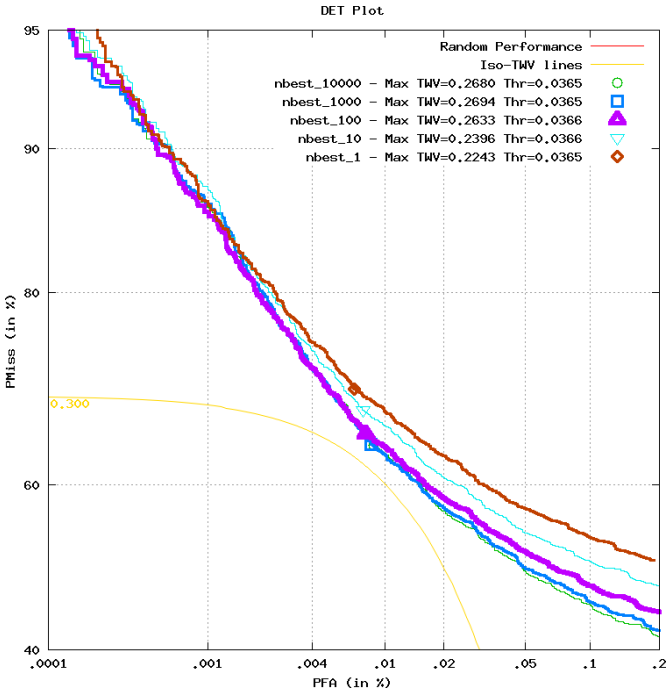


Fig. 3: DET curves with p2p with different N for Tagalog with STO

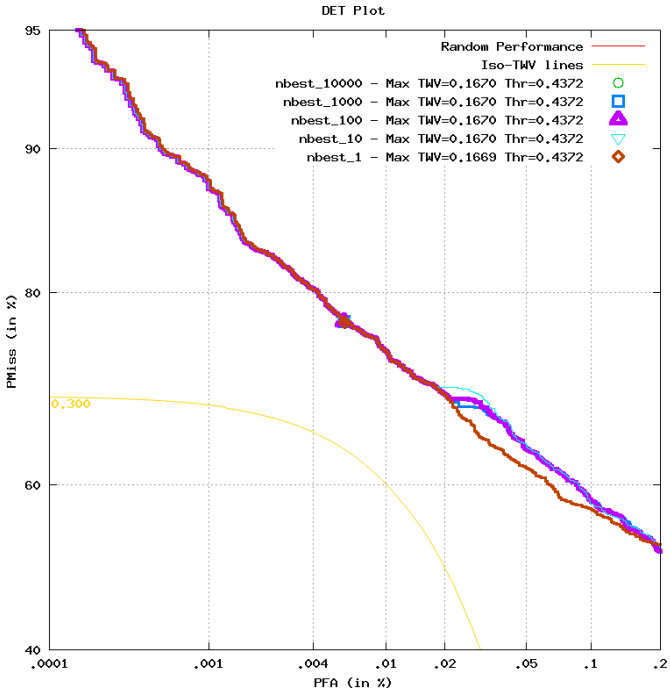


Fig. 4: DET curves with p2p with different N for Tagalog without STO.

[7] U. V. Chaudhari and M. Picheny, "Improved vocabulary independent search with approximate match based on conditional random fields," in *Proc. ASRU*, 2009, pp. 416–420.

[8] B. Ramabhadran, A. Sethy, J. Mamou, B. Kingsbury, and U. Chaudhari, "Fast decoding for open vocabulary spoken term detection," in *Proc. HLT-NAACL*, 2009, pp. 277–280.

[9] C. Parada, A. Sethy, and B. Ramabhadran, "Query-by-example spoken term detection for OOV terms," in *Proc. ASRU*, 2009, pp. 404–409.

[10] D. Can, E. Cooper, A. Ghoshal, M. Jansche, S. Khudanpur, B. Ramabhadran, M. Riley, M. Saraclar, A. Sethy, M. Ulinski, and C. White, "Web derived pronunciations for spoken term detection," in *Proc. SIGIR*, 2009, pp. 83–90.

[11] B. Kingsbury et al, "A high-performance Cantonese keyword search system," in *Proc. ICASSP*, 2013, pp. 8277–8281.

[12] D. A. James, "A system for unrestricted topic retrieval from radio news broadcasts," in *Proc. ICASSP*, 1996, pp. 279–282.

[13] S. Srinivasan and D. Petkovic, "Phonetic confusion matrix based spoken document retrieval," in *Proc. SIGIR*, 2000, pp. 81–87.

[14] J. Mamou, B. Ramabhadran, and O. Siohan, "Vocabulary independent spoken term detection," in *Proc. SIGIR*, 2007, pp. 615–622.

[15] P. C. Woodland, S. E. Johnson, P. Jourlin, and K. Sparck Jones, "Effects of out of vocabulary words in spoken document retrieval," in *Proc. SIGIR*, 2000, pp. 372–374.

[16] C. Allauzen, M. Mohri, and M. Saraclar, "General indexation of weighted automata - application to spoken utterance retrieval," in *Proceedings of the HLT-NAACL Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval*, 2004, pp. 33–40.

[17] E. Arisoy, D. Can, S. Parlak, H. Sak, and M. Saraçlar, "Turkish broadcast news transcription and retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 874–883, 2009.

[18] S. Parlak and M. Saraçlar, "Performance analysis and improvement of Turkish broadcast news retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 731–741, 2012.

[19] P. Karanasou, L. Burget, D. Vergyri, M. Akbacak, and A. Mandal, "Discriminatively trained phoneme confusion model for keyword spotting," in *Proc. Interspeech*, 2012.

[20] G. Chen, S. Khudanpur, D. Povey, J. Trmal, D. Yarowsky, and O. Yilmaz, "Quantifying the value of pronunciation lexicons for keyword search in low resource languages," in *Proc. ICASSP*, 2013, pp. 8560–8564.

[21] F. Metze, X. Anguera, E. Barnard, M. Davel, and G. Gravier, "The spoken web search task at MediaEval 2012," in *Proc. ICASSP*, 2013, pp. 8121–8125.

[22] D. Can and M. Saraclar, "Lattice indexing for spoken term detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2338–2347, 2011.

[23] D.R.H. Miller, M. Kleber, C-L. Kao, O. Kimball, T. Colthurst, S.A. Lowe, R.M. Schwartz, and H. Gish, "Rapid and accurate spoken term detection," in *Proc. Interspeech*, 2007, pp. 314–317.

[24] H. Soltau, G. Saon, and B. Kingsbury, "The IBM Attila speech recognition toolkit," in *Proc. SLT*, 2010, pp. 97–102.

[25] B. Kingsbury, T.N. Sainath, and H. Soltau, "Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization," in *Proc. Interspeech*, 2012.