INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Enhancing Data-Driven Phone Confusions Using Restricted Recognition

Mark Kane[1], *Julie Carson-Berndsen*[2]

[1] Daon, Ireland
[2]School of Computer Science, University College Dublin, Ireland
mark.kane@daon.com, julie.berndsen@ucd.ie

Abstract

This paper presents a novel approach to address data sparseness in standard confusion matrices and demonstrates how enhanced matrices, which capture additional similarities, can impact the performance of spoken term detection. Using the same training data as for the standard phone confusion matrix, an enhanced confusion matrix is created by iteratively restricting the recognition process to exclude one acoustic model per iteration. Since this results in a greater amount of confusion data for each phone, the enhanced confusion matrix encodes more similarities. The enhanced phone confusion matrices perform demonstrably better than standard confusion matrices on a spoken term detection task which uses both HMMs and DNNs.

Index Terms: Phone-based speech recognition, confusion matrices, *a posteriori* phone likelihoods, spoken term detection

1. Introduction

Despite the continuing improvements to speech recognition algorithms from Hidden Markov Models (HMMs) [1] to Deep Neural Networks (DNNs) [2], variation continues to represent a challenge for speech technology applications. Variation can arise as the result of the pronunciation of the speaker (both native and non-native) [3–5], noise [6, 7] or even as a result of inaccuracies produced by the phone-based recogniser [8, 9], all of which may cause a phone to deviate from its canonical specification. From the linguistic perspective, this variation can be modelled with phonological rules which map between phonetic variants (or allophones) and phonological forms (or phonemes). In speech technology, this variation is addressed to some extent through the use of confusion statistics which are calculated by the speech recogniser and applied *a posteriori* to improve the recognition performance. The confusions capture the extent to which one phone is confused with another during recognition and is thus regarded as representing some notion of similarity among the confused phones. Applications of phone-based confusions within speech technologies are often found in Information Retrieval (IR) systems, such as spoken document retrieval and spoken term detection [10–15].

A standard tool used in spoken term detection and speech recognition for quantifying variation is the phone confusion matrix [14] [16–19] which captures the confusion statistics between phones thus providing a way of defining commonalities or groups [20–22]. However, a confusion matrix can suffer from data sparseness due to the fact that although some phones may be phonetically similar, only a small number of confusions may be found with one or more other phones. This paper presents an enhancement to the standard confusion matrix which addresses data sparseness by incrementally constraining the recognition space, a process which will henceforth be referred to as *restricted recognition*. It will be shown that the enhanced confusion matrix based on *restricted recognition* better represents phone similarity compared to the standard confusion matrix. The contribution of this paper lies in the improvement in quantification of *a posteriori* phone confusions helping to address variation in speech technology. An evaluation of the enhanced phone confusion matrix benchmarked with respect to the standard phone confusion matrix and a matrix of binary confusions is carried out in the context of a spoken term detection experiment across two different corpora using HMMs and DNNs.

The remainder of the paper is structured as follows: in Section 2 the method for calculating the standard phone confusion matrix and the enhanced phone confusion matrix based on *restricted recognition* is presented. Section 3 describes the spoken term detection experiment used to evaluate the two types of confusion matrix with the results of the experiment presented in Section 4. Section 5 discusses the resultant enhanced phone confusion matrix with respect to sparseness and conclusions are drawn in Section 6.

2. Confusion Matrix Calculation

For standard phone-based confusion matrix calculation, the hypothesised strings of phones per utterance from a phone recogniser are compared against a reference phone transcription so that differences can be quantified. A confusion matrix, CM, is an N *by* M matrix where N is the quantity of phone *types* applied to the recognition process i.e. the acoustic phone models, and M is the quantity of phone types within the reference transcriptions. In a standard phone confusion matrix, N and M are equal as all phone types in the reference transcriptions are represented by acoustic phone models. By comparing the recognised string of phones for an utterance to that of its reference transcription, the matrix is then populated with the frequency of occurrence of confused phones. In this experiment, phone-based confusion matrices are creating using HTK's HResults [23].

2.1. Enhancement with Iterative Restricted Recognition

A strong biasing factor of phone confusion generation is the accuracy of the phone recogniser, whereby increasing accuracy results in a decreasing number of confusions. However, with *restricted recognition*, it is possible to generate a larger number of confusions compared to standard recognition. Within *restricted recognition* confusions can be regarded as primarily intentional rather than as a result of recognition error. *Restricted recognition* removes the acoustic model for one phone from the recognition process in order to ascertain what other phones are recognised in its place. Hence that particular phone will never be recognised, instead another phone is recognised that is considered to be similar. *Restricted recognition* is undertaken iteratively, removing one acoustic phone model at a time building

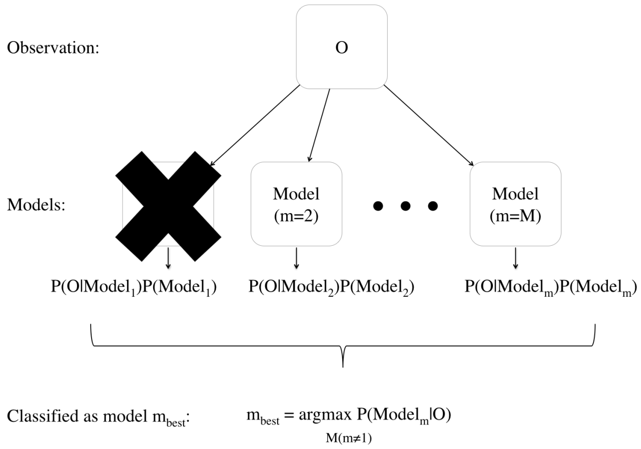
http://dx.doi.org/10.21437/Interspeech.2016-489

the confusion matrix incrementally i.e. for each acoustic model,

1. exclude the model from the recognition process and recognise using the remaining models
2. insert the phone confusions for the transcription phone represented by the excluded model into the enhanced confusion matrix

This means that for the acoustic phone model which has been excluded from the recognition process, an alternative model will always be used to recognise that phone within an utterance. In practice, for each iteration of *restricted recognition* a standard confusion matrix is calculated, however only row *m* is used to populate the enhanced confusion matrix where *m* is the index of the acoustic model excluded during that recognition cycle.In Figure 1, *restricted recognition* is illustrated within a Bayesian framework where a particular acoustic phone model, $Model_m$ where $m = 1$, is excluded during the recognition of an observation. This results in the recognition process always selecting an alternative acoustic phone model in place of the excluded model thus generating substitutions for that phone with respect to the reference transcription whilst maintaining the integrity (recognition accuracy) of the other acoustic phone models.

Figure 1: *Restricted recognition excluding model $Model_m$ where $m = 1$*



Since *restricted recognition* is applied to each acoustic phone model, there are M iterations of the process. When the iterative *restricted recognition* is complete, the enhanced confusion matrix is fully populated. The diagonal of the enhanced confusion matrix contains only zeros since the excluded acoustic phone model meant that the reference transcription phone could never be recognised correctly as itself. However, *restricted recognition* ensures that the maximum amount of confusions are generated for each phone.

Clearly *restricted recognition* requires a corpus in order to train the acoustic models and a speech recognition method to generate the matrix.

2.2. Training Corpus and ASR

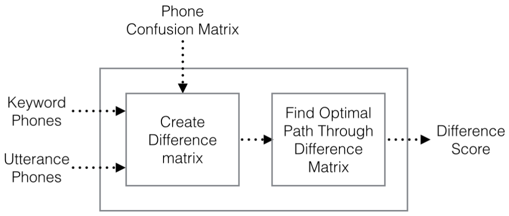
The training corpus used to create the confusion matrices is the widely-used TIMIT training data set [24] as each audio file comes with a gold standard, hand-verified phone transcription. This corpus comes pre-divided into a phonetically balanced training and test set comprised of 630 speakers of 8 major American-English dialects. The training data set contains 3696 audio files. Each audio file contains a sentence length utterance. Prior to phone recognition and phone confusion matrix generation, acoustic phone models are trained using the audio belonging to the training data. The tool used for training the acoustic models is the Hidden Markov Model Toolkit version 3.41 (HTK) [23]. For completeness and repeatability, the training strategy used within HTK is based on Cantab Research's training script [25], and the tutorial within the HTK manual, [23]. This configuration was found to yield a phone recognition accuracy of approximately 95.29% based on training and evaluating within the same training dataset. This accuracy indicates that the acoustic models are sufficiently trained with respect to the training data. Furthermore, the accuracy reported here indicates the quantity of available confusion data within a standard confusion matrix is approximately 5%. Of course phone recognition accuracy on test sets is considerably lower; however, confusion likelihoods determined from the training set are used here to address phone variation within the spoken term detection test sets. The following section describes how confusion matrices are evaluated using spoken term detection.

3. Confusion Matrix Evaluation

3.1. Spoken Term Detection Using Confusion Matrices

In this section a spoken term detection experiment is presented which evaluates the performance of 3 types of phone-based confusion matrices. The detection of spoken terms as keywords is implemented using dynamic programming by aligning keyword and utterance phones so that an alignment difference score can be calculated as illustrated in Figure 2.

Figure 2: *Overview of Dynamic Programming Environment.*



A keyword-utterance difference matrix, D, is defined as a X *by* Y matrix, where X is the total number of phones in an utterance containing several words and Y is the total number of phones in a keyword. For each entry within D, a difference score is calculated between the keyword phone, $k[y]$, and an utterance phone, $u[x]$. The probability of a keyword phone being confused with an utterance phone is calculated using the contents of an N by M confusion matrix, CM, as shown below:

$$P(phone[m]|phone[n]) = \frac{CM[n,m]}{\sum_{i=1}^{N} CM[i,m]} \quad (1)$$

For $phone[m] = k[y]$ and $phone[n] = u[x]$ the difference matrix value is calculated as:

$$D[x,y] = 1 - P(k[y]|u[x]) \quad (2)$$

Note that where $k[y]$ and $u[x]$ are the same then $P(k[y]|u[x]) = 1$ i.e. 100% similar. Once a difference matrix is fully populated, the optimal alignment path through the difference matrix is found using a recursive search governed by linear constraints: $D[x+1, y+1], D[x+1, y], D[x, y+1]$.

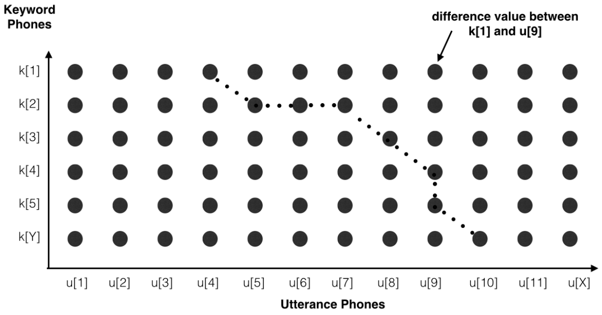


Figure 3: *Optimal path within difference matrix*

This search ensures that the optimal alignment path is always found. The optimal alignment path within a difference matrix is the subset of the path which yields the best alignment between the keyword and the utterance phones as illustrated in Figure 3. The difference matrix values at the coordinates within the optimal alignment path are summed to determine the detection difference score. For each utterance within the search space, a detection score is calculated for a keyword. In the following section, different search spaces and phone recognition algorithms to decode these search spaces are presented.

3.2. Search Space

The spoken term detection experiment is repeated using different phone recognition algorithms and different acoustic search spaces. The latter includes the test set of the TIMIT and NTIMIT [26] corpora. Each test set contains 1344 utterances. NTIMIT was collected by passing the original TIMIT dataset through multiple channels across a telephone network over a telephone handset and thus has degraded acoustic quality with respect to TIMIT which increases the difficulty of the recognition process resulting in more recognition errors. These search spaces are decoded into phones using two different phone recognition algorithms: HTK's HMM [23], and KALDI's DNN [27]. For repeatability, KALDI's TIMIT recipe is used. The phone sequences for the keywords are determined from a canonical lookup table, namely the TIMIT word-to-phone dictionary. The keywords in this search space have the following properties 1) they appear at least five times within TIMIT's (and NTIMIT's) test set and 2) canonically they contain at least five phones. In total this yields 202 keywords within each of the test sets. The average word count of each of the keywords is 7.43 out of a possible $11,025$ word instances across 1344 utterances.

4. Results

In this section, false alarm and miss probability Equal Error Rates (EERs) from a detection error trade-off (DET) curve [28] are reported for the spoken term detection experiment using binary, standard and enhanced phone confusion matrices. The binary matrix, which does not take degrees of similarity into account, is included in order to illustrate the impact of using *a posteriori* phone confusion likelihoods. A summary of the keyword detection results and their associated search space accuracies with respect to different confusion matrices can be found in Table 1.

Table 1: *Spoken term detection EERs for different confusion matrices across different datasets and speech recognition algorithms. Values in parenthesis are discussed in Section 5.*

Acoustic Search Space Information			Keyword Detection EER (%)		
Search Space	Rec. Alg.	Rec. Acc. (%)	Binary	Standard	Enhanced
TIMIT	Manual	100	8.59	3.39 (3.52)	2.06
TIMIT	DNN	75.8	13.54	8.98 (9.05)	6.71
TIMIT	HMM	64.63	16.53	14.46 (13.97)	10.93
NTIMIT	DNN	63.2	22.06	17.2 (16.6)	13.6
NTIMIT	HMM	49.13	26.9	21.76 (21.72)	18.53

This table also includes EERs for a spoken term detection experiment using an ideal search space based on TIMIT's test reference phone transcriptions to simulate a recogniser with perfect accuracy. The results in Table 1 show that the EER of the spoken term detection experiment is proportional to several factors: the quality of the speech within a dataset, the accuracy of the recognition engine and the accuracy of the confusion matrix. The impact on the EER from the speech quality is evident between the TIMIT and NTIMIT datasets whereby, for each corresponding recognition algorithm and confusion matrix, the acoustically degraded NTIMIT dataset always performs worse. The impact of recognition accuracy is evident across the entire table whereby the better the recognition accuracy of the search space, the better the keyword detection result. Finally, for each confusion matrix across different quality datasets, recognition algorithms and search space recognition accuracy, the keyword detection EER from the enhanced confusion matrix is consistently better. False alarms and miss probabilities beyond that of the EERs given in Table 1 are presented in Figure 4, Figure 5, Figure 6 and Figure 7 via NIST DET curve plotting tools [29]. From these DET curves, the spoken term detection experiment using the enhanced confusion matrix performs the best over a range of values.

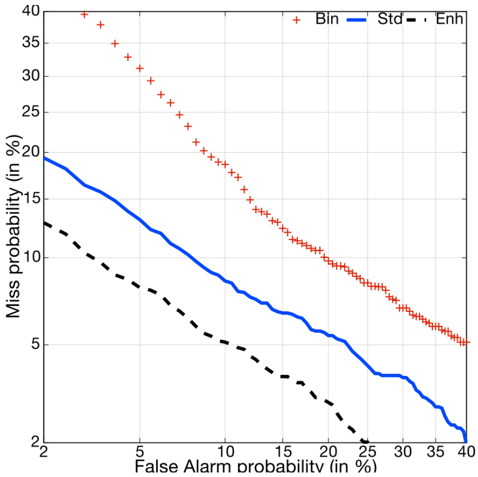


Figure 4: *DET Curves For TIMIT Test Search Space Using DNNs. EER (%) Per Confusion Matrix: Binary = 13.54%, standard = 8.98%, enhanced = 6.71%*

5. Discussion

In order to demonstrate that the enhanced phone confusion matrix not only performs better on the spoken term detection task, but also offers a solution to the data sparseness problem associated with the standard confusion matrix, it is necessary to look

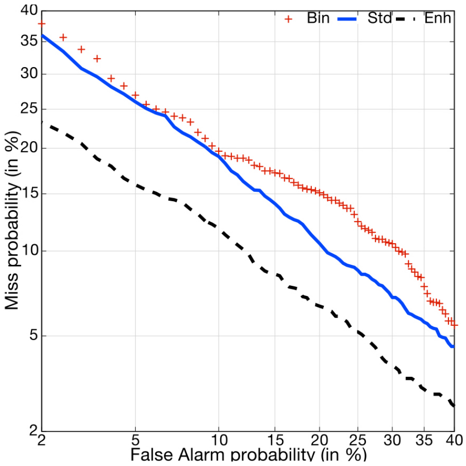


Figure 5: *DET Curves For TIMIT Test Search Space Using HMMs. EER (%) Per Confusion Matrix: Binary = 16.53%, standard = 14.46%, enhanced = 10.93%*

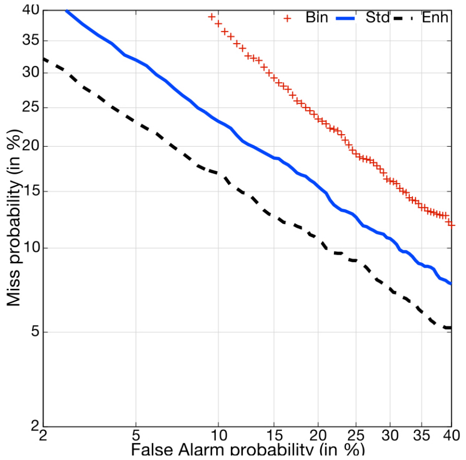


Figure 6: *DET Curves For NTIMIT Test Search Space Using DNNs. EER (%) Per Confusion Matrix: Binary = 22.06%, standard = 17.2%, enhanced = 13.6%*

more closely at the matrices themselves. Since both matrices are too large to present in full here, for the purposes of illustration, only the top confusion for each phone is presented in Table 2. From this table it is clear that the number of confusions captured in the enhanced confusion matrix (using 100% of confusions) is far greater than the standard confusion matrix (using only ≈5% of confusions). Whilst there are some differences between the top confusion labels, it is clear that the confusion quantities are very different and have an impact on the results of the spoken term detection task. For low resource spoken term detection applications, low quantities of confusions as found in the standard confusion matrix are likely to impact performance; this can be overcome through the use of *restricted recognition* to calculate an enhanced confusion matrix.

To increase the percentage of confusions captured within a standard confusion matrix, a development test set was also investigated for standard confusion matrix calculation. Spoken term detection tasks were then repeated with a standard confusion matrix using the TIMIT test set as a best case development set for confusion matrix calculation. In Table 1 spoken term detection EERs based on a standard confusion matrix calculated from a development set are presented within parenthesis. However, the enhanced confusion matrix still results in better EERs.

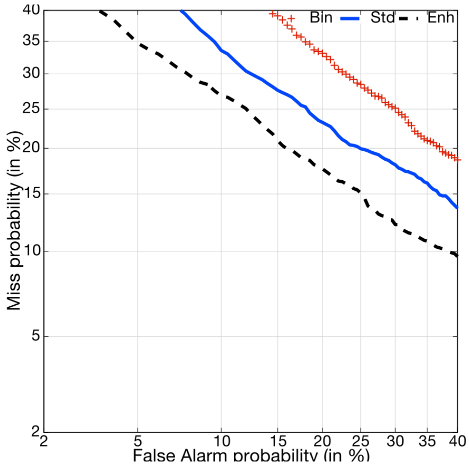


Figure 7: *DET Curves For The NTIMIT Test Search Space Using HMMs. EER (%) Per Confusion Matrix: Binary = 26.9%, standard = 21.76%, enhanced = 18.53%*

Table 2: *Reference phone quantities within the TIMIT training set and their top confusions and quantities within the standard (Std) and enhanced confusion (Enh) matrices.*

	Top Confusion Qty.			Top Confusion Qty.	
Ref. Phone	Std	Enh	Ref. Phone	Std	Enh
aa (2256)	ao,ah (5)	ao (1043)	ix (7370)	ih(71)	ih (3895)
ae (2292)	eh (10)	eh (236)	iy (4626)	ix (9)	ix (1103)
ah (2266)	ax (20)	ax (500)	jh (1013)	ch (10)	ch (636)
ao (1865)	aa (6)	aa (878)	k (4489)	g (17)	g (2128)
aw (728)	aa (4)	aa (209)	l (4425)	el (13)	el (3549)
ax (3535)	ix (55)	ix (1162)	m (3442)	n,em (4)	em (2089)
axh (357)	ix (15)	ix (85)	n (6219)	nx (20)	en (2417)
axr (2445)	r (13)	er (1391)	ng (1194)	n (5)	n (602)
ay (1934)	ah (3)	aa (280)	nx (677)	n (13)	n (487)
b (2429)	p (18)	p (753)	ow (1653)	eh,ax (2)	l (288)
ch (820)	jh (13)	jh (501)	oy (304)	eh (1)	ey (85)
d (4451)	t (61)	t (2218)	p (2929)	b (13)	b (885)
dh (2376)	th (14)	th (911)	q (2685)	t (14)	hv (230)
dx (1864)	d (18)	d (661)	r (4681)	axr (28)	axr (2850)
eh (3277)	ix,ih (12)	ae (1520)	s (6176)	z (41)	z (5159)
el (951)	l (18)	l (809)	sh (1317)	s (7)	zh (528)
em (124)	m (3)	m (72)	t (6367)	d (47)	d (2720)
en (630)	n (18)	n (486)	th (745)	dh (6)	dh (206)
eng (26)	()	ng (17)	uh (500)	ix (8)	ax (104)
epi (908)	t,f (4)	t (85)	uw (529)	ux,ax (4)	ux (133)
er (1693)	axr (7)	axr (1074)	ux (1423)	ix (13)	ix (430)
ey (2271)	iy (4)	iy (903)	v (1994)	f (22)	f (589)
f (2215)	p (7)	v (709)	w (2216)	r (3)	l (949)
g (1458)	k (13)	k (779)	y (995)	iy (3)	iy (323)
hh (937)	hv (8)	hv (449)	z (3682)	s (52)	s (3010)
hv (723)	hh (7)	hh (491)	zh (149)	sh,dx (1)	sh (77)
ih (4248)	ix (84)	ix (2593)			

6. Conclusions

This paper has shown that a novel approach using *restricted recognition* to enhance phone confusion matrices addresses the data sparseness problem of standard confusion matrices and provides a data-driven resource which captures degrees of similarity between phones. As a resource, this enhanced confusion matrix was shown to perform better on a spoken term detection task using different corpora and speech recognition methods.

7. References

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[3] H. Strik and C. Cucchiarini, "Modeling pronunciation variation for asr: A survey of the literature," *Speech Communication*, vol. 29, no. 2, pp. 225–246, 1999.

[4] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris *et al.*, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10, pp. 763–786, 2007.

[5] S. Greenberg, "Speaking in shorthand–a syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, vol. 29, no. 2, pp. 159–176, 1999.

[6] M. Akbacak, L. Burget, W. Wang, and J. Van Hout, "Rich system combination for keyword spotting in noisy and acoustically heterogeneous audio streams," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 8267–8271, 2013.

[7] G. Chen, C. Parada, and T. N. Sainath, "Query-by-example keyword spotting using long short-term memory networks," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5236–5240, 2015.

[8] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2, pp. 95–108, 2000.

[9] A. Alvarez, H. Arzelus, and P. Ruiz, "Long audio alignment for automatic subtitling using different phone-relatedness measures," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6280–6284, 2014.

[10] N. Moreau, H.-G. Kim, and T. Sikora, "Phonetic confusion based document expansion for spoken document retrieval," *INTERSPEECH*, 2004.

[11] P. Zhang, J. Shao, J. Han, Z. Liu, and Y. Yan, "Keyword spotting based on phoneme confusion matrix," *Proc. of ISCSLP*, vol. 2, pp. 408–419, 2006.

[12] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, pp. 421–426, 2009.

[13] M. Larson and G. J. Jones, "Spoken content retrieval: A survey of techniques and technologies," *Foundations and Trends in Information Retrieval*, vol. 5, no. 45, pp. 235–422, 2012.

[14] M. Saraclar, A. Sethy, B. Ramabhadran, L. Mangu, J. Cui, X. Cui, B. Kingsbury, and J. Mamou, "An empirical study of confusion modeling in keyword search for low resource languages," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 464–469, 2013.

[15] A. Mandal, K. P. Kumar, and P. Mitra, "Recent developments in spoken term detection: a survey," *International Journal of Speech Technology*, vol. 17, no. 2, pp. 183–198, 2014.

[16] S. Srinivasan and D. Petkovic, "Phonetic confusion matrix based spoken document retrieval," *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 81–87, 2000.

[17] R. G. Wallace, R. J. Vogt, and S. Sridharan, "A phonetic search approach to the 2006 NIST spoken term detection evaluation," 2007.

[18] G. Chen, O. Yilmaz, J. Trmal, D. Povey, and S. Khudanpur, "Using proxies for OOV keywords in the keyword search task," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 416–421, 2013.

[19] D. Xu, Y. Wang, and F. Metze, "Em-based phoneme confusion matrix generation for low-resource spoken term detection," *IEEE Spoken Language Technology Workshop (SLT)*, pp. 424–429, 2014.

[20] A. Žgank, B. Horvat, and Z. Kačič, "Data-driven generation of phonetic broad classes, based on phoneme confusion matrix similarity," *Speech Communication*, vol. 47, no. 3, pp. 379–393, 2005.

[21] P. Scanlon, D. P. Ellis, and R. B. Reilly, "Using broad phonetic group experts for improved speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 803–812, 2007.

[22] C. Lopes and F. Perdigão, "Broad phonetic class definition driven by phone confusions," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, pp. 1–12, 2012.

[23] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The HTK book," *Cambridge University Engineering Department*, vol. 3.4.1, 2009.

[24] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, 1993.

[25] D. Ellis, "HTK training for TIMIT from cantab research," *http://www.cantabResearch.com/HTKtimit.html*, vol. v1.3, 2006.

[26] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 109–112, 1990.

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec. 2011.

[28] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," *EUROSPEECH*, p. 18951898, 1997.

[29] NIST, "DETware v2.1: DET-curve plotting software for use with MATLAB," *http://www.nist.gov/itl/iad/mig/tools.cfm*.

[30] International-Phonetic-Association, *The principles of the International Phonetic Association: being a description of the International phonetic alphabet and the manner of using it.* International Phonetic Association, 2005.