---

**Proceedings Paper:**
Rath, S.P., Knill, K.M., Ragni, A. orcid.org/0000-0003-0634-4456 et al. (1 more author) (2014) Combining tandem and hybrid systems for improved speech recognition and keyword spotting on low resource languages. In: INTERSPEECH 2014 : 15th Annual Conference of the International Speech Communication Association. INTERSPEECH 2014 : 15th Annual Conference of the International Speech Communication Association, 14-18 Sep 2014, Singapore. International Speech Communication Association (ISCA) , pp. 835-839.

---

# Combining Tandem and Hybrid Systems for Improved Speech Recognition and Keyword Spotting on Low Resource Languages

*Shakti P. Rath, Kate M. Knill, Anton Ragni and Mark J. F. Gales*

Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, UK
{shakti.rath, kate.knill, ar527, mjfg}@eng.cam.ac.uk

## Abstract

In recent years there has been significant interest in Automatic Speech Recognition (ASR) and Key Word Spotting (KWS) systems for low resource languages. One of the driving forces for this research direction is the IARPA Babel project. This paper examines the performance gains that can be obtained by combining two forms of deep neural network ASR systems, Tandem and Hybrid, for both ASR and KWS using data released under the Babel project. Baseline systems are described for the five option period 1 languages: Assamese; Bengali; Haitian Creole; Lao; and Zulu. All the ASR systems share common attributes, for example deep neural network configurations, and decision trees based on rich phonetic questions and state-position root nodes. The baseline ASR and KWS performance of Hybrid and Tandem systems are compared for both the "full", approximately 80 hours of training data, and limited, approximately 10 hours of training data, language packs. By combining the two systems together consistent performance gains can be obtained for KWS in all configurations.

**Index Terms**: keyword spotting, deep neural network, Tandem, Hybrid

## 1. Introduction

In recent years there has been an increasing interest in Automatic Speech Recognition (ASR) and Key Word Spotting (KWS) for low resource languages. The task of KWS is to find occurrences of a particular word or a phrase (a.k.a. query) in audio recordings. The state-of-the-art KWS systems are based on the word lattices generated by an ASR system for the query search. One of the driving forces for this research direction is the IARPA Babel project [1]. The objective of the project is to develop robust KWS (the primary task) and underlying ASR technologies for any human language utilising *limited* amount of data for the ASR engine training.

In this paper two Deep Neural Network (DNN) based ASR systems [2, 3] are investigated - Tandem [4, 5] and Hybrid [6] for both ASR and KWS. In the Tandem configuration the DNN operates as a feature extractor that provides input to the back-end HMM-GMM classifier. In contrast, in the Hybrid configuration it plays the role of the acoustic model [2, 6] itself. A stacked version of the Hybrid system is investigated in this work. Here the features extracted from the Tandem system are used as the input to the Hybrid system. This presents an interesting contrast as the features for both the GMM-HMM system and the Hybrid system are the same, it is only the form of classifier to obtain the observation likelihoods that differs.

The combination of these Tandem and stacked Hybrid systems for both ASR and KWS systems is also considered. This is based on the assertion that the two systems have diverse operational mechanism and hence they likely to have complementary advantages. The combination of ASR systems is well established, using approaches such as ROVER combination [7] and Confusion Network Combination (CNC) [8]. This concept has also been applied to KWS. Some of the early work that combine systems to improve the KWS performance are [9, 10, 11], which combine KWS systems that use word and sub-word models. The recent works in this direction are [12, 13, 14, 15, 9]. [13] combines results from ASR systems with diverse components, such as acoustic model, decoding strategy and audio segmentation, to improve the KWS performance.

This paper examines the combination of Deep Neural Network based ASR and KWS systems in a consistent frame-work for five languages: Assamese, Bengali, Haitian Creole, Lao, and Zulu. Performance is contrasted with each of the individual systems, as well as examining the correlation between the ASR performance and KWS performance. The next section discusses the nature of Babel KWS task and the data. This is followed by a brief description of the ASR and KWS systems used. Finally the experimental results are presented.

## 2. Task Description

The work reported in this paper is based on the IARPA Babel [1] project, which aims to foster research on speech recognition and keyword spotting for low resource languages. The Babel speech corpus covers a range of diverse languages and is distributed under two configurations for each language: a "full" language pack (FLP) comprising of approximately 80 hours of transcribed audio; and the "limited" language pack (LLP) comprising of about 10 hours of transcribed audio. The data is primarily conversational telephone speech, recorded over a range of acoustic conditions, such as mobile phone conversation made from car. The FLP and LLP share the same development set of 10 hours of conversational speech. The phone set and phonetic lexicon are supplied for every language and the lexicon contains only those words occurring in the training data.

This work was supported in part 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. The authors would like to thank the Lorelei team for providing the KWS infrastructure and morphological decomposition.

In the Option Period 1 (OP1) phase of the project, audio from five languages have been released: Assamese; Bengali; Haitian Creole; Lao; and Zulu. The ASR and KWS experiments reported in this paper are conducted on the OP1 languages (both FLP and LLP), and the performance is evaluated on the development data defined for the evaluation 2014. The official metric to measure the accuracy of the system performance has been defined to be Maximum Term Weighted Value (MTWV), which is the best term weighted value [16] (TWV) that can be achieved over all choices of detection threshold. The TWV is defined as

$$TWV(\theta) = 1 - [P_{miss}(\theta) + \beta P_{fa}(\theta)] \quad (1)$$

where $P_{miss}(\theta)$ and $P_{fa}(\theta)$ denote the probabilities of miss and false alarm, respectively, $\theta$ is the detection threshold, and $\beta$ decides the relative weight given to each type of errors.

| Language | Release |
|---|---|
| Assamese | IARPA-babel102b-v0.5a |
| Bengali | IARPA-babel103b-v0.4b |
| Haitian Creole | IARPA-babel201b-v0.2b |
| Lao | IARPA-babel203b-v3.1a |
| Zulu | IARPA-babel206b-v0.1e |

This work made use of the IARPA Babel Program language collection releases shown above.

## 3. ASR System Description

The core ASR toolkit, used for parameterisation, clustering, decoding and GMM-based acoustic model training, is an extended version of the HTK-3.4.1 [17] toolkit. The multi-layer perceptron (MLP) training used an extended version of ICSI's QuickNet [18], which allows deeper network configurations to be used, to train both Tandem and Hybrid systems.

The ASR acoustic models for both Tandem and Hybrid systems shared the same underlying attributes for all language packs. The underlying context-dependent states were specified using state [19, 20, 21], rather than phone-state, roots of the decision tree. Questions involving X-SAMPA attribute and position of the phone in the word were then used. This was found to provide additional robustness to the rare phones, for example the X-SAMPA phone /kx/ in Zulu. With no phone mappings and phone-state roots these would be modelled as monophones. To further improve the ability to model rare phones, diphthongs were split into their constituent parts, with additional markers added to indicate that the unit was derived from a diphthong.

In addition all systems were based on deep neural networks. Two configurations were used. The Tandem configuration used a single neural network with PLP and pitch features as the input. The output of this network was then used in a Hybrid system yielding a stacked configuration. This is illustrated in Figure 1. Both the Tandem and stacked Hybrid MLPs were initialised using layer-by-layer discriminative pre-training. Further details of the two acoustic models are given below.

The language models (LM) for all systems were built using the vocabulary and training data from the audio transcriptions. For all systems trigram class-based language models, interpolated with the word-based language models, were used.

### 3.1. Tandem System

The development of Tandem system is based on [22]. An MLP was trained using cross-entropy, and context dependent targets

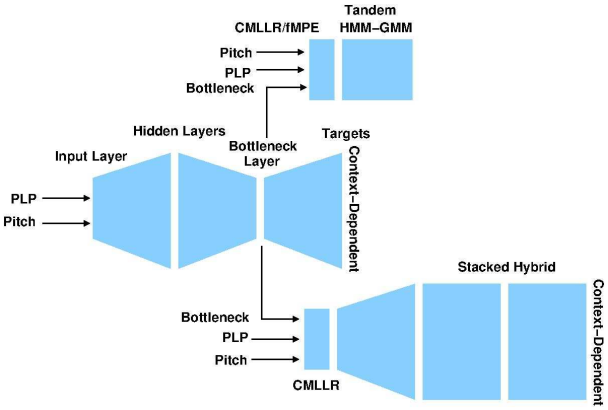


Figure 1: Tandem and Stacked Hybrid systems

were defined by a phonetic decision tree. The input to the network was 9 frames of PLP with pitch [1] appended, and delta, delta-deltas and triples added. This yields a total input vector size of 504. All systems had a bottleneck layer of 26 nodes.

The 26 dimensional bottleneck features were then transformed using a global semi-tied covariance matrix [23] and then appended to HLDA projected PLP features (39 dimensions) and pitch with delta and delta-delta parameters. This yields a complete feature of 68 dimensions. These are the baseline features for the Tandem system below.

A speaker adaptive training (SAT) system using global constrained maximum likelihood linear regression (CMLLR) at a speaker level [24] was then constructed, followed by both Minimum Phone Error (MPE) [25] and feature-space MPE (fMPE) [26] training. The CMLLR transforms were estimated using maximum likelihood (ML) on the ML estimated acoustic models. These were then fixed, and MPE and fMPE training were applied using the CMLLR normalised features.

A multi-pass decoding and adaptation process was used:

1. speaker-independent (SI) decoding with a PLP-based MPE system [2];
2. a global CMLLR transform was estimated for each speaker using the Tandem ML-SAT model;
3. global CMLLR and MLLR transforms were estimated using the Tandem-SAT fMPE+MPE acoustic model;
4. speaker adapted decoding using the Tandem-SAT fMPE+MPE system and a bigram word-based LM;
5. lattice rescoring with a class-based language model and confusion network (CN) generation.

The configurations of the two language packs were tuned to the quantities of data available. The details are as follows:

**Full Language Pack**: the number of target states was set at about 6000 for both the MLP and HMM systems. Five hidden layers, including the bottleneck layer were used, and the network configuration was (including input and target layers):

[1] Initial experiments showed that using pitch as an input to the MLP significantly improved the performance of tonal languages such as Lao, with smaller improvements for non-tonal languages.

[2] Though the performance was worse than a Tandem SI system, the final adapted performance was better because of cross-system effects.

$504$x$1000^4$x$26$x$6000$.

**Limited Language Pack**: the number of target states was set at about 1000 for both the MLP and HMM systems. Four hidden layers, including the bottleneck layer were used. The network configuration including input and target layers was: $504$x$1000$x$500^2$x$26$x$1000$. To moderate the impact of the limited training data either Vocal Tract Length Perturbation (VTLP) [27, 28] was used (Bengali, Haitian Creole, and Lao) or semi-supervised approaches [29, 30, 31] (Assamese and Zulu) were used to train the MLP. For VTLP 8 warp factors for each of the training speakers were used, effectively increasing the size of the training corpus to be the same as the FLP. For the unsupervised data, all the untranscribed audio was recognised and confidence based thresholds used to select 50% of the data.

### 3.2. Stacked Hybrid System

As shown in Figure 1, the hybrid system was trained in a stacked fashion. First the bottleneck MLP for this Tandem system was constructed. These features were speaker normalised using the ML estimated CMLLR transforms from the ML Tandem-SAT system. Again 9 vectors, each of 68-dimensions, were then stacked together to yield a total input vector of 612 features. Speaker adapted decoding with the Hybrid system used the CMLLR transforms generated at stage (2) of the Tandem decoding process to speaker normalise the features. A bigram language model was used for the Hybrid decoding, which was followed by the lattice rescoring using the class-based language model and CN generation as in step (5) of the Tandem decoding.

The configurations of the two language packs were tuned to the quantities of data available. The details are as follows:

**Full Language Pack**: the number of target states was set at about 6000 for the MLP. Five hidden layers were used, and the network configuration including input and target layers was: $612$x$1000^5$x$6000$.

**Limited Language Pack**: the default number of states was set at about 1000 for the MLP. Four hidden layers were used, and the default network configuration including input and target layers was: $612$x$1000$x$500^3$x$1000$. Due to time constraints, data augmentation was only applied to the Zulu Hybrid system. Here semi-supervised training [29, 30, 31] was used in the same fashion as it was in the case of the Tandem system, and the number of target states was increased to 3000.

### 3.3. System Combination

Given the different forms of classifier being used for the Tandem and Hybrid systems, they may be expected to be complementary to one another. To investigate this, the confusion networks generated by the Tandem and Hybrid systems are combined to produce a merged CN [8]. Before combining the two systems, the posterior probability associated with the CN of each system, based on the arc posteriors from the lattice, were mapped to remove any biases in the confidence measures.

## 4. KWS System Description

The KWS system is based on the weighted finite state transducer (WFST) [12]. First the ASR system is used to generate word lattices. These lattices are then processed to generate the word indices for the in-vocabulary (IV) search and phonetic indices to accommodate out-of-vocabulary (OOV) search. The timing information is pushed to the output labels of the arcs of the resulting WFSTs. The arcs in the WFST after the push operation can be expressed as a 5-tuple $(p, i, o, w, q)$, where $p$ and $q$ indicate the start and end states, $i$ denotes the input label, which is a word in case of IV search or a phoneme in case of OOV search, $w$ indicates the posterior probability associated with the input label, and finally $o$ denotes the output label.

The IV queries are searched in the word index, whereas the OOV queries are searched in the phonetic index. More specifically, for the IV search, each query is converted to a word weighted finite state acceptor (WFSA) and a composition operation is carried out with the word index in order to retrieve the hit list (a.k.a posting list) for the query. Each hit list is identified by the name of the audio file, the starting time of the query, duration and the score, which is the posterior probability derived from the WFST. On the other hand, for the OOV search, each query is first expanded to a reasonable phonetic representation using a grapheme-to-phoneme converter, which may not give accurate pronunciation for all queries. The resulting pronunciation is then represented as a phonetic WFSA, and a composition with the phonetic WFST is carried out to retrieve the hit lists for the OOV terms. To boost the OOV search performance a large number of query expansion using a phone-to-phone (P2P) confusion model (NBestP2P) [13] was incorporated. The values of the NBestP2P were set in the range of 1000 to 50000. In addition, it was observed that the OOV performance can be further improved by zeroing the language model score. The IV queries that did not return hits were searched again in the phoneme index which is known as the *cascaded search*. Finally, the IV, OOV and cascaded search hit lists are combined and sum-to-one (STO) [12] score normalisation is applied to make sure that sum of all normalised detection scores for each query is 1.0.

For some languages that are morphologically rich the number of OOV terms can become very large, adversely impacting the performance. For example for the Zulu LLP 61% of the query terms were OOV, compared to 31% for the Bengali LLP. To address this problem a morphological KWS was used for Zulu [32]. Here initially IV word terms are found. Then IV morph terms are found, finally OOV morph terms are found.

| KWS Process | MTWV | | |
|---|---|---|---|
| | IV | OOV | Total |
| Word | 0.2649 | 0.1338 | 0.1851 |
| Morphological | 0.2615 | 0.2073 | 0.2287 |

Table 1: *MTWV scores comparing morphological and word KWS systems for Zulu LLP.*

The impact of the morphological search on the Zulu LLP is shown in Table 1. It is observed that the overall (cascaded search) MTWV score increases, primarily due to the improvement in the OOV search (this is the morph-level OOV search). The slight degradation in IV word performance is due to a shift in the MTWV operating point. In initial experiments, morphological search yielded a performance gain only for Zulu system.

In this work, a simple merging of the posting lists from the Tandem and Hybrid systems, prior to STO normalisation, was used in order to combine the two KWS systems together, rather than a more complicated approach such as MTWV-weighted CombMNZ method discussed in [12]. In initial experiments, there was a slight degradation in performance by using this merging, rather than CombMNZ, but it simplified the pipeline.

## 5. Experimental Results

For all experiments the development data audio and keyword spotting list are associated with the sets shown in Section 2. For each language the development set has approximately 10 hours of audio, and 2000 terms to search for. In this paper Token Error Rate (TER), rather than WER, is used when discussing ASR results. For the broad range of languages investigated under the Babel programme, some languages, for example Vietnamese, do not have references at the word level. Thus TER removes the concept of word (though measured in the same fashion).

It is worth emphasising that given the objective of the project to be a KWS performance of greater than 0.3, wherever choices of system configuration have been made, they were based on the KWS performance, not on the ASR performance.

| Language | Id | LP | TER (%) | | |
|---|---|---|---|---|---|
| | | | Tandem | Hybrid | CNC |
| Assamese | 102 | FLP | 54.2 | 55.1 | 52.8 |
| | | LLP | 65.1 | 67.8 | 64.3 |
| Bengali | 103 | FLP | 54.9 | 56.6 | 54.3 |
| | | LLP | 67.0 | 69.5 | 66.8 |
| Haitian Creole | 201 | FLP | 48.7 | 50.3 | 48.2 |
| | | LLP | 60.5 | 63.4 | 60.4 |
| Lao | 203 | FLP | 48.5 | 51.9 | 48.9 |
| | | LLP | 61.2 | 65.8 | 61.3 |
| Zulu | 206 | FLP | 62.1 | 64.4 | 61.2 |
| | | LLP | 71.5 | 74.1 | 70.6 |

Table 2: *%TER with CN decoding for Tandem and Hybrid and CNC for Full (FLP) and Limited (LLP) Language Packs.*

Table 2 shows the ASR system performance on each of the languages, and each configuration. There are some general trends. For these DNN systems, the Tandem system consistently outperforms the Hybrid configuration. Part of this difference in performance may be because of the use of cross-entropy for the training, rather than sequence training. The difference in performance is also greater for the LLP than the FLP. This can partly be attributed to the use of data augmentation for the Tandem system, but not for the Hybrid system. In general the combination of the Tandem and Hybrid ASR systems helped to improve the performance. The outlier for this was Lao, where the difference in the performance between the Tandem and Hybrid systems was the greatest among all languages.

| Language | Id | LP | MTWV | | |
|---|---|---|---|---|---|
| | | | Tandem | Hybrid | Merge |
| Assamese | 102 | FLP | 0.4660 | 0.4730 | 0.4946 |
| | | LLP | 0.2569 | 0.2360 | 0.2771 |
| Bengali | 103 | FLP | 0.5151 | 0.5121 | 0.5388 |
| | | LLP | 0.2992 | 0.2615 | 0.3100 |
| Haitian Creole | 201 | FLP | 0.6387 | 0.6329 | 0.6602 |
| | | LLP | 0.4648 | 0.4336 | 0.4867 |
| Lao | 203 | FLP | 0.5951 | 0.5881 | 0.6149 |
| | | LLP | 0.4262 | 0.3790 | 0.4439 |
| Zulu | 206 | FLP | 0.3770 | 0.3654 | 0.4084 |
| | | LLP | 0.2287 | 0.1924 | 0.2366 |

Table 3: *MTWV for Tandem and Hybrid and their combination for Full (FLP) and Limited (LLP) Language Packs.*

Table 3 shows the performance of the KWS system on each of the languages and language packs. For the FLPs the performance of both the Tandem and Hybrid systems is very similar, for Assamese the Hybrid system yielded the best performance. For the LLPs there is still a gap in performance, with the Tandem system outperforming the Hybrid. This was also true when comparing the Tandem with no data augmentation to the Hybrid system. In contrast to the ASR system combination, merging posting lists improved the KWS performance in every configuration, even for Lao LLP where the difference in the KWS performance is large. For the combined system, 8 out of the 10 configurations achieved the program goal of 0.3 TWV, although there was a slight degradation when the threshold is automatically determined rather than using the MTWV.

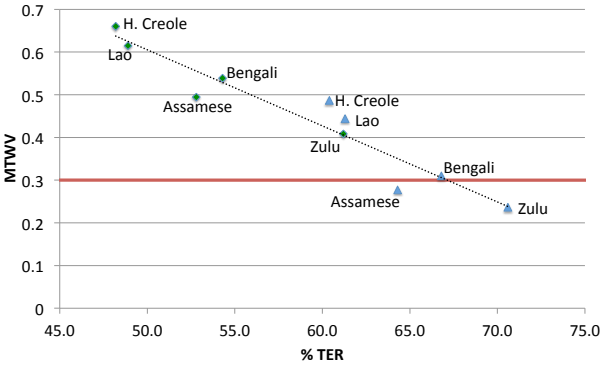


Figure 2: MTWV against TER, ◊ indicates FLP, △ LLP

Given that the 10 configurations have been run in a consistent framework, it is interesting to examine the correlation between the ASR performance and the KWS performance. Figure 2 shows the plot of MTWV against TER (%) for all five option period 1 languages for both LLP and FLP configurations. Here the CNC TER% and the Merged MTWV scores are given. From the plot it is observed that the correlation between the two is high (Pearson Correlation Coefficient -0.945, $R^2$ value 0.911). It is also clear that some languages, such as Haitian Creole and Lao, are simpler than the others at least for this task. Further, the KWS performance of Assamese on the development Keyword List is lower for both FLP and LLP than what is expected considering its ASR performance.

## 6. Conclusions

In this paper the use and combination of deep neural network based Tandem and Hybrid systems were investigated for both ASR and KWS on low resource languages. The systems were evaluated on five languages from the Babel Program using both the Full (FLP, about 80 hours) and Limited (LLP, about 10 hours) Language Pack configurations. The baseline Hybrid systems yielded comparable performance for KWS, the primary task, as the Tandem systems for the FLP configuration. For the LLP the Hybrid performance was poorer. However in combination, the two different forms of classifier yielded complementary KWS systems and gains in MTWV were observed for all languages and both FLP and LLP configurations. Similar trends were observed for the secondary ASR task. However for Lao, where there was the greatest difference in ASR performance between the Tandem and Hybrid systems, slight degradations in the combined performance were observed.

## 7. References


[1] Mary Harper, "IARPA Babel Program," http://www.iarpa.gov/Programs/ia/Babel/babel.html.

[2] G. Hinton, L. Deng, et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[3] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. of ASRU*, Dec 2011.

[4] H. Hermansky, D. Ellis, and S. Sharma, "Tandem Connectionist Feature Extraction for Conventional HMM Systems," in *Proc. of ICASSP*, 2000.

[5] Frantisek Grezl, Martin Karafiat, and Milos Janda, "Study of probabilistic and bottle-neck features in multilingual environment," in *Proc. of ASRU*, 2011.

[6] H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*, Kluwer Academic Publishers, Norwell, MA, USA, 1993.

[7] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recogniser Output Voting Error Reduction (ROVER)," in *Proc. of ASRU*, 1997.

[8] G. Evermann and P. C. Woodland, "Large vocabulary decoding and confidence estimation using word posterior probabilities," in *Proc. of ICASSP 2000*.

[9] D. R. H. Miller, M. Kleber, et al., "Rapid and accurate spoken term detection," in *Proc. of Interspeech*, 2007.

[10] D. Vergyri, I. Shafran, et al., "The SRI/OGI 2006 spoken term detection system," in *Proc. of Interspeech*, 2007.

[11] I. Szoke, L. Burget, J Cernocky, and M. Fapso, "Sub-word modeling of out of vocabulary words in spoken term detection," in *Proc. of SLT 2008*.

[12] J. Mamou et al., "System combination and score normalization for spoken term detection," in *Proc. of ICASSP*, 2013.

[13] L. Mangu, H. Soltau, H.-K. Kuo, B. Kingsbury, and G. Saon, "Exploiting diversity for spoken term detection," in *Proc. of ICASSP*, 2013.

[14] B. Kingsbury et al., "A high-performance Cantonese keyword search system," in *Proc. of ICASSP*, 2013.

[15] D. Karakos, R Schwartz, S. Tsakalidis, L. Zhang, et al., "Score normalization and system combination for improved keyword spotting," in *Proc. of ASRU 2013*.

[16] J. G. Fiscus et al., "Results of the 2006 Spoken Term Detection Evaluation," in *Proc. ACM SIGIR Workshop on Searching Spontaneous Conversational Speech*, 2007.

[17] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book (for HTK version 3.4.1)*, Cambridge University, `http://htk.eng.cam.ac.uk`, 2009.

[18] David Johnson et al., "QuickNet," http://www1.icsi.berkeley.edu/Speech/qn.html.

[19] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. of Eurospeech*, 1999.

[20] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings ARPA Workshop on Human Language Technology*, 1994, pp. 307–312.

[21] Borislava Mimer, Sebastian Stüker, and Tanja Schultz, "Flexible deciscion trees for grapheme based speech recognition," in *Proc. 15th Conference Elektronische Sprachsignalverabeitung (ESSV)*, Cottbus, Germany, 2004.

[22] J. Park et al., "The Efficient Incorporation of MLP Features into Automatic Speech Recognition Systems," *Computer Speech and Language*, vol. 25, pp. 519–534, 2010.

[23] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transaction of Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, May 1999.

[24] M. J. F. Gales, "Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[25] D. Povey and P. C. Woodland, "Minimum Phone Error and I-smoothing for improved discriminative training," in *Proc. of ICASSP*, 2002.

[26] D. Povey et al., "fMPE: Discriminatively trained features for speech recognition," in *Proc. of ICASSP*, 2005.

[27] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc of ICML*, 2013.

[28] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," in *Proc. of ICASSP*, 2014.

[29] G. Zavaliagkos and T. Colthurst, "Utilizing untranscribed training data to improve performance," in *Proc. of DARPA Broadcast news transcription and understanding workshop*, 1998.

[30] L. Lamel and J.-L. Gauvain, "Lightly supervised and unsupervised acoustic model training," *Computer speech and language*, vol. 16, pp. 115–129, 2013.

[31] M. J. F. Gales, D. Y. Kim, P. C. Woodland, H. Y. Chan, D. Mrva, R. Sinha, and S. E. Tranter, "Progress in the CU-HTK broadcast news transcription system," *IEEE Tran ASLP*, vol. 14, no. 5, pp. 1513–1525, 2006.

[32] M. S. Rasooli, N. Habash, O. Rambow, and T. Lippincott, "Unsupervised morphology-based vocabulary expansion," in *The 52nd Annual Meeting of the Association for Computational Linguistics*, 2014.