Data augmentation for low resource languages

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

Department of Engineering, University of Cambridge
Trumpington Street, Cambridge CB2 1PZ, UK
{ar527,kmk1001,spr38,mjfg}@eng.cam.ac.uk

Abstract

Recently there has been interest in the approaches for training speech recognition systems for languages with limited resources. Under the IARPA Babel program such resources have been provided for a range of languages to support this research area. This paper examines a particular form of approach, data augmentation, that can be applied to these situations. Data augmentation schemes aim to increase the quantity of data available to train the system, for example semi-supervised training, multi-lingual processing, acoustic data perturbation and speech synthesis. To date the majority of work has considered individual data augmentation schemes, with few consistent performance contrasts or examination of whether the schemes are complementary. In this work two data augmentation schemes, semi-supervised training and vocal tract length perturbation, are examined and combined on the Babel limited language pack configuration. Here only about 10 hours of transcribed acoustic data are available. Two languages are examined, Assamese and Zulu, which were found to be the most challenging of the Babel languages released for the 2014 Evaluation. For both languages consistent speech recognition performance gains can be obtained using these augmentation schemes. Furthermore the impact of these performance gains on a down-stream keyword spotting task are also described.

Index Terms: data augmentation, speech recognition, babel

1. Introduction

A large amount of transcribed training data is usually needed to enable accurate speech recognition [1, 2]. Although for some languages, such as English and Mandarin, these resources may be sourced, for others, termed *low resource languages*, it may not always be feasible. This has recently created lots of interest in the approaches that can be applied to these situations [3, 4, 5]. To facilitate research in this direction, consistent packs of limited resources for a range of languages have been provided under the IARPA Babel program. The goal of the program is to provide effective search capabilities to efficiently process real-world recorded speech. This is effectively a spoken term detection task, where speech recognition systems are assessed based on keyword search (KWS) performance rather than more conventional transcription accuracy. Though improvements in speech recognition performance may not necessarily translate into improved KWS capacity [3], a certain positive correlation does exist [3, 4, 5], which motivates the work on building accurate speech recognition systems.

A common issue that arises from the use of limited resources in speech recognition systems is robust parameter estimation. A range of approaches can be applied to address robustness issues. These include standard statistical approaches, such as maximum a posteriori (MAP) estimation [6], and *data augmentation* [7, 8, 9, 10, 11, 12]. The MAP estimation introduces a prior on model parameters into training objective function [6, 13]. However, this approach is ill-suited in situations when there is no training data or informative prior distribution available. The data augmentation aims to increase the quantity of training data. This approach has an important theoretical advantage of being able to produce data when real examples are not available [8]. Common schemes include semi-supervised training [7, 2, 11], multi-lingual processing [3, 11, 14], acoustic data perturbation [9, 10, 12] and speech synthesis [8].

The previous work with data augmentation has mostly focused on individual schemes. Not much work has been done on contrasting and examining whether the schemes are complimentary. This paper examines and combines semi-supervised training and acoustic data perturbation on two languages, Assamese and Zulu, found to be the most challenging of the Babel languages released for the 2014 Evaluation.

The rest of this paper is organised as follows. Section 2 provides an overview of commonly used data augmentation schemes including semi-supervised training and acoustic data perturbation. Section 3 discusses options available for training speech recognition systems on augmented data. Section 4 provides individual and combined results on using semi-supervised training and acoustic data perturbation for the two Babel program languages, Assamese and Zulu. Finally, Section 5 presents conclusions drawn from this work.

2. Data augmentation

The data augmentation refers to the schemes that aim to increase the quantity of data available to train speech recognition systems. The schemes can be split based on the type of produced data, such as unsupervised, synthesised and other language data.

2.1. Unsupervised data

The unsupervised data refers to data which lacks correct transcriptions. This also includes data having only rough transcriptions, such as closed captions [7, 2]. The unsupervised data may be adopted by recognising it with an existing or boot-strapped system, filtering out those utterances that fail to decode/pass confidence threshold [15, 16] and re-training the system on su-

This work was 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.

pervised and filtered unsupervised training data. This is commonly referred to as a semi-supervised training. The main advantage of unsupervised data is that it is generally possible to collect vast amounts of such data, e.g., radio and television news broadcasts [7], covering all sorts of speaker and noise conditions. The main disadvantage of this type of data is the lack of correct transcriptions. This limits possible gains from the approaches particularly sensitive to the accuracy of supplied transcriptions, such as discriminative training [17] and speaker adaptation based on discriminative criteria [18].

2.2. Synthesised data

The synthesised data may refer to existing but perturbed in a certain way data as well as new artificially generated data. One major advantage of synthesised data is that, similar to unsupervised case, it is possible to collect vast amounts of such data. Furthermore, different to unsupervised case, the correctness of associated transcriptions is usually guaranteed. A major disadvantage of this type of data could be its quality.

There are numerous options how data can be perturbed. These include vocal tract length perturbation (VTLP) [9, 10] and stochastic feature mapping (SFM) [12]. The VTLP scheme attempts to alter vocal tract length during extraction of standard speech parametrisations such as Mel-frequency cepstral (MFCC) and perceptual linear prediction (PLP) coefficients. Essentially, a single warping parameter is modified either stochastically [9] or deterministically [12]. This results in a simple synthesis process yet the data is perturbed in a non-linear way. Though the original motivation for VTLP was to learn multi-layer perceptrons (MLP) robust to changes in vocal tract length [9, 10, 12], the scheme could be of a wider interest, for instance, to boot-strap systems used for recognising unsupervised data. In contrast to VTLP, the SFM is a general methodology for stochastically mapping features from one domain to another [19, 12]. When applied to speakers, the SFM essentially yields a simplified voice morphing [20] scheme. One simple approach to map utterances of one speaker to another is to apply global constrained maximum likelihood regression (CMLLR) transform [21] estimated from statistics of the other speaker [12]. The issue with this approach is that a simple global transform is applied to every observation in the sequence which may not be powerful enough to yield accurate mapping.

Rather than perturbing existing data it is possible [8] to artificially generate new examples using speech synthesis approaches, such as concatenative or statistical [22]. The concatenative approach attempts to synthesise speech by concatenating existing waveform segments into a sequence. The statistical approach usually adopts acoustic models, such as hidden Markov models (HMM), to produce speech parameter sequences maximising likelihood [22]. These model-based schemes may be particularly useful since speech parameter sequences, such as MFCC or PLP, rather than waveforms are required. Thus many of the waveform production issues [23] are not relevant. Furthermore, these schemes permit model-based adaptation/compensation approaches [24] to be used for synthesising data with target [25] and new [26] speaker and environment characteristics. In contrast to acoustic data perturbation schemes, the use of speech synthesis offers flexibility in generating data for arbitrary given transcription. For instance, it is possible to generate data for targeting only particular confusions using schemes such as acoustic code-breaking [27].

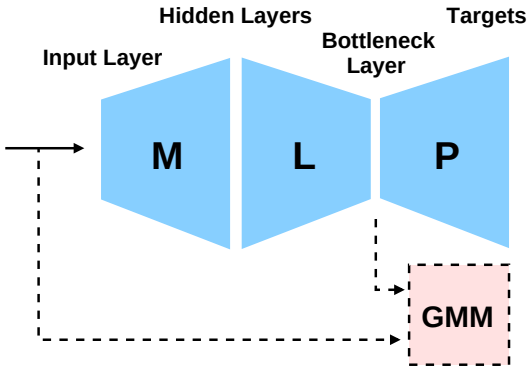

Figure 1: *Schematic diagram of a tandem approach*

2.3. Other language data

Though for many languages there are only limited or no resources, for some languages sufficient resources are available. This has prompted lots of interest in using this type of augmenting data [3, 11, 14]. Furthermore, the use of unsupervised other language data has also been considered [11]. Compared to synthesised data, this type of augmenting data, similar to unsupervised data, is real. However, its use may also be more complicated as it is not obvious what is the best way to exploit it [3]. There have been proposed several approaches. One group relies on the use of a universal phone set to accomplish mapping of one language to another [28, 3]. Another group relies on an alternative form of mapping, such as phone-to-phone [29] or hidden layer unit-to-target as in the MLP-based work of [30, 14]. Both directions have their own advantages and disadvantages [3]. In particular, it is not obvious how to ensure sufficient coverage in the training data for approaches based on universal phone sets, map phones between languages such as English and Cantonese in approaches based on phone-to-phone mappings or ensure that targets are optimally ordered in approaches that map hidden layer units to targets.

3. Augmentation modes

Given augmenting data, an important question is how to best exploit it. The answer to this question will ultimately depend on the particular architecture the speech recogniser adopts and the nature and amount of augmenting data used. There are numerous configurations possible, such as standard Gaussian mixture model (GMM) based HMM [31], tandem [32], hybrid [33], stacked versions of the tandem [34] and hybrid [33] architectures. This section considers several of these, putting a particular emphasis on the tandem architecture adopted in Section 4.

The tandem architecture may be illustrated by Figure 1 which shows a MLP and GMM-based HMM speech recogniser. Three types of MLP layers are shown: input, hidden and bottleneck. The standard parametrisation, such as MFCC or PLP, optionally de-correlated and transformed, is fed into the input layer which is followed by hidden layers where it undergoes a series of non-linear transformations until it reaches bottleneck layer where MLP-derived features are extracted. These features, also called bottleneck (BN) features, are then concatenated with the standard parametrisation, optionally de-correlated and transformed, and used within the standard GMM-based HMM speech recogniser. The hybrid architecture may also be illustrated by Figure 1, although in this case the dashed part is not present and the bottleneck layer is replaced with an extra hidden layer. The posterior probabilities of tar-

gets at the final layer after proper scaling are adopted in place of GMM likelihoods within the standard HMM speech recogniser [33]. The stacked architectures are based on replacing the dashed block in Figure 1 by another MLP of tandem or hybrid configuration. Though all these architectures are based on MLPs, the final speech recognisers often show different error behaviours. This is where system combination approaches [35, 16] may yield further gains in transcription accuracy.

Such MLP-based architectures offer flexibility into the use of augmenting data. For instance, there are options how it can be exploited in the tandem architecture. One option is to only re-train the GMM whilst keeping MLP parameters fixed to the estimates obtained from the supervised data. Another option is to only re-train the MLP. The third option is to re-train both, the GMM and MLP. For hybrid architectures it is common to re-train the whole system on the augmented data [12], although it is possible to consider fine-tuning on the supervised data only. The stacked architectures offer more flexibility though for simplicity they were not investigated in this paper.

In addition to architecture, the optimal approach will also depend on the nature and amount of the particular data used. For instance, it is not obvious which parts are best kept unilingual and which are better to train multi-lingual in case of augmenting data from other languages. Also, there is a clear limit to the usefulness of schemes such as VTLP. Furthermore, some of the augmenting data types may not combine well in practice.

4. Experiments

Experiments were conducted on two limited language packs released by IARPA Babel program: Assamese and Zulu.[1] The data is recorded in real conditions, such as conversational telephone speech in a range of acoustic conditions. There are also provided phone set and phonetic lexicon, which contains only words that appear in the supervised training data transcriptions. The amount of supervised data is 12 and 14 hours for Assamese and Zulu respectively. The underlying transcriptions were used to create a bigram language model (LM) for discriminative training [31] and trigram LM for decoding. The development sets for both tasks contain approximately 10 hours of data. The experiments were conducted using an extended version of CUED's HTK-3.4.1 toolkit [31] providing GMM-based HMM speech recognition techniques, an extended version of ICSI's QuickNet toolkit [36] providing MLP techniques and IBM's proprietary KWS system [37] for keyword searching.

4.1. Speech recognition system

The tandem architecture was selected for investigation. A consistent procedure was used to create tandem systems. This largely followed [38] and consists of three stages. In the *first stage* a speaker-independent GMM-based HMM is built based on PLP. This applies maximum likelihood (ML) training, heteroscedastic linear discriminant analysis (HLDA) and discriminative, minimum phone error (MPE) [39], training. The HMM states were phonetic decision tree clustered into 1000 unique states following the procedure in [3]. The first stage system was used to produce hypotheses for adaptation by running a Viterbi decoding with the trigram LM over the development sets. The *second stage* is the MLP training. A simple MLP topology with 3 hidden, 1 input and 1 bottleneck layer was adopted. The targets were set to 1000 unique states derived for the first stage system. The input to MLP was a 504-dimensional stack of 4 past, 1 current and 4 future vectors, where each vector was a 13-dimensional PLP feature vector augmented with pitch [40], its delta (Δ), delta-delta (Δ^2) and triples (Δ^3). The MLP was pre-trained layer-wise and fine-tuned using a cross-entropy criterion [33]. The first stage system was used to provide targets. This MLP was used to provide 26-dimensional BN features for training and development data. These BN features were concatenated with 52-dimensional PLP+Δ+Δ^2+Δ^3 and 3-dimensional pitch+Δ+Δ features. The *third stage* is the tandem build. This stage (re-)estimates HLDA transform for PLP and global semitied transform [41] for BN. This reduces dimensionality of tandem features from 81 to 68. The SI tandem system is then built similar to the first stage SI system. In addition to SI, the third stage also performs CMLLR-based speaker adaptive training (SAT) first using ML [21] and then feature-space MPE (fMPE-SAT) criterion [42]. The CMLLR transforms during fMPE-SAT were fixed to ML estimates and not re-estimated. This final system was used for decoding. Prior to this, CMLLR and MLLR transforms were estimated using initial hypotheses produced by the first stage system. These transforms were then used in Viterbi decoding with the bigram LM to produce lattices. Though these lattices could be rescored with more advanced LMs, this was not done in this initial investigation. The accuracy of speech recogniser was assessed based on token error rate (TER) in percentage points (%).[2] The TER performance of this system on Assamese and Zulu was 69.4 and 78.4%, as shown in the first row of Tables 3 and 2.

4.2. Keyword search system

For the 2014 Evaluation the IARPA Babel program required each submitted system to be assessed in keyword search capacity. The task was to find all the exact matches of a query in the development set. The KWS performance is measured according to the maximum term weighted value (MTWV), a metric that takes into account the probabilities of misses and false alarms with larger MTWV values corresponding to better KWS performance. The supplied queries were split into in-vocabulary (IV) and out-of-vocabulary (OOV) parts. For the IV queries the set of word lattices is searched [43] to retrieve the list of hits. For the OOV queries a different approach is used. This operates at the phone level by converting OOV queries into phonetic representation using a grapheme-to-phoneme converter [44]. A soft search, which may improve recall whilst degrade precision, is then performed by expanding the obtained phonetic OOV query representation using phone-to-phone confusion matrix [44]. Only 100 representations with the highest score were retained. Furthermore, the language model scores during search were zeroed as this was found to improve KWS performance. The same approach was also adopted with those IV queries that produced no hits (IV-OOV). The combined list of hits including IV, OOV and IV-OOV parts after sum-to-one normalisation [37] is used to compute MTWV.

4.3. Data augmentation

Two data augmentation schemes, semi-supervised (`semi`) training and VTLP (`vtlp`), were considered. The unsupervised

[1] The precise code identifiers are `IARPA-babel102b-v0.5a` and `IARPA-babel206b-v0.1e`. These releases additionally contain full language packs, where the amount of transcribed data is roughly 70 hours.

[2] The TER is used for consistency of reporting performance for all Babel program languages, such as Assamese and Zulu, where token is a word, or Vietnamese, where token is a syllable or foreign word, or Cantonese, where token is a character.

data is provided by the limited language pack release and conversational portion of the full language pack. The unsupervised data was selected as described in Section 2.1. Here, the tandem system was used to produce lattices. These lattices were then converted into confusion networks to yield word confusion scores [16]. The confusion scores were weighted by the average number of frames to yield the final score for data selection. The data selection process followed [17] and retained half of the unsupervised data. This corresponds to the threshold of 0.4 and 0.3 for Assamese and Zulu respectively as can be seen from Figure 2. The perturbed data, `sup+vtlp` and `semi+vtlp`,

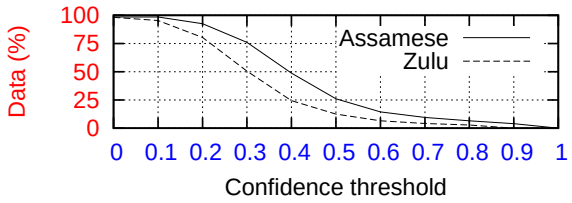

Figure 2: *Percentage of unsupervised data retained for semi-supervised training at different confidence threshold values*

was obtained as discussed in Section 2.2. Here, the original data, `sup` and `semi`, was perturbed 4 and 8 times for Assamese and Zulu respectively as the former was found sensitive to the larger amount of perturbed data. One perturbation factor was fixed to 1, which yields the original data, and the rest, 3 and 7, were randomly sampled from $[0.8, 1.2]$ range for each side. The perturbed semi-supervised (`semi+vtlp`) data was created in the same way. This provides with an approach to increase the amount of unsupervised data where confidence in the accuracy of the underlying transcriptions is above 0.4 and 0.3 thresholds. The amounts of supervised (`sup`) and other types of augmenting data are summarised in Table 1.

Table 1: *Augmenting data quantity in hours*

Type	Assamese	Zulu
`sup`	12	14
`sup+vtlp`	48	110
`semi`	42	49
`semi+vtlp`	152	213

4.4. Results

A range of experiments was conducted to assess the usefulness of augmenting data starting with the more challenging language, Zulu. The first experiment examined the impact of re-training MLP only. The first block of TER results in Table 2 third column shows that the use of augmenting data yields gains over the supervised data. In particular, the combined approach, (`semi+vtlp`, the fourth line) yields the largest 1.7% absolute improvement. The next experiment assessed whether increasing the complexity of GMM system may further improve the results. The tandem system was retrained on the semi-supervised data starting from the first stage. The number of unique states was increased to 3000. This yields 76.9% TER performance as shown on the first line of the second block in Table 2. The use of additional data for training MLP in this case gives small improvement. Re-training the tandem on perturbed semi-supervised data with 5000 unique states yields no additional improvement in TER performance (last line in Table 2). The second series of experiments was conducted on Assamese. The results in Table 3 show a pattern similar to that of Zulu apart from a rather limited usefulness of perturbed compared to unsupervised data.

Table 2: *Zulu*

GMM	MLP	TER	MTWV
`sup`	`sup`	78.4	0.1362
`sup`	`sup+vtlp`	77.1	0.1496
`sup`	`semi`	77.7	0.1468
`sup`	`semi+vtlp`	76.7	0.1446
`semi`	`semi`	76.9	0.1490
`semi`	`semi+vtlp`	76.1	0.1441
`semi+vtlp`	`semi+vtlp`	76.1	0.1454

Table 3: *Assamese*

GMM	MLP	TER	MTWV
`sup`	`sup`	69.4	0.2286
`sup`	`sup+vtlp`	69.3	0.2355
`sup`	`semi`	67.6	0.2309
`sup`	`semi+vtlp`	68.3	0.2341
`semi`	`semi`	66.9	0.2221
`semi`	`semi+vtlp`	66.9	0.2291

Although the above results indicate that data augmentation schemes may be useful for improving TER performance, the ultimate measure of interest in the IARPA Babel program is MTWV. The KWS results in the fourth column of Tables 2 and 3 show that consistent gains in MTWV are also possible. These results also illustrate that improvements in TER do not necessarily translate into improvements in MTWV [3]. For both languages the best MTWV is obtained with the GMM trained on supervised data and MLP trained on perturbed supervised data. The use of standard and perturbed semi-supervised data for training MLP yields a slightly lower MTWV. However, in this case re-training GMM on the semi-supervised data may hurt performance. This indicates that the approach is also sensitive to the accuracy of training transcriptions.

5. Conclusions

Providing accurate speech recognition and keyword searching capabilities for low resource languages is a challenging task. This paper examined an approach, data augmentation, that aims to increase the quantity of available data. Particular schemes discussed were semi-supervised training, acoustic data perturbation, speech synthesis and multi-lingual processing. This paper also discussed various ways to exploit this data in tandem and hybrid architectures. Two of these schemes, semi-supervised training and acoustic data perturbation, individually and in combination were applied within the tandem architecture for two low resource languages, Assamese and Zulu. Speech recognition performance gains were observed from the use of both scheme, with the combined scheme yielding largest gain only for Zulu. Keyword search results showed that gains are also possible, however, the use of semi-supervised training yielded mixed results in this case suggesting sensitivity of the approach to the accuracy of training data transcriptions.

6. Acknowledgements

The authors are grateful to IBM for the KWS system.

7. References

[1] G. Evermann, H. Y. Chan, M. J. F. Gales, T. Hain, A. Liu, D. Mrva, L. Wang, and P. C. Woodland, "Development of the 2003 CU-HTK conversational telephone speech transcription system," in *ICASSP*, vol. 1, 2004, pp. 249–252.

[2] 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.

[3] K. M. Knill, M. J. F. Gales, S. P. Rath, P. C. Woodland, C. Zhang, and S.-X. Zhang, "Investigation of multilingual deep neural networks for spoken term detection," in *ASRU*, 2013.

[4] R. Hsiao, T. Ng, F. Grezl, D. Karakos, S. Tsakalidis, L. Nguyen, and R. Schwartz, "Discriminative semi-supervised training for keyword search in low resource languages," in *ASRU*, 2013, pp. 440–445.

[5] M. Saraclar, A. Sethy, B. Ramabhadran, L. Mangu, X. Cui, B. Kingsbury, and J. Mamou, "An empirical study of confusion modeling in keyword search for low resource languages," in *ASRU*, 2013, pp. 464–469.

[6] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation of multivariate gaussian mixture observations of markov chains," *IEEE Tran SAP*, vol. 2, no. 2, pp. 291–298, 1994.

[7] L. Lamel and J.-L. Gauvain, "Lightly supervised and unsupervised acoustic model training," *Computer speech and language*, vol. 16, pp. 115–129, 2002.

[8] M. J. F. Gales, A. Ragni, H. AlDamarki, and C. Gautier, "Support vector machines for noise robust ASR," in *ASRU*, 2009, pp. 205–210.

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

[10] N. Kanda, R. Takeda, and Y. Obuchi, "Elastic spectral distortion for low resource speech recognition with deep neural networks," in *ASRU*, 2013, pp. 309–314.

[11] Y. Qian, K. Yu, and J. Liu, "Combination of data borrowing strategies for low-resource LVCSR," in *ASRU*, 2013, pp. 404–409.

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

[13] D. Povey, M. J. F. Gales, D. Y. Kim, and P. C. Woodland, "MMI-MAP and MPE-MAP for acoustic model adaptation," in *Eurospeech*, 2003, pp. 1981–1984.

[14] Z. Tüske, J. Pinto, D. Wilett, and R. Schlüter, "Investigation on cross- and multilingual MLP features under matched and mismatched acoustical conditions," in *ICASSP*, 2006, pp. 7349–7353.

[15] G. Zavaliagkos and T. Colthurst, "Utilizing untranscribed training data to improve performance," in *BNTU*, 1998, pp. 301–305.

[16] G. Evermann and P. C. Woodland, "Large vocabulary decoding and confidence estimation using word posterior probabilities," in *ICASSP*, vol. 3, 2000, pp. 1655–1658.

[17] L. Wang, M. J. F. Gales, and P. C. Woodland, "Unsupervised training for Mandarin broadcast news and conversation transcription," in *ICASSP*, vol. 4, 2007, pp. 353–356.

[18] L. Wang and P. C. Woodland, "Discriminative adaptive training using the MPE criterion," in *ASRU*, 2003, pp. 279–284.

[19] H. Zen, Y. Nankaku, and K. Tokuda, "Continuous stochastic feature mapping based on trajectory HMMs," *IEEE Tran ASLP*, pp. 417–430, 2010.

[20] H. Ye and S. J. Young, "High quality voice morphing," in *ICASSP*, vol. 1, 2013, pp. 9–12.

[21] 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.

[22] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based approach to multilingual speech synthesis," in *Text to speech synthesis: new paradigms and advances*, S. Narayanan and A. Alwan, Eds. Prentice Hall, 2004, ch. 7, pp. 135–153.

[23] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, pp. 1039–1064, 2009.

[24] M. J. F. Gales, "Model-based approaches to handling uncertainty," in *Robust speech recognition of uncertain or missing data. Theory and applications*, D. Kolossa and R. Haeb-Umbach, Eds. Springer, 2011, ch. 5, pp. 101–126.

[25] K. Ogata, M. Tachibana, J. Yamagishi, and T. Kobayashi, "Acoustic model training based on linear transformation and MAP modification for HSMM-based speech synthesis," in *Interspeech*, 2006, pp. 1328–1331.

[26] T. Yoshimura, K. Tokuda, T. Masuko, and T. Kobayashi, T. Kitamura, "Speaker interpolation in HMM-based speech synthesis system," in *Eurospeech*, 1997, pp. 2523–2526.

[27] V. Venkataramani and W. Byrne, "Lattice segmentation and support vector machines for large vocabulary continuous speech recognition," in *ICASSP*, 2005, pp. 817–820.

[28] T. Schultz and A. Waibel, "Fast bootstrapping of LVCSR systems with multilingual phoneme sets," in *Eurospeech*, 1997, pp. 371–373.

[29] P. Beyerlein, W. Byrne, J. M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, and W. Wang, "Towards language independent acoustic modellings," in *ASRU*, 1999.

[30] K. Vesely, M. Karafiat, F. Grezl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *SLT*, 2012, pp. 336–341.

[31] 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)*. `http://htk.eng.cam.ac.uk`: University of Cambridge, 2009.

[32] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *ICASSP*, vol. 3, 2000, pp. 1635–1638.

[33] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *Sig Proc Mag*, vol. 29, pp. 82–97, 2012.

[34] C. Plahl, R. Schlüter, and H. Ney, "Hierarchical bottle neck features for LVCSR," in *Interspeech*, 2010, pp. 1197–1200.

[35] J. Fiscus, "A post-processing system to yield reduced word error rates: Recog- niser output voting error reduction (ROVER)," in *ASRU*, 1997, pp. 347–352.

[36] D. Johnson, "Quicknet," in *ICSI*, Berkeley, USA, 2004.

[37] J. Mamou, J. Cui, X. Cui, M. J. F. Gales, B. Kingsbury, K. M. Knill, L. Mangu, D. Nolden, M. Picheny, B. Ramabhadran, R. Schlüter, A. Sethy, and P. C. Woodland, "System combination and score normalization for spoken term detection," in *ICASSP*, 2013, pp. 8272–8276.

[38] J. Park, F. Diehl, M. J. F. Gales, M. Tomalin, and P. C. Woodland, "The efficient incorporation of MLP features into automatic speech recognition systems," *Computer speech and language*, vol. 25, pp. 519–534, 2011.

[39] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, Cambridge University, 2003.

[40] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science B. V., 1995, ch. 14, pp. 495–518.

[41] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Tran SAP*, vol. 29, pp. 82–97, 2012.

[42] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," in *ICASSP*, vol. 1, 2005, pp. 961–964.

[43] C. Allauzen, M. Mohri, and M. Saraclar, "General indexation of weighted automata - application to spoken utterance retrieval," in *IASIR*, 2004, pp. 33–40.

[44] L. Mangu, H. Soltau, H.-K. Kuo, B. Kingsbury, and G. Saon, "Exploiting diversity for spoken term detection," in *ICASSP*, 2013, pp. 8282–8286.