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

Improving the Lwazi ASR baseline

Charl van Heerden[12]*, Neil Kleynhans*[12] *and Marelie Davel*[12]

[1]Multilingual Speech Technologies, North-West University, South Africa.
[2]NWU-CAIR, CSIR Meraka, South Africa.
cvheerden@gmail.com

Abstract

We investigate the impact of recent advances in speech recognition techniques for under-resourced languages. Specifically, we review earlier results published on the Lwazi ASR corpus of South African languages, and experiment with additional acoustic modeling approaches. We demonstrate large gains by applying current state-of-the-art techniques, even if the data itself is neither extended nor improved. We analyze the various performance improvements observed, report on comparative performance per technique – across all eleven languages in the corpus – and discuss the implications of our findings for under-resourced languages in general.
Index Terms: speech recognition, Lwazi, Lwazi ASR corpus, phone recognition, South African languages.

1. Introduction

Automatic speech recognition (ASR) of under-resourced languages is a topic that has garnered increasing interest over the past decade [1]. Targeted data collection efforts such as Globalphone [2], Babel [3] and others [4, 5] have steadily increased the language coverage of available speech corpora. At the same time, freely available tools for data collection [6, 7, 8] have made small localized corpus development much easier, also contributing to the growing pool of curated ASR training data. Still, the majority of sub-Saharan African languages remain under-resourced, with limited or no speech resources available for the study of many of these languages.

In parallel with work targeted at dealing with issues specific to under-resourced languages, recent developments in mainstream ASR research have resulted in clear performance improvements. Specifically, the application of deep neural networks [9], sub-space Gaussian modeling [10] and the packaging of many of these techniques within the Kaldi toolkit [11], have contributed to improved performance in well-resourced ASR systems.

In this study, we revisit earlier baselines obtained on the Lwazi ASR corpus [12], a small freely available corpus of telephony speech in the eleven official languages of South Africa, and determine how these baselines are affected by recent developments. We consider performance trends across a range of languages, in order to better understand the implications for smaller ASR corpora in general.

2. Background

As background to this work, we provide an overview of the Lwazi corpus (Section 2.1), discuss earlier baselines obtained on this corpus (Section 2.2), and touch on those recent developments in ASR that we focus on in this study (Section 2.3).

2.1. The Lwazi corpus

The Lwazi project [13] was originally conceptualized to demonstrate the potential of speech technologies in providing access to information [14]. At the end of the first phase (2006–2009), basic speech recognition and text-to-speech systems were developed in all eleven of South Africa's official languages. (For the majority of these languages, this was the first time such technologies were developed.) In addition, resources developed included annotated speech corpora [12] and electronic pronunciation dictionaries [15], all of which were made available freely via the South African Resource Management Agency [16].

The languages included in the corpus are listed in Table 1, with nine of the eleven languages from the Southern Bantu (SB) family. Per language, the ISO 639-3:2007 language code, language family and estimated number of first language speakers in South Africa are shown. The majority of Southern-Bantu languages are from two language families – Nguni and Sotho-Tswana – with Tshivenda and Xitsonga from two additional language families.

Table 1: *Languages in the Lwazi ASR corpus [17, 18].*

language	ISO code	# million speakers	language family
isiZulu	zul	11.6	SB:Nguni
isiXhosa	xho	8.2	SB:Nguni
Afrikaans	afr	6.9	Germanic
English	eng	4.9	Germanic
Sepedi	nso	4.6	SB:Sotho-Tswana
Setswana	tsn	4.0	SB:Sotho-Tswana
Sesotho	sot	3.8	SB:Sotho-Tswana
Xitsonga	tso	2.3	SB:Tswa-Ronga
siSwati	ssw	1.3	SB:Nguni
Tshivenda	ven	1.2	SB:Venda
isiNdebele	nbl	1.1	SB:Nguni

The Lwazi 1 corpus was the first set of resources available in all of South Africa's official languages, but is very small in today's terms – consisting of between 4 and 10 hours or speech per language (see Table 2).

2.2. Earlier baselines

Various earlier results have been published with regard to the Lwazi corpus [12, 19, 20, 21, 22], not all directly comparable to the work here. The first recognition results on the Lwazi corpus [12] utilized an earlier version of the corpus, and two follow-up papers experimented with concept recognition [19] and data pooling [20], respectively. Given the construction of the corpus (prompts selected from a limited set of government documents) word-based recognition is heavily biased towards

http://dx.doi.org/10.21437/Interspeech.2016-1412

Table 2: *Size of the released Lwazi ASR corpus.*

language	# total minutes	# speech minutes	# distinct phones
afr	254	217	37
eng (SA)	299	253	44
nbl	612	499	46
nso	568	434	29
sot	426	334	28
tsn	474	363	33
tso	517	400	54
ssw	632	501	40
ven	431	335	39
xho	559	439	51
zul	526	417	44

the construction of the language model. We therefore follow the approach of [12] and [22] to report primarily on phone recognition results. That is, we do not perform word recognition using a language model, but phone recognition using a flat phone recognition grammar. (All phones are considered equally likely to occur at any given point.) This provides a clear indication of acoustic model improvement without additional language modeling effects influencing results.

The closest comparisons available are [21] and [22]: both publications report (in addition to other results) on phone recognition rates, with exactly the same train/test set distributions as used here. (Note that different development sets were used in all studies.) All systems were developed using HTK [23]. Henselmans *e*t al. [21] experimented with additional language modeling approach, adding additional data and reported on both phone and word recognition rates. As comparative results from [21] (using unigram phone recognition) are slightly poorer than reported on in [22], we use [22] as baseline for this study.

2.3. Recent developments

Automatic speech recognition systems are continuously extended to include additional techniques that improve performance. Recent techniques range from feature processing to better speech modeling, and are rapidly made accessible to the speech research community through software toolkits such as Kaldi [24]. Kaldi is an open-source toolkit for speech recognition research. It contains near current speech recognition techniques and provides packaged database-specific *recipes* that demonstrate the use thereof. Acoustic modeling techniques provided by Kaldi include: speaker-adaptive training (SAT), subspace Gaussian modeling (SGMM), deep neural networks for probability estimation (DNN) and system combination techniques.

SAT was introduced to improve acoustic modeling by allowing the models to focus on capturing the intra-speaker variability and reducing the effect of inter-speaker variability [25]. Kaldi supports both model-space and feature-space adaptation. For our SAT experiments we utilized the feature-space maximum likelihood linear regression (fMLLR) approach. In the Kaldi training recipe, the speaker-specific fMLLR transforms were composed with a linear discriminant analysis (LDA) transform and a maximum likelihood linear transform (MLLT) [26]. LDA is applied to spliced features (nine consecutive frames) to reduce the dimension to 40, followed by a MLLT transform that regularizes these features to better fit the diagonal covariance assumption.

An SGMM [10] system models context-dependent Hidden Markov Model state parameters indirectly by mapping from a vector space. The Gaussian mixture model structure has a number of Gaussians per state and shares a component's covariance, weights and means across all states. The state-specific parameters are derived by mapping weights and means using a state-specific vector.

The Maximum Mutual Information (MMI) objective function is used to discriminately train acoustic models. Povey *et al.* [27] makes two modifications to this training procedure; (1) boosting path likelihoods proportional to the difference between the hypothesized and true utterances, and, (2) removing shared paths in the numerator and denominator lattices.

DNNs have recently improved many machine learning applications' accuracies and have been applied successfully in speech recognition [9]. Kaldi has two main DNN recipes but for our experiments we utilized the approach detailed in Veselý *et al.* [28], which utilizes Restricted Boltzmann Machine (RBM) pre-training. In addition, the DNNs were further optimized by using sequence-discriminative training within a state-level minimum Bayes risk (SMBR) criterion framework.

Kaldi implements a score combination technique that makes use of minimum Bayes risk decoding to re-score a union of lattices [29]. This lattice is constructed by combining the lattices from various speech recognition systems – in the case of Kaldi these are the SGMM+MMI and DNN systems. The various word options are scored based on posteriors and the best path is chosen.

3. Approach

The Lwazi corpus is associated with a set of dictionaries that have continued to evolve over the past few years, with minor corrections accumulating over time. In order to ensure that we only analyze the improvements introduced using newer acoustic modeling techniques, we first rerun the same baseline systems as before, using the latest available release of the corpus and dictionaries.

Only once we have ascertained that the baselines are repeatable, newer techniques are introduced, one at a time, using implementations made available via the Kaldi toolkit. The specific techniques considered include:

- Speaker-adaptive training (SAT),
- Subspace Gaussian modeling (SGMM),
- Deep neural networks for probability estimation (DNNs), and
- System combination.

Once each technique has been applied, performance is measured using the standard definition of phone error rate (the sum of all substitutions, deletions and insertions, as a factor of the total number of phonemes in the reference utterance). This definition is however not fully unambiguous, as discussed in Section 5.1.

4. Data

The Lwazi corpus consists of prompted telephone – mobile & landline – speech recordings in all eleven of South Africa's official languages. For each language, approximately 200 speakers were recorded, with each speaker reading 30 prompts on average; this results in approximately 4–10 hours of audio per language, as shown in Table 2.

For the purposes of this paper, the same evaluation sets as in [22] were used. However, development sets for tuning were not

used in [22], therefore a smaller development set of 20 speakers was partitioned from the training set for tuning language model weights and for optimizing iterative training techniques (our training set is thus slightly smaller than that used for training in [22]). Male and female speakers were balanced across development and evaluation sets. The number of speakers per set, as well as the number of male and female speakers per language, are shown in Table 3.

All the data (including train, development and test set partitioning) as well as the pronunciation dictionaries used here are available for download at https://sites.google.com/site/lwazispeechcorpus.

Table 3: *Number of speakers in the train (trn), development (dev) and evaluation (tst) sets of the Lwazi ASR corpus. The total number of speakers is also shown together with a gender breakdown.*

language	trn	dev	tst	total	male	female
afr	140	20	40	200	101	99
eng	136	20	40	196	92	104
nbl	140	20	40	200	99	101
nso	130	20	40	190	92	98
tso	154	20	40	214	103	111
tsn	143	20	40	203	96	107
ssw	136	20	40	196	92	104
sot	142	20	40	202	90	112
ven	138	20	40	198	98	100
xho	150	20	40	210	101	109
zul	139	20	40	199	98	101

5. Results

5.1. Repeating the baseline

The baseline results described in [22] are the closest published results to the experiments reported on in this paper, both in terms of approach as well as in the training and test sets that were used. As discussed in Section 4, no development set was used in [22]; as a development set is necessary for tuning in the Kaldi recipes, we defined our own development sets from the respective training sets. In order to ensure that the results in this paper are as directly comparable as possible to results obtained with the techniques and toolkits reported in [22], the experiments in [22] were repeated using the new training and development sets. As can be seen from Table 4, the new results are comparable to those described in [22], with minor degradations attributable to the smaller training sets, and minor improvements due to the improved dictionaries.

While analyzing the baseline, it was observed that the standard error reporting tools made available via the HTK and Kaldi toolkits produced different results. Specifically, even though the definition of phone error rate remains the same, the error rate is dependent on the scoring matrix used during hypothesis and reference alignment. In HTK, the cost of one correct match, one insertion and one deletion is lower than two substitutions during alignment; the opposite choice is made in Kaldi. (That is, even though the same definition is used to score PER, different scoring strategies are used during reference and hypothesis alignment.) This implies that Kaldi and HTK scoring results are not directly comparable, and the HTK baseline was therefore rescored using the Kaldi approach in order to produce the baseline reported on in the remainder of this section.

5.2. Trends per language

Results for the best and worst performing Sotho-Tswana languages – Sesotho (sot) and Setswana (tsn) – are shown in Fig. 1, and for the best and worst performing Nguni languages – siSwati (ssw) and isiZulu (zul) – in Fig. 2. Across all of these languages, a clear improvement in PER is observed from the *triphone* to *SGMM+MMI* system – with a large gain observed when introducing SAT. (Exact results per language are shown in Table 4.)

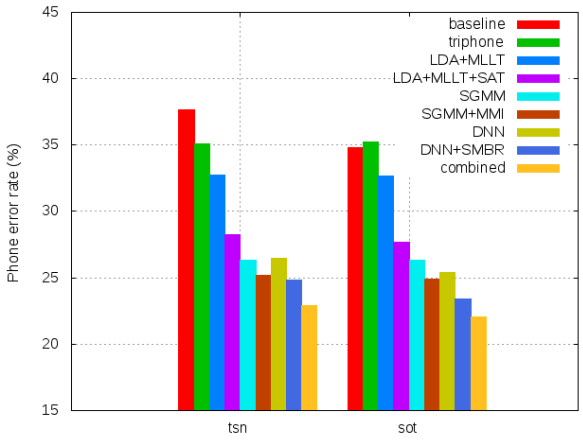

Figure 1: *Phone error rate using different acoustic modeling techniques for two languages from the Sotho-Tswana language family.*

Within the training process there is a split in system development when either *DNN* training or *SGMM+MMI* training commences. Surprisingly, the *DNN* results are comparable to that of *SGMM+MMI* in the majority of cases, where the latter system is on par or slightly out-performing the *DNNs* – except for English and Afrikaans (see below). Consistently, across languages, the *DNN+SMBR* approach produces the best results for a single system. Lastly, the best results for all languages are the *combined* systems, that rescores the combined lattices produced by both the *SGMM+MMI* and *DNN+SMBR* systems.

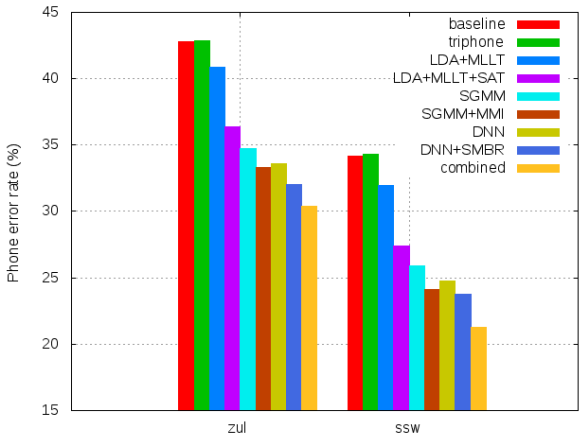

Figure 2: *Phone error rate using different acoustic modeling techniques for two languages from the Nguni language family.*

Table 4: *A comparison of phone error rates across languages, using different acoustic modeling techniques.*

	eng	**zul**	**afr**	**tsn**	**nbl**	**tso**	**nso**	**sot**	**xho**	**ssw**	**ven**
baseline	48.35	42.66	41.30	37.69	35.26	35.12	36.75	35.72	34.03	34.24	32.25
triphone	50.73	41.72	41.86	35.42	35.27	36.41	33.84	35.28	35.25	34.77	32.15
LDA+MLLT	46.11	40.87	38.50	32.74	32.49	33.59	31.17	32.86	31.60	32.56	29.35
LDA+MLLT+SAT	40.77	36.35	33.49	28.86	27.52	27.93	27.16	28.07	27.48	28.20	23.99
SGMM	38.04	34.77	30.84	27.12	24.83	26.21	24.40	26.44	24.93	26.33	21.96
SGMM+MMI	37.27	33.31	30.03	25.52	23.12	24.63	24.33	25.02	23.27	23.63	20.87
DNN	34.57	33.57	27.82	26.35	23.79	24.85	25.16	25.64	23.75	24.76	20.68
DNN+SMBR	33.23	32.05	26.56	24.91	22.39	22.88	23.52	24.20	21.78	23.73	19.24
combined	32.24	30.36	25.29	23.05	20.43	21.55	21.27	22.16	21.08	21.40	17.96

5.3. Results across languages

For completeness, all language results are shown in Table 4. In addition, results from four selected systems – the initial baseline, trained triphones once speaker-adaptive training has been applied (*LDA+MLLT+SAT*), the best single system and the best combined results – are shown in Fig. 3. In all cases the best single system results are obtained using the *DNN+SMBR* system.

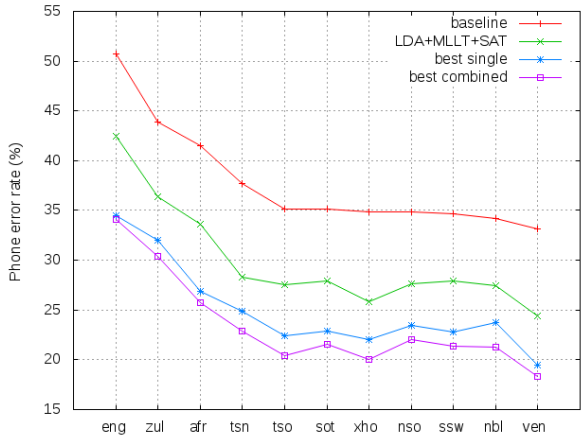

Figure 3: *Comparing language-specific performance across the four main systems.*

6. Discussion

From Table 4 we can clearly see that the more sophisticated techniques produce better results, with large gains observed when compared to previously published baselines. However, these techniques generally require extensive computing resources. Besides providing new baselines for the Lwazi telephony corpus, the speech recognition trends lay out a rough guide to selecting the best technique given technological constraints. In under-resourced environments gaining access to high performance computing – such as GPUs for DNN training and decoding – may be difficult. Therefore backing off to certain techniques might be adequate for an application, and our trends provide a proxy for expected performance.

The same trends were observed across all languages studied, even though absolute accuracies differed substantially. The differences among languages can be attributed to various factors, including the quality of the data itself, but is most heavily influenced by the phonotactic perplexity per language, as also observed in [19]: the lower the perplexity, the better the results. Given the significant differences in the linguistic characteristics of the languages studied, similar gains can be expected for other under-resourced languages. This is especially significant, given the large increases in data sizes typically required to gain reductions in error rates (when keeping modeling techniques constant) [30].

The techniques used in this study have all matured to the extent that limited tweaking is required when applying standard versions. Existing techniques related to both monolingual and multilingual bottleneck features have not yet been applied here: these are techniques we would like to explore in future work.

7. Conclusion

In this paper we investigated the implication of recent advances in acoustic modeling for under-resourced languages, by revisiting existing baselines published on the Lwazi ASR corpus of South African languages. This corpus includes nine under-resourced languages from the Southern Bantu family. To facilitate future comparisons, all data used here, including train, development and test set partitioning, are available for download at https://sites.google.com/site/lwazispeechcorpus.

We demonstrate large gains (16-25% relative) by applying speaker-adaptive training, and additional large gains (12-20% relative) from applying either SGMMs or DNNs for probability estimation. System combination of the best single system results resulted in an additional 1-10% relative gain. This results in an overall improvement of 31-45% relative gain when compared to the previous baselines.

While the corpora used here are small, the reduction in error rate is significant, especially if the data is intended for the development of seed systems used to bootstrap speech technology applications. The same trends were observed across all languages studied, even though absolute accuracies differed substantially. This bodes well for other under-resourced languages, where similar gains are to be expected.

8. Acknowledgments

We would like to thank Brno University of Technology (BUT) and our gracious hosts – Jan (Honza) Černocký, Martin Karafiát, Karel Veselý and team – for support during our visit to BUT, and for access to the BUT computing environment where most of these experiments were conducted.

9. References

[1] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, pp. 85–100, 2014.

[2] T. Schultz, "Globalphone: a multilingual speech and text database developed at Karlsruhe University." in *Proc. Interspeech*, Denver, CO, USA, September 2002, pp. 345–348.

[3] M. P. Harper, "Babel," http://www.iarpa.gov/index.php/research-programs/babel, accessed: 2016-01-20.

[4] T. Hughes, K. Nakajima, L. Ha, P. Moreno, and M. LeBeau, "Building transcribed speech corpora quickly and cheaply for many languages," in *Proc. Interspeech*, Makuhari, Japan, September 2010, pp. 1914–1917.

[5] C. van Heerden, M. H. Davel, and E. Barnard, "The semi-automated creation of stratified speech corpora," in *Proc. PRASA*, Johannesburg, South Africa, December 2013, pp. 115–119.

[6] N. J. De Vries, M. H. Davel, J. Badenhorst, W. D. Basson, F. De Wet, E. Barnard, and A. De Waal, "A smartphone-based ASR data collection tool for under-resourced languages," *Speech Communication*, vol. 56, pp. 119–131, 2014.

[7] I. Lane, A. Waibel, M. Eck, and K. Rottmann, "Tools for collecting speech corpora via Mechanical-Turk," in *Proc. NAACL HLT*, Los Angeles, CA, USA, June 2010, pp. 184–187.

[8] M. H. Davel, C. J. van Heerden, and E. Barnard, "Validating smartphone-collected speech corpora," in *Proc. SLTU*, Cape Town, South Africa, May 2012, pp. 68–75.

[9] 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," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[10] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafit, A. Rastrow, and R. Rose, "The subspace Gaussian mixture model - a structured model for speech recognition," *Computer speech and language*, vol. 25, no. 2, pp. 404–439, 2011.

[11] 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," in *Proc. ASRU*, Big Island, Hawaii, USA, December 2011.

[12] E. Barnard, M. Davel, and C. van Heerden, "ASR corpus design for resource-scarce languages," in *Proc. Interspeech*, Brighton, UK, September 2009, pp. 2847–2850.

[13] CSIR Meraka, "Lwazi: increasing the impact of speech technologies in South Africa," http://www.meraka.org.za/lwazi, accessed: 2016-03-20.

[14] E. Barnard, M. H. Davel, and G. B. van Huyssteen, "Speech technology for information access: a South African case study," in *Proc. AAAI Spring Symposium on AI-D*, Palo Alto, CA, USA, March 2010, pp. 8–13.

[15] M. Davel and O. Martirosian, "Pronunciation dictionary development in resource-scarce environments," in *Proc. Interspeech*, Brighton, UK, September 2009, pp. 2851–2854.

[16] Resource Management Agency, "RMA Home," http://rma.nwu.ac.za, accessed: 2016-03-20.

[17] E. Barnard, M. H. Davel, C. J. V. Heerden, F. D. Wet, and J. Badenhorst, "The NCHLT speech corpus of the South African languages," in *Proc. SLTU*, May 2014, pp. 194–200.

[18] Statistics South Africa, "Census 2011: Census in brief," De Bruyn Park Building, 170 Thabo Sehume Street, Pretoria, 0002, Tech. Rep. 03-01-41, 2012. [Online]. Available: www.statssa.gov.za

[19] C. van Heerden, M. H. Davel, and E. Barnard, "Medium-vocabulary speech recognition for under-resourced languages," in *Proc. SLTU*, Cape Town, South Africa, May 2012, pp. 146–151.

[20] C. van Heerden, N. Kleynhans, E. Barnard, and M. Davel, "Pooling ASR data for closely related languages," in *Proc. SLTU*, Penang, Malaysia, May 2010, pp. 17–23.

[21] D. Henselmans, T. Niesler, and D. van Leeuwen, "Baseline speech recognition of South African languages using Lwazi and AST," in *Proc. PRASA*, Johannesburg, South Africa, December 2013, pp. 30–35.

[22] C. van Heerden, J. Badenhorst, F. de Wet, and M. Davel, "Lwazi ASR evaluation: Technical report," Meraka Institute, CSIR, Pretoria, South Africa, Tech. Rep. 1, May 2013.

[23] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK book," *Cambridge University Engineering Department*, vol. 3, p. 175, 2002.

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, and M. H. et al., "The Kaldi speech recognition toolkit," in *IEEE Workshop on automatic speech recognition and understanding*, Waikoloa, HI, USA, December 2011.

[25] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer speech and language*, vol. 12, no. 2, pp. 75–98, 1998.

[26] R. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," in *Proc. ICASSP*, Seattle, Washington, USA, May 1998, pp. 661–664.

[27] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proc. ICASSP*, Las Vegas, Nevada, USA, April 2008, pp. 4057–4060.

[28] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. Interspeech*, Lyon, France, August 2013, pp. 2345–2349.

[29] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum Bayes risk decoding and system combination based on a recursion for edit distance," *Computer speech and language*, vol. 25, no. 4, pp. 802–828, 2011.

[30] R. K. Moore, "A comparison of the data requirements of automatic speech recognition systems and human listeners," in *Proc. Interspeech*, Geneva,Switzerland, September 2003, pp. 2581–2584.