Improving Language Recognition with Multilingual Phone Recognition and Speaker Adaptation Transforms

Andreas Stolcke Murat Akbacak Luciana Ferrer Sachin Kajarekar
Colleen Richey Nicolas Scheffer Elizabeth Shriberg

Speech Technology and Research Laboratory
SRI International, Menlo Park, CA, USA
stolcke@speech.sri.com

Abstract

We investigate a variety of methods for improving language recognition accuracy based on techniques in speech recognition, and in some cases borrowed from speaker recognition. First, we look at the question of language-dependent versus language-independent phone recognition for phonotactic (PRLM) language recognizers, and find that language-independent recognizers give superior performance in both PRLM and PPRLM systems. We then investigate ways to use speaker adaptation (MLLR) transforms as a complementary feature for language characterization. Borrowing from speech recognition, we find that both PRLM and MLLR systems can be improved with the inclusion of discriminatively trained multilayer perceptrons as front ends. Finally, we compare language models to support vector machines as a modeling approach for phonotactic language recognition, and find them to be potentially superior, and surprisingly complementary.

1. Introduction

Language recognition (or language identification, LID) systems are commonly based on a combination of two main modeling approaches. One the one hand, short-term cepstral features (typically including shifted delta features [1]) are modeled using Gaussian mixture models (GMMs) with joint factor analysis (JFA) [2, 3]. Second, one or more language-specific, unconstrained phone recognizers are used to tokenize the speech into phone sequences, which are then modeled using statistical language models [4] ([parallel] phone recognition language modeling, [P]PRLM). Scores from both kinds of system are combined and calibrated using a variety of techniques, such as Gaussian back ends [5] or multiclass logistic regression [6].

In this paper we take a closer look at approaches based on phone recognition, and investigate if recent advances in automatic speech recognition (ASR), as well as corresponding phone-based methods developed for speaker recognition, can be leveraged for improved language recognition. This work was in part inspired by work at LIMSI [7, 8] showing that advanced ASR techniques, such as lattice decoding,[1] speaker-adaptive training, and context-dependent phone modeling can result in significant improvements for phonotactic LID systems and yield state-of-the-art performance. We build on these techniques here and investigate additional improvements. First, we explore the use of language-independent phone sets and corresponding phone recognizers, which should give better coverage of a wide acoustic-phonetic space as found in language recognition. Second, we investigate phone recognition based on acoustic front end features that are themselves estimated by neural networks and trained discriminatively for phone recognition. Such features have yielded substantial improvements for language-specific speech recognition [10], have been shown to generalize to languages not used in training [11], and have been incorporated in language-specific PRLM systems [6].

Additional improvements come from techniques that have been found to work well for speaker recognition, but are not yet generally used in language recognition. In this category we have models that use speaker adaptation transforms, estimated by maximum likelihood linear regression (MLLR), as features [12]. This technique provides an alternative cepstral modeling approach that is informed by speech recognition (different transforms for different phone classes) and should in principle be able to model language differences as well. MLLR modeling was already shown to work well for native versus nonnative accent discrimination [13]. A second question inspired by developments in speaker recognition concerns the best modeling approach for phonotactic N-grams. Results in phone-based speaker recognition showed that discriminative models in the form of support vector machines (SVMs) are more accurate than statistical language models [9]. This suggests using SVMs for phonotactic LID as well. Like neural-network-based front ends, this was previously investigated with a language-specific

[1] Lattice-based phonotactic modeling was independently found to give improvements in speaker recognition [9].

PRLM system [6]; here we revisit SVM N-gram modeling in the context of the aforementioned other techniques, including multilingual phone recognition.

2. Method

2.1. Data

All experiments reported here use the NIST Language Recognition Evaluation (LRE) 2005 data set.[2] The evaluation set comprises the seven target languages: English, Mandarin, Spanish, Hindi, Japanese, Korean, and Tamil. All training data was drawn from the LDC CallFriend corpus and comprised about 56 hours each for English, Mandarin, and Spanish, and about 26 hours each for Hindi, Japanese, Korean, and Tamil. The test data consists of 3662 samples of conversational speech of about 30 seconds each. For evaluation purposes we split the test set into two subsets and perform twofold crossvaluation, i.e., the system combination weights and calibration thresholds are estimated on the complement of the half of the data that is being scored. The primary metric used in our experiments is the macro-average of the seven per-language equal error rates (avgEER).

Phone recognition and other ASR models were trained on other datasets described further below, which were separate from the LRE training materials used to collect language recognition model statistics.

2.2. Cepstral system

We wanted to be able to evaluate different phone-based LID methods and systems not just in isolation but also in combination with a state-of-the-art cepstral baseline system. To this end, we developed a GMM-JFA system based on a 56-dimensional feature vector consisting of energy and a shifted-delta cepstrum in 7-1-3-7 configuration [1]. Deltas were computed over the whole file (including nonspeech regions), and feature frames were removed based on SRI's speech/nonspeech segmenter. We estimated 2048-component GMMs with the LRE training data in the seven target languages. The same data was used to train a 300-dimensional eigenchannel subspace [14]. Scores for each test sample were produced using dot product scoring [15].

Combination of raw system scores (if multiple systems are used) and their calibration is performed using multiclass logistic regression as implemented in Niko Brümmer's FoCal Multi-class toolkit [16].

The baseline system achieved an avgEER of 2.87% on the test data.

3. Phonotactic Language Modeling

3.1. Phone recognizers

As a baseline for phonotactic language recognition, we build a PPRLM system utilizing three language-specific recognizers (for American English, Spanish, and Levantine Colloquial Arabic), trained on available conversational telephone corpora available from the LDC. Language-dependent properties of these systems are listed in Table 1. Apart from gender dependence (which is enabled by larger training data size) all systems use similar acoustic modeling: perceptual linear prediction (PLP) front end with up to third-order difference features, vocal tract length normalization, dimensionality reduction via heteroscedastic linear discriminant analysis (HLDA), and triphone acoustic models trained using the minimum phone error (MPE) criterion. Decoding uses an open phone loop (no phonotactic constraints) and generates phone lattices, from which phone N-grams with posterior probability weighted frequencies are extracted. Following [8], we also perform constrained MLLR speaker adaptation prior to decoding.

Language-dependent recognizers by their nature will model different languages more or less accurately, and might not perform as well for recognizing languages that are outside the acoustic or phonotactic space covered by the languages chosen. An alternative is to train a single recognizer that incorporates data from a variety of languages. While still possibly biased, such a recognizer might achieve better generalization for LID purposes if commonalities among all languages are modeled to some extent.

To test the suitability of multilingual recognizers for LID, we developed a system that is based on a shared set of 52 phones that give a reasonable representation of four fairly diverse languages (American English, Mandarin, Spanish, and Egyptian Arabic), while glossing over fine-grained distinctions in the language-specific phone sets (such as tone in Mandarin).[3] For training purposes, the word-level transcripts in each language were mapped to the multilingual phone set, and the training data was pooled. Training statistics for Spanish and Egyptian Arabic were given extra weight to achieve a more balanced coverage of all languages in the final models. Also, American English data was selected to roughly equate the amount of data from native and nonnative speakers, since the raw training corpora tend to be dominated by the former. The composition of the multilingual training corpus is given in Table 2. Note that the combined multilingual training set is 370h, or less than one quarter of the largest language-dependent (English) training set.

Language recognition scores are obtained by com-

[2] At the time of this study, this was the most recent publicly available LRE dataset from the Linguistic Data Consortium (LDC).

[3] The choice of the Egyptian colloquial dialect of Arabic (ECA) was determined by the fact that vowelized transcripts for ECA are available, but are not for other dialects.

Table 1: Properties of language-specific phone recognizers.

Language	Phoneset size	Training data	Gender dependence
American English	47	1400 h	yes
Spanish	33	18 h	no
Levantine Arabic	39	61 h	no

Table 2: Data used for multilingual phone recognizer training.

Language	Sources	Duration	Weighting
American English (native)	Fisher, Switchboard, CallHome	123h	1x
American English (nonnative)	Fisher	108h	1x
Mandarin Chinese	CallHome	103h	1x
Spanish	CallHome	19h	3x
Egyptian Arabic	CallHome	17h	3x

Table 3: Results with phone recognition language modeling

Systems used	%avgEER
American English	4.17
Levantine Arabic	4.91
Spanish	5.49
Am.Eng.+Levant.	2.99
Am.Eng.+Levant.+Span.	2.76
Multilingual	3.01
Am.Eng.+Levant.+Span.+ML	2.09

puting the length-normalized log likelihood ratio of the target phone language model L relative to the nontarget language model, given the test sample phone decoding output X:

$$s = \frac{1}{|X|} \frac{\log P(X|L)}{\log P(X|\bar{L})} \qquad (1)$$

where $|X|$ is the number of phone tokens in the sample, and $\bar{L}$ is the union of all nontarget languages.

3.2. Language-dependent versus multilingual PRLM

Table 3 shows results with various PRLM and PPRLM configurations. All results were obtained with 3-gram phone language models, since 4-grams did not give better results. English is the single best language-dependent system (4.17%), and combining two or three such systems gives substantial gains (to 2.99% and 2.76%, respectively), albeit with diminishing returns. The multilingual PRLM by itself gives almost the same level of performance (3.01%) as the combination, and by adding it to the combination a further 24% relative gain (to 2.09% absolute avgEER) can be achieved.

Table 4: Results with phone recognition enhanced with MLP features

System	Front end	%avgEER
Multilang. PRLM	PLP	3.01
Multilang. PRLM	PLP+MLP	2.82
PPRLM	PLP	2.09
PPRLM	PLP+MLP	1.77

3.3. Use of discriminative MLP features

In prior work we have made extensive use of acoustic features computed by multilayer perceptrons (MLPs, or neural networks), trained to perform phone discrimination at the frame level, with excellent gains in word recognition accuracy [10]. More recently, the Brno team has experimented with a similar hybrid MLP/HMM phone recognizer for PRLM language recognition [6].

Although the MLP features used here were not trained specifically for the languages of our phone recognizers (or the languages of the LRE task), we had previously found that training such MLPs on a large English database yielded features that could help recognition in other languages as well [11]. This suggests appending the English-trained MLP features to the standard PLP features of our multilingual recognizer. We observed that phone recognition accuracies increase between 2 and 4% absolute with the modified recognizer. The question is whether this improvement also translates into improved language recognition.

Table 4 shows results for both PRLM and PPRLM (combined with language-dependent PRLMs). As shown, the avgEER is reduced for both PRLM, for a relative gain of 6%, and for PPRLM, by 15% relative, compared to otherwise similar systems that use only the stan-

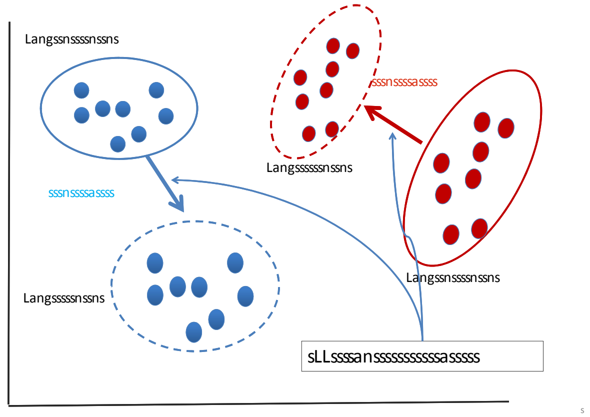

Figure 1: Feature extraction via MLLR model adaptation

dard PLP acoustic front end. (Note that only the multilingual PRLM system is augmented with MLP features in these experiments.)

4. Speaker Adaptation Transform Models

4.1. Feature extraction and modeling

In speaker recognition, approaches based on modeling an affine transform that adapts speaker-independent models to speaker-dependent models have been very effective [12]. In the language recognition setting, we can similarly estimate MLLR transforms that map from a language-independent speech model to a language-dependent one, and model the transform parameters as features (Figure 1). In fact, as with PRLM, the underlying speech models can be derived from any specific language, or using multiple languages, as long as transforms are estimated consistently for target language training and test data.

Following the general recipe in [12], the modeling proceeds as follows. Given a speech sample, phone-loop ASR is used to assign frames to phone classes, and an MLLR transform is estimated for each class. The transform for the nonspeech class is discarded, and the coefficients for all other transforms are concatenated into a linear vector. Each vector component is rank-normalized [17] using the combined target language training data as the reference distribution. For each target language, an SVM model with linear kernel function is trained with target language samples as positive instances and all others as negative instances. Given a test sample, the (signed) distance between the test feature vector and the SVM hyperplane serves as the raw (pre-calibration) language recognition score.

4.2. Results

We implemented this MLLR-SVM approach using three MLLR reference acoustic models: an English male-speaker model, an English female-speaker model, and a gender-independent multilingual model. In all cases the acoustic feature dimension is 39, and 8 MLLR transforms (phone classes) were used, yielding an SVM feature vector of length $8 \times 39 \times 40 =$ 12,480.

Table 5: Results with MLLR transform modeling

MLLR model	%avgEER
English, female	12.98
English, male + female	10.25
Multilingual	7.47

Results are given in Table 5. Comparing the first and last result, we again observe that a system based on a single speech model trained on multilingual data fares much better than a similar system trained on monolingual data. The gender-dependence of the English speech models actually affords an advantage, because the feature vectors from the male and female MLLR versions (regardless of test speaker gender) can be concatenated. This approach is effective for speaker recognition [18] and also gives a gain for language recognition, as shown in the second line of Table 5. However, the combination of gender-dependent English MLLR transforms still does not yield as good a result as the gender-independent multilingual MLLR.

4.3. Improvements

The performance of MLLR models is still substantially below that of the cepstral and phonotactic models. However, we can improve the general approach in several ways. First, we note that so far we have extracted one feature vector per training/test sample, regardless of speech duration. This is unlike the cepstral GMM and PRLM systems, which obtain more data points the longer the speech duration. While test samples are roughly 30 seconds in length, training samples consist of much longer recordings (up to several minutes long). We can therefore partition the training samples into shorter segments of roughly 30 seconds each, and obtain multiple training vectors per conversation. Not only does this increase the amount of training samples, it should also do a better job of modeling the variability inherent in short speech excerpts.

Second, we can optimize the number of Gaussians in the MLLR reference models, which was originally chosen for good phone recognition performance, for best language recognition results instead. We were hoping that a more compact reference model would improve language discrimination, as it forces the MLLR transforms to do more of the work of matching the reference model to the language sample (if the MLLR GMM is too large it can do so in part by simply choosing among its Gaussians).

Table 6: Effect of improved MLLR transform modeling

MLLR system	%avgEER
One transform per training sample	7.47
Multiple transforms per training sample	5.98
+ Reduced number of Gaussians	5.19
+ NAP	4.54
+ MLP features	3.96

Third, we can estimate the intra-language variability of the feature space, and apply nuisance attribute projection (NAP) [19] to remove the subspace that has the highest nuisance variability. (This is the same method used to remove intra-speaker variability in SVM-based speaker recognition systems.) NAP is especially attractive in combination with the training-set partitioning method described above, as it allows NAP to estimate and remove intra-session variability, in addition to between-speaker variability.

Finally, as with PRLM, we augment our standard PLP front end with MLP features optimized for (English) phone discrimination. In our case, this adds 25 dimensions to the acoustic feature space. To limit the increase in MLLR parameters we estimate block-diagonal adaptation transforms, increasing the number of MLLR coefficients by $8 \times 25 \times 26$, for a total of 17,680.

Results from implementing these modifications incrementally are shown in Table 6. The largest improvement (20% relative) comes from splitting of training samples, with smaller gains from the other measures. Reducing the number of Gaussians in the MLLR phone models model by a factor of 4 (from 64 to 16), applying NAP, and added MLP features each give about 13% relative error reduction.

5. Phonotactic Modeling with Support Vector Machines

In this section we revisit the modeling choices in phonotactic language recognition. Comparisons on speaker recognition systems based on phone N-grams showed that SVM models can give quite substantial gains over statistical language models [9]. The Brno team has reported good results with SVM modeling of phone N-grams from a Hungarian recognizer [6], so it was important to assess the relative merits of this modeling approach under our multilingual phone recognition framework, leaving all other parameters constant.

We implemented a phonotactic SVM ("PRSVM") system based on the same phone N-gram frequencies as used by the multilingual PRLM system with MLP features (Section 3.3). The lattice-based relative frequencies of the most frequent phone N-grams (as determined on the combined training set) are assembled into a feature vector. Each relative frequency is scaled by the inverse square root of the global frequency of that N-gram, so as to implement the TFLLR pseudo-likelihood-ratio kernel proposed by [20]. The resulting feature vectors are modeled and scored by linear-kernel SVMs as described in Section 4.1. Also, as for MLLR transform SVMs, we split training sessions into 30-second segments to obtain multiple feature vectors per session.

Table 7: Results with multilingual phone N-gram systems

Systems used	%avgEER
Phone N-gram LM	2.82
Phone N-gram SVM (3-gram)	3.01
Phone N-gram SVM (4-gram)	2.74
Phone N-gram LM + SVM	2.42

Table 8: Results with combinations of multiple systems. All noncepstral systems are based on multilingual phone models, but PPRLM also incorporates language-specific models.

Systems used	%avgEER
PRSVM	2.74
Cepstral GMM	2.87
+ MLLR-SVM	2.59
+ PRLM	1.43
+ MLLR-SVM + PRLM	1.19
+ MLLR-SVM + PPRLM	1.24
+ MLLR-SVM + PRLM + PRSVM	1.14

Table 7 summarizes results. The trigram SVM yields an avgEER of 3.01%, slightly worse than the corresponding trigram LM. However, unlike with phonotactic LMs, we found that increasing the maximum N-gram length to 4 lowers the error rate, to 2.74%. We also found that a score-level combination of phonotactic LM and SVM models gives a 12% relative reduction over the single best system, in spite of the two systems differing merely in their modeling.

6. Combining Systems

We now look at score-level combinations of the various systems at our disposal: the cepstral baseline system, phonotactic PRLM and PRSVM, and MLLR-SVM. Results are shown in Table 8. We use the cepstral GMM as our baseline and starting point for all combinations. Adding the MLLR-SVM to the baseline reduces error by 10% relative. Recall that the MLLR-SVM by itself still has a much higher error rate (3.96%), yet the two alternative ways of modeling cepstra apparently give somewhat complementary information. Adding the multilingual PRLM to the cepstral system gives the largest relative error reduction: 50%. This might be explained by the fact that both features and modeling paradigms are very different in these two systems. Adding the MLLR-SVM as a third system gives a surprisingly large gain of 17% relative.

Earlier we saw that adding the three language-dependent PRLM systems to the multilingual PRLM gave a large error reduction. However, this is no longer true once the cepstral and MLLR systems are also included. In fact, there is a slight increase (from 1.19% to 1.24%) in avgEER, which is most likely due to insufficient training data for the logistic regression combiner. (Given sufficient training data, adding another system should never degrade performance.) On the other hand, adding our alternative, SVM-based phonotactic model does give a small (4% relative) additional improvement over the three-way system combination. Overall, we see a 55% relative error reduction over the PRSVM, the single best system.

7. Conclusions and Future Work

Starting from standard cepstral and PRLM systems, we have investigated a number of improved modeling techniques for language recognition, inspired by work in speech recognition and speaker recognition. An alternative form of ASR-mediated cepstral modeling, MLLR-SVMs, while not improving on the cepstral GMM by itself, does give gains in combination with the latter, even after adding phonotactic models. We found that both phonotactic (PRLM) and cepstral MLLR-SVM models work best when based on ASR acoustic models trained with multiple languages and a unified phone set. A multilingual PRLM is much better than any of the language-specific PRLMs, and, by itself, approaches the performance of a three-language PPRLM. Both PRLM and MLLR-SVM can also be improved by adding MLP features trained for phone discrimination to the acoustic front end. Finally, we found that replacing the traditional phonotactic language models with discriminatively trained SVMs over N-gram frequency features can lower error, as the SVMs can apparently benefit from longer N-gram lengths than the statistical language models. Also, combining PRLM and PRSVM models gives additional gains.

We believe that further improvements can be obtained by building on the techniques developed here, and addressing some key open questions. For example, one could train multiple language-specific PRSVMs and combine these systems at either the feature level (vector concatenation) or score level ("PPRSVM"). N-gram feature selection could probably be improved beyond a simple frequency criterion, in order to take advantage of even longer N-grams as long as they provide discriminative information. Finally, an effective form of nuisance variability compensation, which is critical to good cepstral system performance (JFA for GMMs, NAP for SVMs), has so far not been found for phonotactic systems (we tried NAP with PRSVMs but found no gains).

8. Acknowledgments

We thank our SRI colleagues Gokhan Tur and Wen Wang for valuable comments on the work presented here, and the anonymous reviewers for useful suggestions. Thanks also go to Wen Wang and Dimitra Vergyri for help with the language-specific phone recognition systems.

9. References

[1] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. Deller, Jr., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features", in J. H. L. Hansen and B. Pellom, editors, *Proc. ICSLP*, pp. 89–92, Denver, Sep. 2002.

[2] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified", *in Proc. ICASSP*, vol. 1, pp. 637–640, Philadelphia, Mar. 2005.

[3] N. Brümmer, A. Strasheim, V. Hubeika, P. Matějka, L. Burget, and O. Glembek, "Discriminative acoustic language recognition via channel-compensated GMM statistics", *in Proc. Interspeech*, pp. 2187–2190, Brighton, U.K., Sep. 2009.

[4] M. A. Zissman and E. Singer, "Automatic language identification of telephone speech messages using phoneme recognition and N-gram modeling", *in Proc. ICASSP*, vol. 1, pp. 305–308, Adelaide, Australia, 1994.

[5] P. A. Torres-Carrasquillo, E. Singer, W. M. Campbell, T. Gleason, A. McCree, D. A. Reynolds, F. Richardson, W. Shen, and D. E. Sturim, "The MITLL NIST LRE 2007 language recognition system", *in Proc. Interspeech*, pp. 719–722, Brisbane, Australia, Sep. 2008.

[6] P. Matějka, L. Burget, O. Glembek, P. Schwarz, V. Hubeika, M. Fapšo, T. Mikolov, O. Plchot, and J. Černocký, "BUT language recognition system for NIST 2007 evaluations", *in Proc. Interspeech*, pp. 739–742, Brisbane, Australia, Sep. 2008.

[7] J. L. Gauvain, A. Messaoudi, and H. Schwenk, "Language recognition using phone lattices", in S. H. Kim and D. H. Youn, editors, *Proc. ICSLP*, Jeju, Korea, Oct. 2004.

[8] M. F. BenZeghiba, J.-L. Gauvain, and L. Lamel, "Context-dependent phone models and models adaptation for phonotactic language recognition",

in Proc. Interspeech, pp. 313–316, Brisbane, Australia, Sep. 2008.

[9] A. O. Hatch, B. Peskin, and A. Stolcke, "Improved phonetic speaker recognition using lattice decoding", *in Proc. ICASSP*, vol. 1, pp. 169–172, Philadelphia, Mar. 2005.

[10] Q. Zhu, A. Stolcke, B. Y. Chen, and N. Morgan, "Using MLP features in SRI's conversational speech recognition system", *in Proc. Interspeech*, pp. 2141–2144, Lisbon, Sep. 2005.

[11] A. Stolcke, F. Grézl, M.-Y. Hwang, X. Lei, N. Morgan, and D. Vergyri, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons", *in Proc. ICASSP*, vol. 1, pp. 321–324, Toulouse, May 2006.

[12] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, "MLLR transforms as features in speaker recognition", *in Proc. Interspeech*, pp. 2425–2428, Lisbon, Sep. 2005.

[13] E. Shriberg, L. Ferrer, S. Kajarekar, N. Scheffer, A. Stolcke, and M. Akbacak, "Detecting nonnative speech using speaker recognition approaches", *in Proceedings IEEE Odyssey-08 Speaker and Language Recognition Workshop*, Stellenbosch, South Africa, Jan. 2008.

[14] D. Matrouf, N. Scheffer, B. Fauve, and J.-F. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification", *in Proc. Interspeech*, pp. 1242–1245, Antwerp, Aug. 2007.

[15] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, "Comparison of scoring methods used in speaker recognition with joint factor analysis", *in Proc. ICASSP*, pp. 4057–4060, Taipei, Apr. 2009.

[16] N. Brümmer, "Focal multi-class—tools for evaluation, calibration and fusion of, and decision-making with, multi-class statistical pattern recognition scores", http://sites.google.com/site/nikobrummer/focalmulticlass, June 2007.

[17] A. Stolcke, S. Kajarekar, and L. Ferrer, "Nonparametric feature normalization for SVM-based speaker verification", *in Proc. ICASSP*, pp. 1577–1580, Las Vegas, Apr. 2008.

[18] A. Stolcke, L. Ferrer, and S. Kajarekar, "Improvements in MLLR-transform-based speaker recognition", *in Proc. IEEE Odyssey 2006 Speaker and Language Recognition Workshop*, pp. 1–6, San Juan, Puerto Rico, June 2006.

[19] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition", *in Proc. ICASSP*, vol. 1, pp. 629–632, Philadelphia, Mar. 2005.

[20] W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek, "Phonetic speaker recognition with support vector machines", in S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pp. 1377–1384, Cambridge, MA, 2004. MIT Press.