Detecting Nonnative Speech Using Speaker Recognition Approaches

Elizabeth Shriberg[1], *Luciana Ferrer*[2], *Sachin Kajarekar*[1]
Nicolas Scheffer[1], *Andreas Stolcke*[1], *Murat Akbacak*[1]

[1]Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, U.S.A.
[2]Department of Electrical Engineering, Stanford University, Stanford, CA, U.S.A.
{ees,lferrer,sachin,scheffer,stolcke,murat}@speech.sri.com

Abstract

Detecting whether a talker is speaking his native language is useful for speaker recognition, speech recognition, and intelligence applications. We study the problem of detecting nonnative speakers of American English, using two standard speech corpora. We apply approaches effective in speaker verification to this task, including systems based on MLLR, phone N-gram, prosodic, and word N-gram features. Results show equal error rates between 12% and 20%, depending on the system, test data, and choice of training data. Asymmetries in performance are most likely explained by differences in native language distributions in the corpora. Model combination yields substantial improvements over individual models, with the best result being around 8.6% EER. While phone N-grams are widely used in related tasks (e.g., language and dialect ID), we find that it is the least effective model in combination; MLLR, prosody, and word N-gram systems play stronger roles. Overall, results suggest that individual systems and system combinations found useful for speaker ID also offer promise for nonnativeness detection, and that further efforts are warranted in this area.

1. Introduction

Automatic detection of nonnative speech is both theoretically interesting and practically important. Theoretical interest derives from the question of how speakers' native (first) language (L1) influences a second language in which they are not native (L2). For the present paper, however, our main motivation comes from prior work in speech and speaker recognition. It is well-known that nonnative speakers pose problems for speech recognition, typically degrading performance because of mismatch to the (largely native) training speakers. Similarly, we found that nonnatives adversely affect speaker verification performance, as a result of systematic shifts in score distributions relative to native speakers. In both cases, knowing the nativeness status of a test speaker would enable adaptation techniques to mitigate the mismatch between training and test data. Finally, identifying nonnative speech has utility in its own right in many scenarios, such as automated customer service and intelligence applications.

Previous work most closely related to nonnativeness detection has been done under the label of "accent identification", usually defined as the task of classifying a speech sample as belonging to one of several (typically 3 to 6) native and/or nonnative accents. A special case involving native varieties of a language also goes under the name of "dialect identification". Features used for these tasks include cepstral vectors (typically modeled by Gaussian mixture models or hidden Markov models), phone strings (typically modeled by language models), and a variety of prosodic features [1, 2, 3, 4, 5, 6]. The work of Schultz *et al.* [7] comes closest to our notion of nonnativeness detection; however, their work examines only a single L1 (Japanese), rather than an open set.

The work presented here differs in both task and method from previous work. First, we look at the binary classification of native/nonnative identification, as opposed to multiclass "accent" or "dialect" identification. Also, prior work has typically relied on small, home-grown databases. In this study we examine two large speech corpora that are widely used for speech and speaker recognition research. (As noted below, this also meant having to hand-label a substantial subset of data ourselves.) We focus on features and models that were found to be effective in speaker recognition, and that are designed to capture a range of phenomena from low-level, acoustic to high-level, stylistic features. The models used can be roughly mapped to specific linguistic phenomena that should be helpful in characterizing nonnative speech, including pronunciation, prosody, vocabulary, and grammar.

2. Nonnativeness Detection

We examine data from people speaking (or, in some cases, attempting to speak) American English, primarily because we had data and suitable models available for that language from prior work in speech and speaker recognition. In principle the study could be carried out for other languages or dialects. We define nonnative speakers for this study as follows:

- A speaker whose first language is any dialect of American English, and who is speaking in that dialect in the conversations studied, is considered to be a native speaker.

- Talkers whose first language is not English, are considered nonnative speakers when speaking English.

Two further caveats should be noted. To focus the task, talkers whose first language is a non-American dialect of English (e.g., British, Australian, Indian) are removed from consideration, since they typically are not trying to modify their accent when speaking English (i.e., they are speaking their native dialect, posing no significant difficulties). Second, bilingual and multilingual speakers who reported American English as one of their native languages were also removed from the study.

Nonnativeness detection poses particular challenges beyond those of speaker or dialect ID. Nonnative speech is affected by the speaker's L1, the relationship of L1 to L2, and the speaker's proficiency in L2. In this study, because of limited data resources, we group all nonnative speakers together, regardless of L1 or proficiency; better performance could probably be achieved by conditioning on these factors.

3. Data and Experimental Protocol

Data for our study is drawn from two large corpora: the LDC Fisher (phase 1) corpus [8], and the NIST 2006 speaker recognition evaluation (SRE) set, which in turn is a subset of the Mixer collection [9]. For experiments we selected subsets of these corpora, identified as *SRI-FSH* and *SRI-SRE06*, respectively.

3.1. SRI-FSH

The LDC Fisher Phase 1 corpus contains a large number of speakers, a small, but significant portion of which are nonnative speakers. All conversations are conducted in English. All speakers who did not declare English as their native language (according to LDC's corpus documentation) are included in our database as nonnatives. For natives, we chose a random subset of American English native speakers about equal in size to the nonnative set. The final selection comprised 749 nonnative speakers and 741 native speakers. For each speaker an average of 1.9 conversation sides were available, resulting in 1,512 conversation sides from native speakers and 1,503 conversation sides from nonnative speakers. Fisher conversations are 10 minutes in length, yielding about 5 minutes of speech per conversation side on average. Note that Fisher speaker classification relies on self-reported L1s, according to LDC's corpus documentation. (This means that it is likely that some speakers with very high English proficiency will be labeled as nonnatives.) Table 1 shows the distribution of L1 languages for nonnative speakers.

Table 1: Distribution of L1 for the SRI-FSH and SRI-SRE06 databases. The table includes all L1s found in SRI-SRE06 and all the frequent L1s in SRI-FSH (66% of all conversations). L1 labels are as found in LDC documentation and in some cases (e.g., Chinese and Mandarin) are not orthogonal classes.

L1	SRI-FSH	SRI-SRE06
Spanish	17.90	0
Chinese	10.65	77.78
Russian	8.05	9.82
Hindi	8.05	0.48
Mandarin	3.99	4.99
German	3.99	0
Cantonese	3.39	0
Korean	3.33	0.48
French	3.06	0
Arabic	2.59	0.64
Urdu	1.06	0.48
Thai	0.20	2.09
Other	0	3.22

3.2. SRI-SRE06

The NIST SRE 2006 data set contains a mix of languages, albeit with English as the dominant language. Certain speakers occur in multiple conversations, and some of these conversations are conducted in non-English languages. This suggests that many of the English conversations involve nonnative speakers. For this study, we listened to a total of 2590 conversation sides involving 595 distinct speakers, and recorded nativeness judgments; 280 speakers (1,604 conversations sides) were labeled as native American English, 315 (986 conversation sides) as nonnative. SRE 2006 conversations are 5 minutes long, giving about 2.5 minutes of speech per side on average. Labeling for this database was done by listening to a random subset of segments from each speaker and assigning the labels based on the perceived accent and fluency. The resulting labels, then, might consider an actual nonnative as native if the speaker had no detectable accent or fluency difficulties, but it is acceptable in our view that the models mimic the decisions of a human listener.

Table 1 shows the distribution of native languages (L1) for a subset of the SRI-SRE06 nonnative speakers for which L1 could be reasonably inferred from the corpus documentation. Since the actual L1 information for each speaker was not available, we looked at the range of all languages spoken by a given speaker throughout the corpus. If a speaker had conversations in both English and another language L, and the speaker had been labeled a nonnative, we assumed that the speaker's L1

is L. This allowed us to infer L1 for 621 conversation sides. Clearly, the distribution of L1 for SRI-SRE06 is very different from that of the Fisher data. In particular, SRI-SRE06 data contains mostly Chinese speakers, while SRI-FSH shows a much broader and more balanced selection of L1s.

3.3. Experimental protocol

We evaluate nonnativeness detection systems using one of two training and test protocols:

- **Matched**: Training and test data comes from the same corpus. To make efficient use of all available data, we employ 10-fold cross-validation. Speakers are randomly assigned to ten roughly equal partitions. Each partition in turn is used as the test set, while the other nine partitions are used to train models. Overall results are computed by averaging the outcomes over the ten test partitions.
- **Mismatched**: Training and test data come from different corpora: we train on SRI-FSH and test on SRI-SRE06, and vice versa. We use this protocol to simulate a more realistic application dealing with an unknown population of speakers, and test the generalization of our models under such conditions.

Performance is measured in terms of equal error rate (EER), i.e., the operating point at which false nativeness and false nonnativeness decisions occur with equal frequency.

4. Models and Systems

Since the task in the precise form defined here is novel, we investigated modeling approaches inspired by two extensively studied, related tasks:

- Language recognition/identification (LID): a nonnative English speaker can be seen as a special subtype of language.
- Speaker verification/identification (SID): nativeness or nonnativeness may be viewed as a generalization of speaker identity, and similar binary classification techniques can be applied.

The approaches studied here were ultimately adapted from a subset of systems found in SRI's speaker recognition system [10], plus one system jointly developed by SRI and ICSI [11]. Apart from expediency, our choice was motivated by two observations. First, speaker recognition, broadly speaking, employs a superset of features and modeling techniques found in language recognition. Second, the systems studied here are chosen to cover a range of features and time scales, giving good coverage of the documented linguistic manifestations of nonnativeness, as explained further below. In addition to the SVM-based models described below we also experimented with cepstral Gaussian mixtures (GMMs) as nonnativeness classifiers, but found them to be less competitive than our best single system. Since we include other models based on cepstral features, we did not include cepstral GMMs in this study.

4.1. Phone N-gram language model

As a baseline, we tested a popular approach to language recognition, a phone-recognition-based language model (PRLM) [12]. An open-loop phone recognizer is run on each conversation side and the 1-best phone strings are recorded. The recognizer is trained on English data from the Switchboard corpus and uses 45 phone labels based on the ARPABET set. Trigram language models (LMs) are trained based on phone transcripts for native and nonnative training speakers, respectively. At test time, a score is computed that is the length-normalized log likelihood ratio of the two LMs given the phone sequence extracted from test data.

Note that we used the PRLM as a baseline only; a much improved model based on phone N-grams is described below. Because of resource limitations, as well as the relatively poor performance of the PRLM on our data, we did not investigate approaches based on parallel phone recognition for multiple languages (PPLRM) in this study.

4.2. MLLR transform SVM

This model uses the speaker maximum likelihood linear regression (MLLR) adaptation transforms employed by a large-vocabulary automatic speech recognizer (ASR) as features [13, 14]. The system computes two 39×40-dimensional affine transforms for the Gaussian means of a male and female ASR model, respectively. Two versions of the system are studied. The first version relies only on phone-loop recognition and clusters the Gaussians into two phone classes, yielding $2 \times 2 \times 39 \times 40 = 6240$ features per conversation side. The second, more elaborate version uses full word recognition hypotheses to compute MLLR transforms, and clusters Gaussians into eight phone clusters; it thus yields feature vectors of $8 \times 2 \times 39 \times 40 = 24960$ components. The features are rank-normalized to the unit interval along each dimension, and a support vector machine (SVM) with linear kernel is trained to discriminate between native and nonnative conversation sides. The nonnativeness score is the signed distance of the test feature vector from the decision hyperplane.

4.3. Phone N-gram SVM

This model is a phone-sequence classifier based on support vector machines rather than language models [15]. We adapted a variant that has shown good performance

in speaker recognition [11]. As with the PRLM, an open-loop phone recognizer is run on each conversation side, but generating phone lattices rather than just 1-best hypotheses. We then extract expected frequencies for unigrams, bigrams, and trigrams (i.e., N-grams are weighted according to their posterior probability of occurrence in the lattice). The 8,483 most frequent N-grams are retained, giving the dimensionality of the feature vector. The N-gram frequencies are then scaled by the inverse square root of the overall N-gram probabilities. When combined with a linear SVM kernel, this gives the log likelihood ratio kernel of [15]. Prior studies [11] have shown that both the switch from LM to SVM modeling and the generalization from 1-best to lattice recognition give substantial gains in speaker recognition, so we expected the phone N-gram SVM to be superior to PRLM for nonnativeness detection as well.

4.4. Prosodic sequence SVM

This system models syllable-based prosodic features [16]. Features are based on estimated F0, energy, and duration information extracted over syllables inferred via automatic syllabification of ASR output. Pauses, which can be particularly useful for low-proficiency speakers, are also modeled as special tokens in the sequences. Prosodic feature sequences are transformed into fixed-length vectors by a particular implementation of the Fisher score [17]. Features modeling sequences of one, two, and three syllables are used. The resulting feature vector, of dimension 38,314, is first rank-normalized (as in the MLLR-SVM system) and modeled by linear kernel SVM.

4.5. Word N-gram SVM

Models of how frequently speakers use certain words or phrases (idiolect) have been proposed by [18] and, although poor models by themselves, were found to improve speaker recognition systems in combination with other knowledge sources. Here we use a version of such a model based on SVMs, which gave better results than language-model-based approaches [19]. Per-conversation side frequencies for 126k unigrams, bigrams, and trigrams are extracted from the 1-best hypotheses of the ASR system. The (very sparse) feature vectors are rank-normalized along each dimension and modeled by linear kernel SVMs, similar to the MLLR, phonetic, and prosodic systems.

We can roughly associate the above systems with different aspects of nonnative language proficiency. The MLLR and phone N-gram systems model acoustic observations and can potentially capture differences in pronunciations. The prosodic system models patterns of pitch, energy, duration, and pausing. It can thus capture aspects both of fluency and of suprasegmental characteristics of language that are among the most difficult for L2 speakers to master. Finally, the word N-gram SVM models differences in vocabulary, idiom, and grammar usage.

4.6. ASR system

The 8-class MLLR system, the prosodic, and the word N-gram system all rely on word recognition. Our ASR system was a 2-pass conversational telephone recognizer dating from 2003 [20]. No Fisher data was used to train the system, and all acoustic training data came from native American English speakers. Consequently, the average word error rate (WER) for nonnative speakers was much higher than for native speakers (40% versus 27%, as measured against Fisher quick transcriptions provided by LDC). Also, because of the larger temporal distance between training and test data (and attendant bigger mismatch in language and acoustic models) we expect higher error rates on SRE06 than on Fisher; however, we could not verify this in the absence of transcriptions for SRE06 data.

5. Results and Discussion

We look first at the performance of individual systems for the four different conditions comprising the crossing of (1) matched vs. mismatched train and test data, and (2) the choice of test set (SRI-FSH vs. SRI-SRE06). As noted earlier, the two corpora differ in terms of L1 distributions, L2 proficiency, and ASR performance. It is thus interesting to look at results when swapping train and test data, and for both matched and mismatched cases, to understand how performance depends on these factors. Subsequently, in Section 5.2, we look at the results of system combination for the case of the most realistic scenario (mismatched train/test) and testing on the most current data set (SRI-SRE06). We ask whether combination aids performance, and if so, which systems offer the most complementary information.

5.1. Individual systems

Figure 1 shows results for the five individual systems, under the four different train/test conditions. For ease of readability, lines connect results for each condition.

A number of observations can be noted. First, the effects of test corpus and train/test mismatch are remarkably regular across systems plots running almost perfectly parallel. One exception is the prosodic system for the mismatched condition when testing on SRI-SRE06, which is better than expected based on the shape of the trends for the other three conditions. Although further study is necessary, we suspect one reason is that nonnative speakers in SRI-SRE06 often struggled to get words out in English; this may have allowed the prosodic model to utilize features based on pausing patterns that tend to be robust to the higher rate of ASR errors in that corpus.

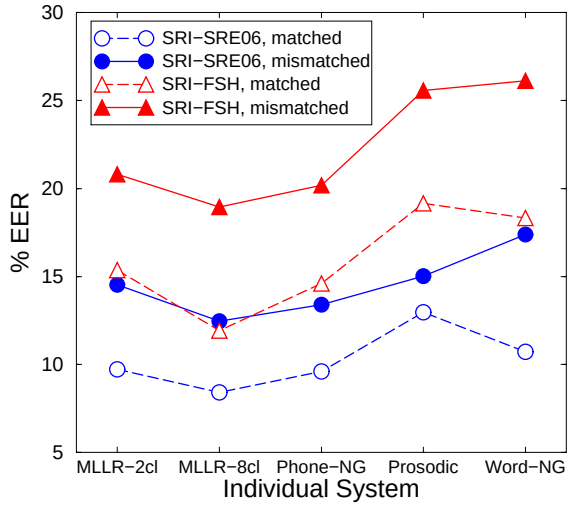

Figure 1: Nonnativeness detection results (in % EER) by individual system ("cl" = class, "NG" = N-gram), test corpus, and matched vs. mismatched training. Matched conditions use 10-fold cross-validation on the same corpus. Mismatched conditions train on one corpus and test on the other. The PRLM baseline system was tested on SRI-SRE06 only, with mismatched training, giving 17.3% EER (not shown in figure).

Second, SRI-SRE06 is the easier test corpus, as can be easily seen from comparing the two cross-validation results in Figure 1. This is most likely due to two factors: the fairly homogeneous nature of L1 for nonnatives in that corpus (nearly 80% Chinese), and the estimated lower proficiency of nonnatives in L2 in that data. Both factors should facilitate native/nonnative discrimination. Note, however, that as mentioned earlier, ASR performance is probably worse on SRI-SRE06 than on SRI-FSH. It is not clear at this point whether lower ASR performance hurts or helps nonnativeness detection (it might help if word error patterns are consistent and informative of nativeness status, similar to the way one of our duration-based speaker ID models benefited from higher word error rates [19]). The answer may depend on the model used, and is of obvious importance to applications. One could investigate the question by running poorer ASR systems on matched results, and observing the effect by model for this task.

Third, as anticipated, train/test corpus mismatch causes large degradations in performance. The effect of mismatch is larger when testing on SRI-FSH than when testing on SRI-SRE06. This is the expected direction when considering the effects of L1 distribution and L2 proficiency, since both factors make training on SRI-SRE06 a poor choice for testing on the more L1-variable and higher L2-proficient SRI-FSH data. But in terms of the effect of ASR, FSH is the better-recognized corpus. As noted above, it is not yet clear how ASR affects results here, and further study is needed.

Fourth, while the pattern of individual system results is fairly stable over conditions, in absolute terms the effect of test corpus and train/test mismatch is larger than that of the individual systems. This highlights the importance of calibrating for such conditions when evaluating future research on nativeness detection.

Finally, results in Figure 1 follow the general pattern from speaker ID in that, when used alone, systems based on acoustic features perform better than those based on longer-range features such as prosody or words. A difference between the two tasks, however, is that for nonnativeness detection, the prosodic and word N-gram systems are much closer in performance to the level of the acoustic systems. As expected, the PRLM baseline system is worse than the phone N-gram SVM (17.3% versus 13.4% EER training on SRI-FSH and testing on SRI-SRE06).

5.2. Combined systems

So far we have only considered the performance of individual models. As in speaker recognition, the combination of multiple systems can yield improvements because of the complementary information captured by the various systems. For this experiment we focus on the most realistic and current condition, namely testing on SRI-SRE06, using mismatched training (SRI-FSH). We also remove the 2-class MLLR system from consideration, since it performs less well than the 8-class system and speaker recognition studies have shown no gain from combining the two systems [14].

The combiner produces a new nativeness score from the scores output by the individual systems and consists of a neural network with a single layer (perceptron). The combiner is trained with 2-fold cross-validation on SRI-SRE06 (half of it is used to train the combiner, the other half is used for testing). The final score is an average over the folds. Table 2 shows results.

As shown, system combination yields improvements over the best individual system for this condition. But, more important, contributions to system combination follow a different ordering than do the individual results. The 8-class MLLR and phone N-gram systems perform well individually, and one of them is necessary to achieve good performance. The phone N-gram, however, is actually the least helpful system to combine with the MLLR system in the two-way combination. Just as in studies in speaker ID [10] the best N-way system includes all systems used in the best $(N-1)$-way system. System contributions can thus be ordered from most to least useful in combination, as follows:

$$\text{MLLR} > \text{prosody} > \text{word N-gram} > \text{phone N-gram}$$

However, all systems contribute complementary informa-

Table 2: System combination results using the mismatched condition, testing on SRI-SRE06. The combiner is trained with 2-fold cross-validation on SRI-SRE06. Results are given in terms of EER %. The corresponding best single-system condition is provided for comparison.

Best single system	
mllr,8class	12.47
2 way combinations	
mllr,8class+prosody	10.35
mllr,8class+wordlm	11.03
prosody+phone-ngram	11.10
mllr,8class+phone-ngram	11.28
wordlm+phone-ngram	11.35
prosody+wordlm	11.66
3 way combinations	
mllr,8class+prosody+wordlm	9.29
mllr,8class+prosody+phone-ngram	9.54
prosody+wordlm+phone-ngram	10.16
mllr,8class+wordlm+phone-ngram	10.22
4 way combination	
mllr,8class+prosody+wordlm+phone-ngram	8.60

tion: the final four-way combination yields an EER of 8.60%, a 31% relative reduction in error from the best single system.

6. Conclusions and Future Work

We have investigated the problem of nonnativeness detection in speech, using two large data sets from standard speech corpora. Nativeness labels were based on self-reported first language information for one data set, and on the results of human listening for the other. Several models that have been successfully employed in past work for automatic speaker recognition were adapted for the task of detecting nonnative speech. Interestingly, relative performance results of the models for nonnativeness detection followed the same general pattern as results for speaker verification, but with a smaller gap between acoustic and stylistic systems in the case of nonnativeness detection. Results from model combination also showed a substantial improvement over results for any individual model.

We also found, however, that in absolute terms, the largest effect on results was the degree of mismatch between training and test corpora. As might be expected, performance is particularly degraded when training data is more homogeneous than is test data, for nonnativeness-related factors. We hypothesize that important factors include the distribution of particular L1s, nonnatives' L2 proficiencies, and ASR accuracy. ASR performance remains a poorly understood variable in our study, in part because reference transcripts were not available for the SRE06 data. Further work is needed to determine the effect of ASR accuracy on individual systems, holding other variables (especially train/test mismatch) constant. It should be the case that better robustness of the ASR system to nonnative speech will lead to overall improvements in nonnativeness recognition.

A long-term goal, given sufficient data, is to investigate explicit modeling of specific L1/L2 combinations. In such work it should also be important to condition on estimates of proficiency in L2. In data collection efforts such as those by NIST, in which subject recruitment tends to involve particular communities of speakers, L1 and proficiency in L2 may be correlated. As stated earlier, based on the hand-labeling of SRI-SRE06 we suspect that the SRI-SRE06 data has a higher rate of lower-proficiency nonnative speakers than does Fisher, although this claim requires more study. If it is true, then the L1 skew of SRI-SRE06 is confounded with lower proficiency, making it difficult to determine which factor affects performance. Our guess is that mismatches in the direction of (1) higher L1 skew in train than test, and (2) lower L2 proficiency in train than test (making training samples more discriminative than they will be in testing), are particularly detrimental to performance.

7. Acknowledgments

We thank two anonymous reviewers and Fred Goodman for helpful comments. This research was supported by NSF award IIS-0544682, by a development contract with Sandia National Laboratories, and by an SRI IR&D project. The views herein are those of the authors and do not reflect the views of the funding agencies.

8. References

[1] J. H. L. Hansen and L. M. Arslan, "Foreign accent classification using source generator based prosodic features", *in Proc. ICASSP*, vol. 1, pp. 836–839, Detroit, May 1995.

[2] L. M. Arslan and J. H. L. Hansen, "Language accent classification in American English", *Speech Communication*, vol. 18, pp. 353–367, July 1996.

[3] C. Teixeira, I. Trancoso, and A. Serralheiro, "Accent identification", in H. T. Bunnell and W. Idsardi, editors, *Proc. ICSLP*, vol. 3, pp. 1784–1787, Philadelphia, Oct. 1996.

[4] M. A. Zissman, T. P. Gleason, D. M. Rekart, and B. L. Losiewicz, "Automatic dialect identification of extemporaneous conversational, Latin American Spanish speech", *in Proc. ICASSP*, vol. 2, pp. 777–780, Atlanta, May 1996.

[5] M. Lincoln, S. Cox, and S. Ringland, "A comparison of two unsupervised approaches to accent iden-

tification", in R. H. Mannell and J. Robert-Ribes, editors, *Proc. ICSLP*, vol. 1, pp. 109–112, Sydney, Dec. 1998. Australian Speech Science and Technology Association.

[6] P. Fung and W. K. Liu, "Fast accent identification and accented speech recognition", *in Proc. ICASSP*, vol. 1, pp. 221–224, Phoenix, AZ, Mar. 1999.

[7] T. Schultz, Q. Jin, K. Laskowski, A. Tribble, and A. Waibel, "Improvements in non-verbal cue identification using multilingual phone strings", *in Proceedings of the ACL-02 Workshop on Speech-to-Speech Translation: Algorithms and Systems*, vol. 7, pp. 101–108, 2002.

[8] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: a resource for the next generations of speech-to-text", *in Proceedings 4th International Conference on Language Resources and Evaluation*, pp. 69–71, Lisbon, May 2004.

[9] A. Martin, D. Miller, M. Przybocki, J. Campbell, and H. Nakasone, "Conversational telephone speech corpus collection for the NIST speaker recognition evaluation 2004", *in Proceedings 4th International Conference on Language Resources and Evaluation*, pp. 587–590, Lisbon, May 2004.

[10] L. Ferrer, E. Shriberg, S. S. Kajarekar, A. Stolcke, K. Sönmez, A. Venkataraman, and H. Bratt, "The contribution of cepstral and stylistic features to SRI's 2005 NIST speaker recognition evaluation system", *in Proc. ICASSP*, vol. 1, pp. 101–104, Toulouse, May 2006.

[11] 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.

[12] 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.

[13] 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.

[14] A. Stolcke, S. S. Kajarekar, L. Ferrer, and E. Shriberg, "Speaker recognition with session variability normalization based on MLLR adaptation transforms", *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, pp. 1987–1998, Sep. 2007.

[15] 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.

[16] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic feature sequences for speaker recognition", *Speech Communication*, vol. 46, pp. 455–472, 2005, Special Issue on Quantitative Prosody Modelling for Natural Speech Description and Generation.

[17] L. Ferrer, E. Shriberg, S. Kajarekar, and K. Sonmez, "Parameterization of prosodic feature distributions for SVM modeling in speaker recognition", *in Proc. ICASSP*, vol. 4, pp. 233–236, Honolulu, Apr. 2007.

[18] G. Doddington, "Speaker recognition based on idiolectal differences between speakers", in P. Dalsgaard, B. Lindberg, H. Benner, and Z. Tan, editors, *Proc. EUROSPEECH*, pp. 2521–2524, Aalborg, Denmark, Sep. 2001.

[19] S. S. Kajarekar, L. Ferrer, E. Shriberg, K. Sonmez, A. Stolcke, A. Venkataraman, and J. Zheng, "SRI's 2004 NIST speaker recognition evaluation system", *in Proc. ICASSP*, vol. 1, pp. 173–176, Philadelphia, Mar. 2005.

[20] A. Stolcke, H. Franco, R. Gadde, M. Graciarena, K. Precoda, A. Venkataraman, D. Vergyri, W. Wang, J. Zheng, Y. Huang, B. Peskin, I. Bulyko, M. Ostendorf, and K. Kirchhoff, "Speech-to-text research at SRI-ICSI-UW", *in DARPA RT-03 Workshop*, Boston, May 2003, http://www.nist.gov/speech/tests/rt/rt2003/spring/presentations/sri+-rt03-stt.pdf.