LANGUAGE-INDEPENDENT CONSTRAINED CEPSTRAL FEATURES FOR SPEAKER RECOGNITION

Elizabeth Shriberg and Andreas Stolcke

SRI International, Menlo Park, CA, USA
{ees,stolcke}@speech.sri.com

ABSTRACT

Constrained cepstral systems, which select frames to match various linguistic “constraints” in enrollment and test, have shown significant improvements for speaker verification performance. Past work, however, relied on word recognition, making the approach language dependent (LD). We develop language-independent (LI) versions of constraints and compare results to parallel LD versions for English data on the NIST 2008 interview task. Results indicate that (1) LI versions show surprisingly little degradation from associated LD versions, (2) some LI constraints outperform their LD counterparts, (3) useful constraint types include phonetic, syllable position, prosodic, and speaking-rate regions, (4) benefits generally hold for different train/test lengths, and (5) constraints provide particular benefit in reducing false alarms. Overall, we conclude that constrained cepstral modeling can benefit speaker recognition without the need for language-dependent automatic speech recognition.

Index Terms**—**language-independent phone recognition, cepstral constraints, speaker verification.

1. INTRODUCTION

Most approaches to speaker recognition use Mel frequency cepstral coefficients (MFCCs) extracted from all regions in a signal that are deemed to contain speech. Recent work on *constrained* cepstral systems has shown that significant improvements in performance can be obtained by creating specialized speaker models that use the same MFCC features, while restricting the modeled frames to only those that match a particular, linguistically motivated “constraint”. In our prior work [1], such constraints have corresponded to a specific phone, syllable position, or location with respect to pauses. The constrained systems can then be combined at the score level with a baseline system and/or with each other. Despite “reusing” the same features as the all-frames baseline system and other constrained systems, gains from combination can be substantial, demonstrating the value of matching cepstral vectors according to their linguistic context in enrollment and test data.

Several previous studies have investigated constraining or selecting cepstral frames to enhance speaker modeling; the most successful ones have used word or phone information, thereby reducing variability associated with phonetic content. For example, the approaches in [2] and [3] condition a cepstral Gaussian mixture model (GMM) on the identities of frequent words or syllables, respectively. The methods described in [4] and [5] assign frames to broad phone classes in order to score them with class-dependent GMMs.

Our constrained cepstral modeling approach differs from these earlier approaches in several key aspects. Our goal is not to partition all frames of speech. On the contrary, the idea is to focus the modeling on only those regions that yield highly consistent or distinctive (for the speaker) features. Similarly, in this approach constraints are not orthogonal; the same frame can be used as a member of different constraints that are either partially overlapped or nested with respect to the speech regions covered. In addition, our frame selection criteria go beyond phone and word information to include prosodic and speaking rate phenomena.

To support such detailed linguistic modeling, our previous work [1] relied on word recognition, making it language dependent (LD) and therefore of restricted value for some applications. In this work we develop language-independent (LI) versions of constraints using multilingual phone recognition and automatic syllabification, and compare results to the LD versions for English data. We also move from eigenchannel-only compensation in prior work to an updated framework based on joint factor analysis (JFA). Finally, we evaluate the method using the most recent NIST speaker recognition evaluation (SRE) cost metric, which focuses on an operating point with very few false alarms.

2. METHOD

2.1. Constrained Cepstral Modeling

The constrained cepstral speaker modeling approach is illustrated in Figure 1. The signal is converted to a stream of MFCC cepstral frames, as usual, followed by a selection of frame subsets based on

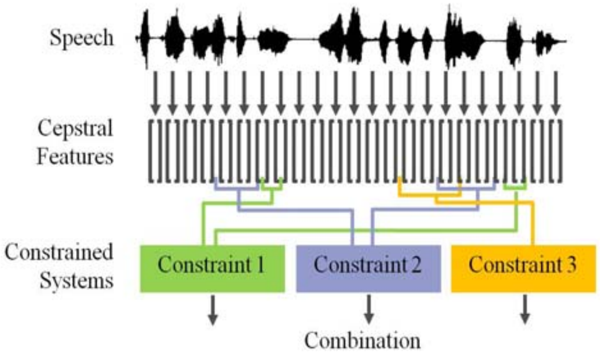

Figure 1: Schematic representation of the approach.

linguistic criteria (explained in more detail below). Currently, our constraint systems are combined at the score level. Each constraint is associated with its own GMM-UBM (universal background model), yielding a verification score. The scores are combined with a baseline GMM that uses all frames, and/or with each other, in general using linear logistic regression. In the present study, given the lack of a suitable training set, we combine each constraint separately with the baseline score, using equal weights.

All constrained systems, as well as the baseline system, use the same front end, based on MFCCs from 24 Mel filters covering 300-3300 Hz. Twenty coefficients were computed (C0-C19). Based on these 20 values the first- and second-order derivatives are calculated, resulting in a 60-dimensional feature vector. This feature vector is mean and variance normalized over the session.

Within-speaker and across-speaker variability was modeled in the JFA framework [6]. The background model consisted of 1024 Gaussians; 300 eigenchannels and 600 eigenspeakers were included in the model (in [1] we had performed only eigenchannel-based modeling). As an expedient, these three model components were estimated for the baseline system, on all speech frames, and then reused for all constrained models. (Prior work [1] found that constrained models or their combination may work better if UBM and JFA matrices are trained for each constraint separately, but in the current effort this was not fully explored.) All model scores are normalized with ZT-norm.

2.2. Word Recognition

Language-dependent constraints were based on the output of a fast and simplified version of SRI's conversational telephone speech recognition system, limited to two decoding and various rescoring passes [7]. The word error rate (WER) on native and nonnative speakers on transcribed parts of the Mixer corpus was 23.0% and 36.1%, respectively. On SRE 2006 microphone sessions we measured a WER of 28.8%.

2.3. Multilingual Phone Recognition

Language-independent constraints that made use of phonetic or syllable information were extracted based on the output of a multilingual phone recognizer. We used a 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). Training data comprised about 232 hours of American English (about equal parts native and nonnative speakers), 103 hours of Mandarin, 19 hours of Spanish, and 17 hours of Egyptian Colloquial Arabic, all from conversational telephone speech corpora. A phone trigram served as the phonotactic language model. The phone recognition error rate on English data was roughly 32% for native speakers and about 40% for nonnative speakers. Note that it is useful for the phone recognizer to have substantial error rates since our goal is to understand loss associated with errorful phone-based recognition on speaker recognition performance.

2.4. Language-independent Syllabification

Since many constraints found to be useful as LD versions in prior work [1] make use of syllable information, we needed a syllabification approach that did not rely on word unit information. To compare results directly between LI and LD versions in this study, we developed an automatic approach based only on phone (not word) information, which was applied both to phones from word recognition (for LD versions) and to phones from multilingual phone recognition (for LI versions). The automatic syllabification uses the maximum onset principle with a list of possible nuclei. Because syllabification intentionally does not reference word or phonotactic information, possible nuclei are restricted to sounds used most often as nuclei rather than onsets or codas. Syllables without vowels in dictionary pronunciations (e.g., words such as “hm”, which contain a syllabic nasal or resonant) are therefore sacrificed. The maximum onset approach also groups long strings of consonants in the LI versions into a single (unpronounceable) onset, and generally produces minimal coda material compared with canonical word-level pronunciations.

2.5. Constraints

We explored a large number of different constraints based on regions defined by phones, phone groups, syllable position, pause context, pitch and energy-based regions, speaking rate regions, and regions defined by more than one of these factors. Due to space limitations, we focus here on only a subset of ten constraints, chosen to represent a range of different feature types. Within each feature type we have chosen fairly well performing constraints based on the results from the LD approach.

Phone-based features are represented here by three constraints: **low vowels**, **non-low front vowels**, and **non-low back vowels**. Two constraints refer to specific phones, but extract frames from all syllables containing a specific phone in any part of the syllable. These include **syllables containing a nasal phone** ([m], [ng], or [n]), and **syllables containing an [r] or [l]**. Syllable position is represented here by two constraints: **syllable nuclei** and **syllable onsets**. (We note that syllable codas here perform rather weakly relative to their performance in [1], as a consequence of the maximum-onset-based syllabification described earlier, which reduces the content available in coda position.) For pause context we include a single constraint: automatically derived **syllables that directly precede a pause** of at least 60 milliseconds. We include a pitch feature: frames associated with a **positive pitch slope**, after a regularization and fitting of pitch to straight-line approximations, and a **speaking rate** feature, corresponding to regions of *slower* speech for that particular talker, given both the talker's speaking rate and the inherent durations of different phones.

2.6. Data

Since LD constraints had worked particularly well on the interview portion of the SRE08 evaluation data (in part due to the availability of longer sessions), we focused our study of LI constraints on that condition. A set of 82 interview speakers was held out from the NIST SRE08 (original and follow-up) data for training purposes, and a test set was created using the remaining SRE08 data.

For each original condition from SRE08 an extended set was created by pairing every available model against every available test sample (except when the model and the test sample used data from the same original recording session). No additional models were created and only samples originally used for testing were used for testing in the extended set. The test sets thus created contained both *short* (3-minute) and *long* (8-minute) speech samples (the length on long sessions was limited to 8 minutes to match the SRE10 evaluation condition). Trial counts are given in Table 1.

Table 1: Evaluation trials by condition.

Train-test condition	Target trials	Impostor trials	Total
Short-short	33743	1108882	1142625
Short-long	10234	336437	346671
Long-short	32248	1054592	1086840
Long-long	9774	319956	329730

The background GMM was trained on SRE04 data. The eigenchannel matrix was trained on SRE04 and SRE05 altmic data; eigenspeaker training additionally included telephone data from SRE05 and Switchboard. Score normalization made use of SRE04, SRE05, and SRE06 telephone and altmic data.

3. RESULTS

3.1. Constraint Sparseness and LD/LI Overlap

A first set of results helpful in interpreting the performance of different constraints with respect to each other and with respect to comparing LD and LI versions is shown in Table 2. For each constraint we list the percentage of segmented frames (after nonspeech removal) matching the constraint, for both LD and LI versions, as well as the overlap in actual frames selected between the LD and LI versions. Because LI versions sometimes outperform LD versions for speaker verification (see Figure 3 and discussion), we compute the frame overlap percentage relative to both LD and LI frames.

Table 2: Frame percentages selected and overlaps between LD and LI constraint regions

Constraint	LD % frames	LI % frames	LI/LD overlap	LD/LI overlap
Low vowel	32.9	28.2	55.4	64.5
Non-low front vowel	8.9	10.4	52.6	40.6
Non-low back vowel	4.3	10.3	55.9	23.5
Syllable with nasal	15.7	16.2	52.6	39.8
Syllable with r or l	13.2	11.7	47.3	53.4
Syllable nucleus	30.3	35.0	71.5	61.8
Syllable onset	21.2	21.9	58.8	58.0
Prepause syllable	23.2	26.6	64.5	56.4
Frame in pitch rise	17.8	18.5	83.6	79.5
Slower speaking rate	17.2	19.5	57.8	51.2

3.2. Verification Performance in Long-Long Condition

Figure 2 shows results for LD versus LI versions of the same constraint. Results are for the condition using long data in both training and testing. Constraints are shown on the x-axis, ordered by the relative performance gain in LD results. The y-axis is the relative performance gain from combination of the constraint with the baseline – the difference between the combined performance and the baseline alone, divided by the performance of the baseline alone. Thus, higher values indicate better constraints in terms of complementary information to the all-frames baseline. The metric being compared is the NIST decision cost function (DCF) in both the old version (oDCF; false alarms are 10 times more costly than false rejections) and the new SRE10 version (nDCF; false alarms are 1000 times more costly than false rejections).

A first observation is that even though constraint systems “reuse” the same cepstral features as used in the baseline system, they provide considerable gains (as high as over 20%) in combination. This is rather impressive given the sparse frame counts for some constraints as shown in Table 2. A second surprising finding is that despite the low rate of overlap between LD and LI frames shown in Table 2, overall, LI versions show rather minimal degradation with respect to LD versions. There are two exceptions: syllables with nasals, and syllables with [r] or [l]. In a few cases, the LI versions seem to outperform the LD versions: for the nucleus constraint, and for the slower speaking rate constraint. But generally speaking, these results suggest that there is not much loss for most constraints when moving from LD to LI phone and syllable extraction, and that this finding seems to hold across a range of different constraint types (phonetic, syllabic, prosodic). The same pattern of minimal loss from LD to LI performance holds for a number of other constraints not shown in Figure 2.

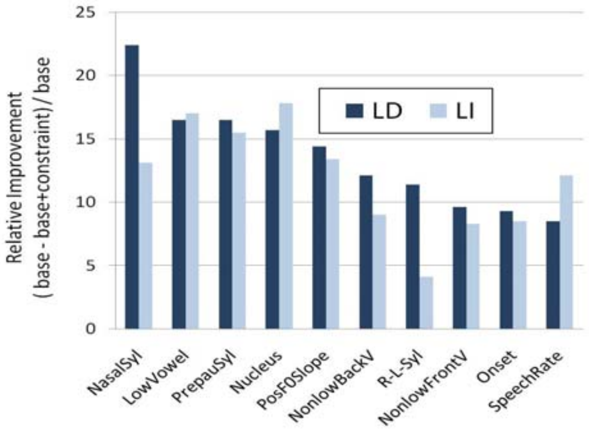

Figure 2: Comparison of LD and LI versions of selected constraints, for relative nDCF reduction from combination of constraint with baseline over baseline alone (long-long condition).

Next, we compare relative improvements from constrained modeling according to oDCF versus nDCF, which penalizes false alarms more severely. Figure 3 plots the two metrics against each other, for both LD and LI constraints.

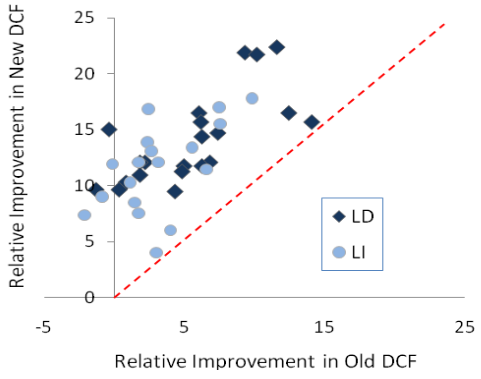

Figure 3: Relative improvement ((base-(base+constraint))/base) for oDCF versus nDCF, for LD and LI constraints (long-long condition). The dashed line indicates y=x.

As shown, the improvement from constraint combination, relative to the result for the baseline system alone, is consistently greater for the nDCF than the oDCF metric, indicating that the constrained modeling is particularly good at reducing false alarms. The interpretation is that adding more linguistic detail to the speaker model makes it less likely to be falsely matched, even by speakers with similar all-frames-based cepstral feature models.

3.3. Verification Performance by Train/Test Length

Because many of our constraints are fairly sparse, we were also interested in the effect of train and test sample length. Results by length are shown for four different constraint types in Figure 4. Nasal syllables, as seen earlier, suffer for LI versions, but this seems to be no worse for short train or test samples than for the long-long condition. Prepause syllables and pitch rises both showed fairly consistent small degradations from LD to LI versions across length conditions. For the nucleus constraint, long samples in training show a slight edge for the LI versions, while short samples in training show a slight benefit for LD versions. Further work is needed to understand these relationships. Overall however, it appears that the difference in gains between LD and LI constraints in the long-long condition is not greatly changed when moving to conditions involving short train and/or test samples.

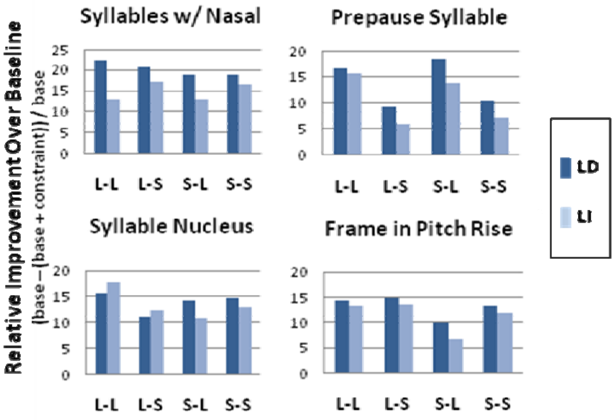

Figure 4: Comparison of LD and LI constraints, for relative nDCF reduction from combination of constraint with baseline over baseline alone, by train (first L or S) and test (second L or S) length. L=long S=short.

4. DISCUSSION AND CONCLUSIONS

Our goal was to investigate whether cepstral constraints originally developed based on language-dependent (English) word recognition could be replaced by corresponding constraints derived from language-independent phone recognition and automatic syllabification. Despite considerable differences in constraint region alignments between LD and LI versions, with few exceptions the LI versions of most constraints showed surprisingly little degradation from their LD counterparts in relative performance gain when fused with a baseline system on the NIST SRE08 interview task. This result held over a range of different constraint types (phonetic, syllable-based, pause-context-based, prosodic, speaking-rate-based). In some cases, such as syllable nuclei, the LI version outperformed the LD version, suggesting that the LI versions may avoid a contaminating effect from the language model or dictionary pronunciations that impact regions extracted using word recognition. While the minimal degradation held best for longer samples in both train and test, LI still held up well when train and/or test lengths were decreased, although the degree of this effect seems to be dependent on the constraint. A comparison of performance over different cost metrics revealed that constraints (both LD and LI) tend to perform well at low false alarm regions; this may be because they provide greater linguistic detail, making it more difficult for an impostor to match a target speaker. Overall these findings suggest that constrained cepstral modeling can benefit speaker recognition without the need for LD automatic speech recognition, thus opening up the possibility of LI speaker modeling with linguistic constraints.

The current study examined only English data, in order to compare directly between LD and LI constraints on a current NIST task. An important next step is to repeat the study for additional languages, and to compare the pattern of LI constraint robustness across languages. Another important goal for future work is to investigate constraint utility when constraints are combined with each other – rather than with only a baseline system. This was not currently feasible due to a lack of sufficient held-out data to train a constraint combiner. We also seek to better understand the relationship between constraint performance, constraint sparseness, and the overlap between LD and LI versions. In general, sparser constraints should lead to longer-term benefit in combination since given that all systems use the same features, sparser systems will by definition be less correlated with the all-frames baseline and with other constraints. On the other hand, sparse constraints may be less robust to extraction, especially for LI versions. Finally, we aim to better predict and explain which constraints show better performance for LI than for LD versions, since such knowledge could benefit not only LI performance, but performance when word recognition is available as well.

ACKNOWLEDGMENTS

We thank Luciana Ferrer, Martin Graciarena, Sachin Kajarekar, Nicolas Scheffer, Harry Bratt, and Colleen Richey. This work was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory (ARL), and by Sandia National Laboratories. All statements of fact, opinion, or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI, the U.S. Government, or Sandia.

REFERENCES

[1] T. Bocklet and E. Shriberg, “Speaker recognition using syllable-based constraints for cepstral frame selection,” in *Proc. of ICASSP*, pp. 4525–4528, Taipei, Taiwan, 2009.

[2] D. E. Sturim, D. A. Reynolds, R. B. Dunn, and T. F. Quatieri, “Speaker verification using text-constrained Gaussian mixture models,” in *Proc. of ICASSP*, vol. 1, pp. 677–680, Orlando, FL, 2002.

[3] B. Baker, R. Vogt, and S. Sridharan, “Gaussian mixture modelling of broad phonetic and syllabic events for text-independent speaker verification,” in *Proc. Eurospeech*, pp. 2429–2432, Lisbon, 2005.

[4] A. Park and T. J. Hazen, “ASR dependent techniques for speaker identification,” in *Proc. ICSLP*, pp. 1337–1340, Denver, 2002.

[5] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, and P. Laface, “Loquendo – Politecnico di Torino's 2006 NIST Speaker Recognition Evaluation System,” in *Proc. Eurospeech*, pp. 1238–1241, Antwerp, 2007.

[6] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Improvements in factor analysis based speaker verification,” in *Proc. of ICASSP*, vol. 1, pp. 113–116, Toulouse, 2006.

[7] S. Kajarekar, N. Scheffer, M. Graciarena, E. Shriberg, A. Stolcke, L. Ferrer, and T. Bocklet, “2008 NIST Speaker Recognition Evaluation: SRI System Description,” in NIST SRE Workshop, Montreal, Canada, 2008.