Dealing with Limited and Noisy Data in ASR: a Hybrid Knowledge-based and Statistical Approach

Abeer Alwan

Department of Electrical Engineering,
University of California, Los Angeles
alwan@ee.ucla.edu

Abstract

In this talk, I will focus on the importance of integrating knowledge of human speech production and speech perception mechanisms, and language-specific information with statistically-based, data-driven approaches to develop robust and scalable automatic speech recognition (ASR) systems. As we will demonstrate, the need for such hybrid systems is especially critical when the ASR system is dealing with noisy data, when adaptation data are limited (for the case of speaker normalization and adaptation), and when dealing with accents.

Index Terms: noise-robust ASR, speaker normalization, speaker adaptation, accented English, limited data, knowledge-based

1. Introduction

The last few decades have witnessed wide-spread use of speech processing devices and tremendous progress in their performance and reliability. Using mathematical models of human speech production and perception has been an important factor in the improved performance of these devices. For example, simplified linear models of speech production form the basis of several speech synthesizers [24] and the most widely-used speech coder today: CELP (Code Excited Linear Prediction) [30]. Simple auditory models have been used successfully in optimizing the performance of speech and audio coders [31, 36, 21] and are embedded in the MPEG audio-coding standards [6]. Using auditory models as preprocessors has resulted in improved performance of automatic speech recognition (ASR) systems in noise [18, 32].

Typical speech recognition systems involve two fundamental steps: short-term spectral analysis, followed by pattern comparison with representative templates (or statistical models of templates). Today, many systems use Mel-Frequency Cepstral Coefficients (MFCCs) and their temporal derivatives for feature extraction, and Hidden Markov Models (HMMs) of the templates for pattern comparison [29]. MFCCs are defined as the Discrete Cosine Transform (DCT) of log spectral estimates obtained with a critical bandwidth-like non-uniform filter bank model [12]. The DCT provides an orthogonal transformation to a vector space with better energy compaction, which therefore requires fewer (and largely decorrelated) coefficients per acoustic vector. The wide use of temporal derivatives indicates that much of the information of the acoustic signal may be represented in the changes that occur over time.

Providing a rigorous stochastic framework, HMMs and the techniques to train and apply them, have led to successful large-vocabulary speaker-independent systems. Unfortunately, the performance of ASR systems still degrades significantly when training data are limited and/or the acoustic environment (amount and type of background noise, reverberation, competing sources, etc.) differs from the training one.

In this paper, we will provide examples on how speech and language-specific knowledge can improve HMM-based ASR performance when training data are limited or noisy.

2. Limited Data: Rapid Speaker Adaptation and Dealing with Accented English

Spectral mismatch between training and testing utterances can cause significant degradation in the performance of ASR systems. Speaker adaptation and speaker normalization techniques are usually applied to address this issue. The maximum likelihood linear regression (MLLR) [25] technique is widely used in speaker adaptation due to its effectiveness and computational advantages. When the adaptation data are sparse, however, MLLR's performance degrades because of unreliable parameter estimation.

In the NSF-funded project, TBALL (Technology Based Assessment of Language and Literacy) we aim to advance the state of the art in children's speech recognition, datamining, and human-computer interface design so that effective child-friendly *conversational* interfaces can be designed and developed. These technologies are researched in a framework of early learning and integrated with an understanding of the components of children's academic performance to develop a literacy assessment system (for an overview, please see [2]). Several K-5 schools in Northern and Southern California are partners in this project. These schools have a diverse economic and ethnic student body with more than half of the population being Hispanic. The project addresses several fundamental research issues including pronunciation modeling and speaker adaptation techniques that are scalable to children who are native and non-native English speakers.

Acoustic data from young children are not as widely available as adult data are, and there is no corpus of Spanish-accented English spoken by young children. Hence, relying only on data-driven techniques is not a viable option. Moreover, since time is of essence in a classroom situation, rapid speaker adaptation with minimal adaptation data is critical. In the following, I will describe our successful efforts in creating rapid adaptation algorithms based on physiological and acoustic constraints. In addition, knowledge of the child's first language (Mexican Spanish in our case) can help improve pronunciation modeling, and hence ASR, significantly.

2.1. Rapid Speaker Adaptation with Limited Adaptation Data

We developed two algorithms for rapid speaker adaptation. The first involves warping the speech spectra by paying particular attention to the third formant frequency (F3), which is more correlated with the speaker's vocal-tract length than F1 or F2 [16, 15]. The second algorithm performs normalization based on the location of the subglottal resonances in the speech spectra. Both techniques are computationally efficient and perform better than maximum-likelihood based vocal tract length normalization (VTLN) [22] for limited adaptation data.

In the first method [11], speech spectra are reshaped by aligning corresponding formant peaks between training and test spectra. There are various levels of mismatch in formant structures. Regression-tree based phoneme- and state-level spectral peak alignment is proposed for rapid speaker adaptation using linearization of VTLN [40]. This method is investigated in an MLLR-like framework, taking advantage of both the efficiency of frequency warping (VTLN) and the reliability of statistical estimations (MLLR). Two different regression trees are investigated: one based on phonetic classes (using combined knowledge and data-driven techniques) and the other based on Gaussian mixture classes. Compared to MLLR and ML-based VTLN, improved performance can be obtained for both supervised and unsupervised adaptations for both medium vocabulary (the RM1 database) and connected digits recognition (the TIDIGITS database) tasks. Performance improvements are largest with limited adaptation data, which is often the case for ASR applications, and these improvements are shown to be statistically significant.

In [38, 39] another speaker normalization technique, based on subglottal resonances, was introduced. Speaker normalization typically focuses on variabilities of the supra-glottal (vocal tract) resonances, which constitute a major cause of spectral mismatch. Recent studies [8, 28] show that the subglottal airways also affect spectral properties of speech sounds. The speaker normalization method is based on estimating the second and third subglottal resonances (hereafter referred to as Sg2 and Sg3, respectively). It should be noted that the algorithm that automatically detects the subglottal resonances, especially Sg2, was calibrated using direct measurements from accelerometer data collected simultaneously with the acoustic recordings.

Since the subglottal airways do not change much for a specific speaker, the subglottal resonances are independent of the sound type (vowel, consonant, etc.) and remain almost constant for a given speaker of a certain age. This context-free property makes the proposed method suitable for limited-data speaker adaptation. The method is computationally more efficient than maximum-likelihood based VTLN in estimating frequency-warping factors, with performance better than ML-based VTLN especially for limited adaptation data in a variety of testing conditions and tasks.

Cross-language variability of Sg2 was then investigated for a number of English and Spanish words spoken by bilingual children. Analysis showed that, as predicted, Sg2 is independent of speech content and language. Based on these observations, a cross-language speaker normalization method using Sg2 was developed. Experimental results showed that Sg2 normalization is more robust across languages than VTLN, with no significant performance difference observed when adaptation data changed from English to Spanish. This language-independent property of Sg2 leads to robust cross-language normalization, whereby acoustic models trained in one language can be adapted with data in another language, which may be useful in ASR applications for second-language learning.

2.2. Dealing with Accented English

There are many children in California who are of Hispanic origin, and their speech exhibits various degrees of accentedness.

A number of standard ASR and machine learning techniques allow us to integrate prior knowledge of accented children's speech into our automatic assessment modules. Acoustic models can be trained for the fundamental sounds of speech based on both accented and unaccented recordings, to cover as much of the variability in pronunciation as is seen in the data. When decoding these acoustic models from an unknown speech signal, we can constrain the results to a closed set of pronunciations that reflect common variants made by speakers in our target population. Finally, in synthesizing a binary reading score from a set of acoustic cues, we can condition our decision on prior knowledge of the child's demographics, native language, grade level, and other factors we would assume teachers to know and use when making the same assessment. These building blocks afford us the methods to assess students fairly, regardless of their native languages or pronunciation idiosyncrasies.

For example, using a mapping of Spanish phonology to English can help predict (and tag) pronunciations and talkers with Spanish accented speech [42]. Therefore, we know that if speakers read a word that way, they are probably reading correctly and if intervention is needed it is in English phonology. Using Spanish letter to sound rules we can predict pronunciations that apply Spanish rules to English words, which diagnoses a reading issue in English, but also flags that the concept of letter to sound rules is being learned. This is a real learning opportunity for the teacher to know about and is far more important than if the recognizer just said 'incorrect'. We also found that all of the optimized vowel letter-sound dictionaries included the Spanish-confusable letter-sound pronunciations, proving that the addition of these unacceptable pronunciations in the dictionary was helpful in detecting pronunciation errors in nonnative speakers. If a child repeatedly makes Spanish-related pronunciations, the system will detect this class of errors, and the teacher can infer from the automatic results that the child may be confusing the letter-sound pronunciations of the two languages. Both letter-sound and word verification results improved when the pronunciation dictionary included Hispanic pronunciations with models trained on in-domain data [4, 37].

3. Noise Robust ASR

Speech recognition systems trained in quiet suffer from performance degradation in the presence of ambient noise. This is mainly due to the mismatch between the clean acoustic models and noisy features. There are two ways to reduce the mismatch to achieve satisfactory performance. One approach is to either denoise front-end feature vectors while keeping the clean models unchanged [5, 14], or to develop noise robust features [19, 9]. The other approach involves adapting the back-end acoustic models according to the noisy environments [25, 17].

Over the years, members of our laboratory have developed a number of front-end and back-end processing techniques for noise robust speech recognition that are inspired by the way humans produce and perceive speech especially in noise. The techniques include: variable frame rate analysis [43, 41], peak isolation [32], threading formant peaks [33], incorporating voicing [35], and harmonic demodulation [45]. Recognition results with the Aurora databases show that a combination of these techniques results in a significant reduction in word error rate for the mismatched case (clean training and noisy testing) when compared to the MFCC front-end, with no significant increase in computational complexity (e.g., [44, 9]).

The **Variable Frame Rate (VFR)** algorithm is motivated by the fact that changes in spectral characteristics are important cues for discriminating and identifying speech sounds [1]. These changes can occur over very short time intervals. Computing frames every 10 ms, as is commonly done in ASR, is not sufficient to capture such dynamic changes. The VFR algorithm increases the frame rate for rapidly-changing segments with relatively high energy and decreases the frame rate for steady-state segments, based on a weighted log energy Euclidean MFCC distance [43] or on computing entropy changes between adjacent frames [41].

The **Peak Isolation** algorithm is a cepstral-processing technique to enhance local spectral peaks. It was motivated by perceptual experiments with spectrally complex speech-like stimuli that indicated that human audition is particularly sensitive to the frequency location of local spectral peaks [23]. Measurements of the timing detail of individual inner hair cell responses reveal firing responses which track dominant spectral features [13], suggesting one possible mechanism for increased sensitivity to local spectral peaks.

Unfortunately, obtaining reliable estimates for formant frequencies and their temporal derivatives in a noisy acoustic signal is a considerable challenge. However, we have shown that a relatively simple tracking of the dominant spectral peak in three spectral regions can provide robust measures which improve the performance of an HMM-based speech recognition system at low signal to noise ratios [33].

Most feature vectors used in speech recognition do not include voicing information. With a single speaker in a clean acoustic environment, there may be few instances where voicing information is necessary for communication: whispering works in quiet situations at close range. However in noisy situations, the temporal regularity of voicing may be a critical cue for the separation of speech from background noises in similar spectral regions but with entirely different temporal structure. Using a temporal pitch-perception model [26], [34] has shown that running autocorrelation measures of amplitude modulation in the pitch-range for voiced speech can predict the perceptual detection of voicing [35], and increase the robustness of a speech recognition system in noise.

The **Harmonic demodulation (HD)** algorithm is a method that aims at reducing the difference between clean and noisy speech spectra, particularly in inter-harmonic valleys. Here, speech production is viewed as amplitude modulation in the frequency domain, with the excitation spectrum being the carrier and the spectrum of the vocal tract transfer function being the modulator. Non-coherent demodulation with non-linear envelope detection is used to recover the spectrum of the vocal tract transfer function [45]. Envelopes of the speech spectra, instead of the speech spectra themselves, are used to compute the acoustic features.

Recognizing speech requires processing and analyzing a hierarchy of structured cues that evolve over widely varying time scales. Our work has focused on the relatively low-level structure from the rate of the formant frequency through the rate of glottal pulses to the rate of the syllable. Our initial results are encouraging. The current challenges are to extend the findings to larger-scale recognition tasks, and to find intelligent ways to use the redundant information available throughout the hierarchy of speech with various time-scales at the feature, phoneme, syllable, word, phrase and sentence level.

4. Summary and Conclusion

The most effective path to improved performance of speech applications, especially ASR, is a hybrid knowledge-based and data-driven one. Large databases (when available) can be used to derive statistically-salient models. Knowledge of speech production, perception, and language structure can help improve ASR performance especially when training data are limited or noisy. Examples from noise-robust ASR, rapid speaker adaptation, and ASR of accented English were provided. Despite significant technical advances, a robust speech recognition system that approaches human performance still does not exist [27]. Research on improving our understanding of human cognitive abilities and on how best to constrain and regularize data-driven approaches, therefore, continues.

Specifically, in the speech production area, better understanding and quantitative characterization of the acoustics and articulatory dynamics of both normal and pathological human speech are needed. In particular, data and models that address articulatory and acoustic variabilities, both within and across speakers are critical with benefits extending to the clinical and linguistic areas.

In the perception area, better models of the human capacity for perceiving speech in noise are critical. While healthy-hearing individuals are remarkably adept at isolating a specific speech signal from the background noise and understanding what is said, the performance of automatic speech recognition and hearing aids degrade significantly in noisy environments. An important direction here would be to understand and model how acoustic cues are weighted differently depending on the acoustic environment. For example, perceptual experiments reported in [7, 20] show that while both the voice-onset-time (VOT), a primarily temporal cue, and the transition of the first formant frequency (F1) play an equally important role in perceiving voicing for plosives in quiet, the role of the VOT is virtually diminished in the presence of additive white noise.

Finally, thinking beyond HMMs to other probabilistic graphical models [3] that are more amenable than HMMs to including rule-based and/or heuristic speech and language constraints might greatly improve ASR performance.

5. Acknowledgments

Work supported in part by past and present NSF and NIH grants. Special thanks to former and current students and collaborators who made important contributions all along, and research, most interesting.

6. References

[1] A. Alwan, J. Lo, Q. Zhu, ”Human and Machine Recognition of Nasal Consonants in Noise,” Proceedings of the 14th International Congress of Phonetic Sciences, Vol. 1 Page 167-170, August 1999, San Francisco.

[2] Alwan, A., Y. Bai, M. Black, L. Casey,M. Gerosa, M. Heritage, M. Iseli, B. Jones, A. Kazemzadeh, S. Lee, S. Narayanan, P. Price, J. Tepperman, and S. Wang, “ A System for Technology Based Assessment of Language and Literacy in Young Children: the Role of Multiple Informa-

tion Sources", IEEE Multimedia Signal Processing Workshop, pp. 26-30, Oct. 2007.

[3] Bilmes, J., and Bartels, C., "Graphical model architectures for speech recognition," IEEE Signal Processing Magazine," Volume: 22, Issue: 5, pp. 89- 100, Sept. 2005.

[4] Black, M., Tepperman, J., Kazemzadeh, A., Lee, S. and Narayanan, S., "Pronunciation Verification of English Letter-Sounds in Preliterate Children." Proceedings of InterSpeech 2008 , Brisbane, Australia, September 2008.

[5] Boll, S. "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[6] Brandenburg, K. and G. Stoll. "The ISO/MPEG Audio Codec: A Generic Standard for Coding of High Quality Digital Audio." Audio Engineering Society Preprint 3336, 1992

[7] Chen M. and Alwan, A. "On the Perception of Voicing for Plosives in Noise," Proc. EUROSPEECH 2001, Aalborg, Denmark, Vol. 1, pp. 175-178.

[8] Chi X. and M. Sonderegger, "Subglottal coupling and its influence on vowel formants," *JASA*, 122(3):1735-1745, 2007.

[9] Cui, X., M. Iseli, Q. Zhu, and A. Alwan, "Evaluation of Noise Robust Features on the Aurora Databases," ICSLP Proceedings, Denver, Colorado, Sep. 2002, Vol.1, pp.481-484.

[10] Cui, X., A. Bernard, and A. Alwan, "A noise-robust ASR back-end technique based on weighted Viterbi recognition," *Proc. of Eurospeech*, pp. 2169–2172, 2003.

[11] Cui, X., and A. Alwan, "Adaptation of Children's Speech with Limited Data Based on Formant-like Peak Alignment," Computer Speech and Language, Vol. 20, Issue 4, pp. 400-419, October 2006.

[12] Davis, S. B., Mermelstein P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. ASSP **28** (1980) 357–366

[13] Delgutte, B. and Kiang, N. Y. S. "Speech coding in the auditory nerve: I. Vowel-like sounds," JASA **75** (1984) 866–878

[14] Deng, L., A. Acero, J. Droppo, and X. Huang, "High-performance robust speech recognition using stereo training data," *Proc. of ICASSP*, 2001.

[15] Eide, E. Gish, H., " A parametric approach to vocal tract length normalization," Proc. ICASSP, 1996. Vol. 1, pp. 346-348

[16] Fant, G. Speech Sounds and Features. MIT Press, Cambridge, MA. (1973).

[17] Gales, M. and S. Young, "HMM recognition in noise using parallel model combination," *Proc. of Eurospeech*, vol. 2, pp. 837–840, 1993.

[18] Ghitze, O. "Auditory nerve representation as a front-end for speech recognition in a noisy environment," Computer Speech and Language **1 (2)** (1986) 109–130

[19] Hermansky, H., and N. Morgan "RASTA processing of speech," IEEE Trans. Speech and Audio Proc. **2 (4)** (1994) 578–589

[20] Jiang, J., Chen, M., and Alwan, A. "On the perception of voicing in syllable-initial plosives in noise", Journal of the Acoustical Society of America, Volume 119, Issue 2, pp. 1092-1105, February 2006.

[21] Johnston, James D. "Transform Coding of Audio Signals Using Perceptual Noise Criteria." IEEE Journal on Selected Areas in Communications **6 (2)** (1988) 314-323

[22] Kamm, T., G. Andreou, and J. Cohen. "Vocal tract normalization in speech recognition: Compensating for systematic speaker variability." Proceedings of the 15th Annual Speech Research Symposium, pages 161–167, Johns Hopkins University, Baltimore, MD, 1995.

[23] Klatt, D. "Prediction of perceived phonetic distance from critical-band spectra: a first step," Proc. IEEEE ICASSP, Paris, 1278–1281, 1982

[24] Klatt, D. "Review of text-to-speech conversion for English," JASA vol. 82, no. 3, pp. 737-793, 1987.

[25] Leggetter L. and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density Hidden Markov Models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[26] Licklider, J.C.R. "A duplex theory of pitch perception," Experientia **7** (1951) 128–134

[27] Lippman, R. P. "Speech recognition by machines and humans," Speech Communication **22** (1997) 1–15

[28] Lulich, S.M. "Subglottal resonances and distinctive features," *J. Phon.*, to appear.

[29] Rabiner, L.R.: "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proc. IEEE **77** (1989) 257–286

[30] Schroeder, M. R., Atal, B. S., Hall, J. L.: "Optimizing digital speech coders by exploiting masking properties of the human ear," JASA **66** (1979) 1647–1652

[31] Schroeder, M.R., Atal, B.S.: "Code-Excited Linear Prediction (CELP): High-Quality Speech at Very Low Bit Rates." Proc. IEEE ICASSP (1985) 937–940

[32] Strope, B., Alwan, A.: "A model of dynamic auditory perception and its application to robust word recognition," IEEE Transactions on Speech and Audio Processing **5 (5)** (1997) 451–464

[33] Strope, B., Alwan, A. "Robust Word Recognition Using Threaded Spectral Peaks," Proc. IEEE ICASSP, Seattle, **II** (1998) 625–629

[34] Strope, B., and Alwan, A.: "Modeling the Perception of Pitch-Rate Amplitude Modulation in Noise", in "Computational Models of Auditory Function", a book edited by Steve Greenberg and Malcolm Slaney, pp. 315-327, IOS Press, NATO Science Series, Netherlands, 2001.

[35] Strope, B., "Modeling auditory perception for robust speech recognition," unpublished Ph.D. dissertation, Department of Electrical Engineering, UCLA, August 1998

[36] Tang, B., Shen, A., Alwan, A., Pottie, G., "A Perceptually-Based Embedded Subband Speech Coder," IEEE Transactions on Speech and Audio Processing, **5 (2)** (1997) 131–140

[37] Tepperman, J., Black, M., Lee, S., Kazemzadeh, A., Gerosa, M., Heritage, M., Alwan, A. and S. Narayanan, " A Bayesian network classifier for word-level reading assessment," Proc. of Interspeech 2007, pp. 2185-2188, Antwerp, Belgium, August 2007.

[38] Wang, S., A. Alwan and S. M. Lulich, “Speaker normalization based on subglottal resonances,” in *Proc. ICASSP*, pp. 4277-4280, 2008.

[39] Wang, S., A. Alwan and S. M. Lulich, “Speaker normalization based on subglottal resonances,” in *Proc. Interspeech 2008*, Brisbane, Australia.

[40] Wang, S., X. Cui and A. Alwan, “Speaker Adaptation with Limited Data using Regression-Tree based Spectral Peak Alignment”, *IEEE TASLP*, Vol. 15, pp. 2454-2464, 2007.

[41] You, H., Q. Zhu, and A. Alwan, ”Entropy-base Variable Frame Rate Analysis of Speech Signals and Its Application to ASR,” in Proc. ICASSP, Pp.549-552, Montreal, Canada, May. 2004.

[42] You, H., A. Alwan, A. Kazemzadeh and S. Narayanan, ”Pronunciation Variation of Spanish-accented English Spoken by Young Children,” Eurospeech 2005, pg. 749-752.

[43] Zhu, Q., and A. Alwan, ”On the use of variable frame rate analysis in speech recognition,” Proc. IEEE ICASSP, Istanbul, Turkey, Vol. III, pp. 1783-1786, June 2000.

[44] Zhu, Q., X. Cui, M. Iseli and A. Alwan, “Noise Robust Feature Extraction for ASR using the Aurora 2 Database,” Proc. EUROSPEECH 2001, Aalborg, Denmark, Vol. 1, pp. 185-188.

[45] Zhu, Q., and A. Alwan, ”Non-linear feature extraction for robust recognition in stationary and non-stationary noise,” Computer, Speech, and Language, 17(4): 381-402, Oct. 2003.