# TOWARD BETTER AUTOMATIC SPEECH RECOGNITION

**D. O'Shaughnessy, W. Wang, W. Zhu, V. Barreaud, T. Nagarajan, and R. Muralishankar**
INRS-EMT, U. of Quebec, 800 de la Gauchetiere west, suite 6900, Montreal Quebec, Canada H5A 1K6 dougo@emt.inrs.ca

## 1. INTRODUCTION

Automatic speech recognition (ASR) performs best when there is a strong correspondence between system training and operating conditions, e.g., when one tests on speech data that is similar in style to that used for training. Mismatch (different environment speakers, or vocabulary) degrades performance. We have developed new model techniques able to adapt to various speech environments without modifying the basic ASR systems: an appropriate feature transformation scheme for the Mel-frequency cepstral coefficients (MFCC), a new speech-processing front-end feature that performs better than the existing MFCC, a log-energy dynamic range normalization technique for ASR in adverse conditions, and a continuous ASR method that exploits the advantages of syllable and phoneme-based sub-word unit models.

## 2. NEW ADAPTATION METHODS

Statistical Data Mapping (SDM) is a new approach for ASR model adaptation. SDM assumes that speech observations are generated by subsets of mutually related random sources. The relationship/model between random sources can be established by a maximum likelihood criterion. Thus speech observations can be mapped through the model to any desired environment without heavy loss of the original information. SDM is a non-linear approach and has the strength to handle non-time-invariant variations, e.g., bandwidth and speech context. SDM has a flexible framework, which can be varied in different applications. We evaluated our algorithms on the Visteon and Superman databases (Scansoft Inc.): a wide-band automobile embedded-speech database, recorded at automobile speeds of low, 50 km/h and 75 km/h, and a narrow-band network speech database. The TIMIT and NTIMIT databases were also used. Finally, we also used the Aurora 2 database, corresponding to TI-DIGITS training data down-sampled to 8 kHz and filtered with a G.712 characteristic. It includes clean and multi-condition training sets. The noisy utterances have SNR from -5 dB to 20 dB. ASR models adapted by our SDM technique have better performance than other models, e.g., non-adapted or adapted by MLLR. The SDM-adapted ASR models have improved performance in noisy environments.

We also dealt with the problem of speech enhancement when only a corrupted speech signal is available for processing. Kalman filtering is known as an effective speech enhancement technique, in which speech signal is usually modeled as autoregressive (AR) model and represented in the state-space domain. Various approaches based on the Kalman filter are presented in the literature. They usually operate in two steps: first, additive noise and driving process variances and speech model parameters are estimated and second, the speech signal is estimated by using Kalman filtering. Sequential estimators are used for sub-optimal adaptive estimation of the unknown a priori driving process and additive noise statistics simultaneously with the system state. The estimation of time-varying AR signal model is based on weighted recursive least square algorithm with variable forgetting factor. The proposed algorithm provides improved state estimates at little computational expense.

In spectral subtraction for speech enhancement, we subtract the magnitude of the noise spectrum from that of the noisy speech, while keeping the noisy phase. Noise spectrum estimated during weak segments at the start of an utterance assumes that noise is stationary. Our approach is to have a sub-band-based speech detector to separate each signal frame into noise or speech. The estimated noise spectrum is updated at each frame with a forgetting factor, thus dealing with unstable noise. We utilize the noise suppression gain information in another way for noise reduction. Spectral valleys are usually more disturbed by noise than spectral peaks. So we emphasize spectral peaks (e.g., formants). We introduced a frequency masking filtering algorithm in the standard MFCC feature extraction algorithm.

## 3. FEATURE SPEAKER ADAPTATION WITH A SMALL FOOTPRINT

Most ASR systems are trained on a great variety of speakers and are thus called ``Speaker Independent'' (SI) systems. Yet training on data specific to the current user (Speaker dependent) gives better performance. We focused on the use of Speaker Normalization methods. They are useful for ASR that lacks memory and processing capacity such as embedded engines. Indeed, normalization amounts to a set of transformations of speech data (at different steps of the front-end). The light computation required by the normalization process can be done on-line. Moreover, the number of parameters describing this set of transformations is small and their set can be considered as a small user footprint. A speaker S uses a desktop dictation application (namely, Scansoft's Dragon Naturally Speaking, DNS). This application derives a speaker profile (the set of the normalization parameters), which would be used again to normalize S's speech when he uses DNS. One solution is to enroll portable speaker profiles on fully available dictation data with a desktop ASR and then use them for embeded recognition of command words. This architecture raises the problem of intra-speaker variability since the user-speaker used a speaking style that varies from enrollment to test.

We did a study of speaker profile portability. This work has been conducted on MREC, the DNS engine. The objective was to see if our profiles are efficient with small adaptation sets and if they can be used on command test data. Using profiles with MREC Engine, these results showed that speaker profiles could be ported from one task (sentence) to another (command). Their efficiency is reduced but the gain is still significant. This suggests that our framework will produce significant improvment in recognition.

## 4. SEGMENTATION INTO SYLLABLES

Many public-domain speech corpora have only an orthographic transcription. In order to use these corpora to build syllable-level models and ASR systems, an efficient automatic speech segmentation algorithm is required. We made a new explicit segmentation algorithm that uses the orthographic transcription, which allows knowing the number of syllable segments in a speech signal a priori. Although the short-term energy (STE) function contains useful information about syllable boundaries, it cannot be directly used to do segmentation due to significant local energy fluctuations. We use an Auto-Regressive model-based algorithm to smooth the STE function using the knowledge of the number of syllable segments required. If the error is more than 40 ms, those segment points are considered as erroneous boundaries. Experiments on the TIMIT corpus show that the error in segmentation is at most 40 ms for 87.84% of the syllable segments.

We propose a new front-end feature, warped discrete cosine transform cepstrum (WDCTC), which achieves better vowel recognition and speaker-identification performances than the MFCCs. The WDCTC has a better performance than the MFCC in a 5-vowel recognition and speaker-identification task. Feature transformation aims to maximize the desired source of information for a speech signal in the front-end feature and to minimize undesired sources such as noise, speaker variability and other speech signals. Frequency domain interpretations of the feature transformation provide useful and interesting information by highlighting the importance of different frequency regions for a particular vowel. Tests are conducted with unknown vowel samples extracted from continuous speech in TIMIT. The average vowel-recognition performance with the feature transformation scheme was 71.2%.

## 5. ENERGY RANGE NORMALIZATION

Cepstral mean normalization (CMN) and Cepstral Variance Normalization (CVN) are simple noise robust post-feature processing techniques. In CMN, the log-energy feature (or C0) is treated in the same way as other cepstral coefficients. Compared with cepstral coefficients, the log-energy feature has quite different characteristics. We try here to find a more effective way to remove the effects of additive noise for the log-energy feature. We propose a log-energy dynamic range normalization (ERN) method to minimize mismatch between training and testing data. Comparing with that of clean speech, characteristics of the log-energy feature sequences of noisy speech are: elevated minimum value, and valleys that are buried by additive noise energy, while peaks are not affected as much. The larger difference for valleys leads to a mismatch between the clean and noisy speech. Obviously, mean normalization is not an optimized solution. We suggest an algorithm to scale the log-energy feature sequence of clean speech, in which we lift valleys while we keep peaks unchanged. The dynamic range of log-energy feature sequences of an utterance is normalized to a target dynamic range.

The proposed algorithm was evaluated on the Aurora 2.0 digit recognition task. The proposed log-energy dynamic range normalization algorithm had overall about a 31.56% relative performance improvement when systems were trained on a clean speech training set. This method does not require any prior knowledge of noise and level. It is effective to improve the performance of speech recognition for eight different noise conditions at various SNR levels. Histogram equalization and non-linear transformation techniques have been reported to achieve 51.81% or 49.71% performance gains in clean-condition training in the literature, but the proposed method can be easily combined with variance normalization to get a better result (54.21%). In addition, the proposed method does not need to estimate feature density functions or to direct transformation functions. It only needs a very small extra computation load.

## 6. CONCLUSION

Feature transformation interpreted as a filter bank throws light on the relative significance of different frequency bands for a particular vowel and helps in understanding vowels from the frequency domain perspective. Inclusion of WDCTC as a front-end feature in the state-of-art ASR systems may improve the overall performance.

## REFERENCES

T. Nagarajan, D. O'Shaughnessy, Explicit segmentation of speech based on frequency-domain AR modeling, paper 1299, INTERSPEECH, Lisbon, Sept. 2005.

V. Barreaud and D. O'Shaughnessy, Experiments on Speaker Profile Portability, # 1832, INTERSPEECH-05

R. Muralishankar, A. Sangwan, D. O'Shaughnessy, Warped Discrete Cosine Transf Cepstrum: EUSIPCO, 2005.

W. Zhu and D. O'Shaughnessy, Log-energy dynamic range normalization for robust speech recognition, ICASSP, Philadelphia, PA, SP-L.11.2, March 2005.

W. Wang, D. O'Shaughnessy, Robust ASR Model Adaptation by Feature-Based Statistical Data Mapping, INTERSPEECH 2004, Oct. 2004, Jeju, Korea.

**ACKNOWLEDGMENTS :** We thank NSERC and PromptQuebec for their funding support.