A channel normalization technique for speech recognition in mismatched conditions

Neil Kleynhans and Etienne Barnard

Department of Electrical, Electronic and Computer Engineering
University of Pretoria, South Africa
ntkleynhans@csir.co.za and
Human Language Technologies (HLT) Group
Meraka Institute
ebarnard@csir.co.za

Abstract

The performance of trainable speech-processing systems deteriorates significantly when there is a mismatch between the training and testing data. The data mismatch becomes a dominant factor when collecting speech data for resource scarce languages, where one wishes to use any available training data for a variety of purposes. Research into a new channel normalization (CN) technique for channel mismatched speech recognition is presented. A process of inverse linear filtering is used in order to match training and testing short-term spectra as closely as possible. Our technique is able to reduce the phoneme recognition error rate between the baseline and mismatched systems, to an extent comparable to the results obtained by the widely-used cepstral mean subtraction. Combining these techniques gives some additional improvement.

1. Introduction

In this paper, we investigate a channel normalization technique that reduces the speech data channel mismatch between varied sources by estimating the average short-term spectral energy and then filtering the speech data with an appropriate mapping filter.

Any mismatch between training and testing speech data significantly degrades the performance of trainable speech-processing systems. The mismatch is introduced by physical processes such as background noise, non-stationary noise, recording transducers and transmission channels, as well as population differences such as speaker dialects, age and gender distributions, etc. Only the combined effect of these varying processes are generally observable in the data; therefore all these effects are treated as one "channel" mismatch process. Once a mismatch has been identified, channel normalization techniques are employed to reduce the effect it has on the speech system. Such issues are often dealt with by recording sufficiently variable training data, but the penalty introduced by the channel mismatch becomes critical when a resource scarce language is used. One of the major problems in dealing with resource scarce languages is that collecting speech data is expensive and the amount of data is not comparable to that traditionally used for global languages. One method to reduce the impact of data scarcity is to use different recording devices such as cellular phones, land-line phones and computer microphones. However, this method would inevitably introduce a channel mismatch. Thus, an effective channel normalization technique is needed to satisfactorily reduce the channel mismatch. Ideally, one would want the speech system to behave as if the speech data originated from one source.

There are many strategies that are used to minimize the effect of channel mismatch. In the fortunate case that speech data is available from all the channels, channel-dependent acoustic models can be trained or existing acoustic models could be adapted to better handle incoming speech data. Even though this strategy works the best, it is rare that enough speech data is available to develop robust acoustic models for each channel. In the speech signal domain, blind channel estimation and inverse filtering have been used to reduce the channel influence on the speech data [1]. However, it is difficult to make assumptions about the channel response and spectral nature of speech data. Experiments have shown that if a non-linear channel response is encountered, the blind channel estimation technique did not provide an increase in recognition accuracy [1].

Feature vector mapping tries to overcome the channel mismatch by treating the channel effect as feature transformation in the model domain [2, 3, 4]. More traditional techniques are Cepstral mean subtraction (CMS) and Relative spectra (RASTA) filtering [5, 6, 1]. CMS subtracts a long-term average cepstral component from each extracted cepstral component. This method has gained significant popularity in speech and speaker recognition systems for removing slow-varying channel changes [7], but a small amount of speech information is also removed [1]. The CMS method can only be used in speech-based systems that use cepstral feature vectors to represent the speech data. The RASTA filtering method applies a filter that rejects spectral components that move too slowly or quickly compared to the normal rate of change of speech spectral components [7]. However, RASTA filtering violates the standard hidden Markov model (HMM) assumption of piecewise stationary [6] and introduces phase distortion [5], which negatively impacts on recognition accuracies. The simple CMS technique has been proved equally as good as phase corrected RASTA for telephony experiments [5].

Based on the previous work done, the three main criteria that were used to develop a new channel normalization technique, were:

- a resource scarce language environment is assumed, therefore generating channel-dependent acoustic models becomes impractical,
- more complex channel normalization techniques afford little benefit over simpler methods, and

- feature vector independence is required in order to benefit a variety of systems.

The CMS technique meets two of the three criteria; therefore it was used as a baseline channel normalization method. The new channel normalization technique should provide a performance gain over no normalization and the resulting error rate of the mismatch data system should be similar to the error rate given by a CMS implementation.

2. Method

As in speech parametrization techniques, which encode short-term speech information, an initial step was to calculate the average short-term spectral energy over the frames of speech. The frame length was chosen to roughly ensure stationarity of the signal, shifted to create overlap between adjacent frames and each frame windowed. Given frames of speech, $X_i^N = \{X_1, X_2, ..., X_N\}$, the average short-term spectral energy is calculated as

$$Y_c(f) = \frac{1}{N} \sum_{i=1}^{N} \mid H_c(f) X_i(f) W_{HAM}(f) \mid^2 \qquad (1)$$

where $H_c(f)$ is the channel frequency response, $X_i(f)$ represents the frame level spectrum and $W_{HAM}(f)$ is the Hamming window frequency response.

It is assumed that the filter response is linear and time-invariant, therefore remaining constant across the frames of speech and speakers in the database. It would be a difficult task to calculate the channel response using just this information, but the goal here is not to determine the most probable frequency response. The desire is to transform the data from a channel, to better match a channel with a different response, through the use of inverse filtering. An approximation of the mapping filter can be found, if the ratio between two average short-term spectral energies is calculated:

$$\tilde{H}_{Inv}(f) = \frac{\mid H_{C1}(f) \mid^2 \sum_{i=1}^{N} \mid X_i(f) \mid^2}{\mid H_{C2}(f) \mid^2 \sum_{j=1}^{N} \mid X_j(f) \mid^2} \qquad (2)$$

If the assumption is made that the speech characteristics are similar across the data collected from varying channels, the difference that is present in the energy distribution is directly as a result of the channel responses. Figure 1, shows the average short-term spectral energy for a subset of data collected from TIMIT and Wall Street journal corpora, which demonstrates a clear difference in the spectral energy distributions.

The assumption that the speech characteristics are similar across corpora could easily be in error. For instance, the phonetic distribution could be skewed, which would result in more energy being present in certain frequency bands. Therefore, as an average short-term spectral energy estimation improvement, confining the estimator and inverse filter calculation to broad phonetic classes should improve the assumption that the speech characteristics of the two sources are similar; we report on experiments involving both the basic idea and the refined approach below.

3. Experiments

A triphone-based HMM phoneme recognizer, developed using the Cambridge University HMM Toolkit (HTK) [10], was used to perform a variety of channel normalization experiments. The task we used for our benchmarking experiments was phone recognition; this allows us to focus on acoustic modelling exclusively. The two corpora chosen for experimentation were TIMIT and Wall Street Journal (WSJ). A difference in channel characteristics can be expected due to the varied recording environments (room acoustics) and setup (type of microphone used to record the utterances). The sample average short-term spectral energy distribution for the two corpora are shown in Figure 1.

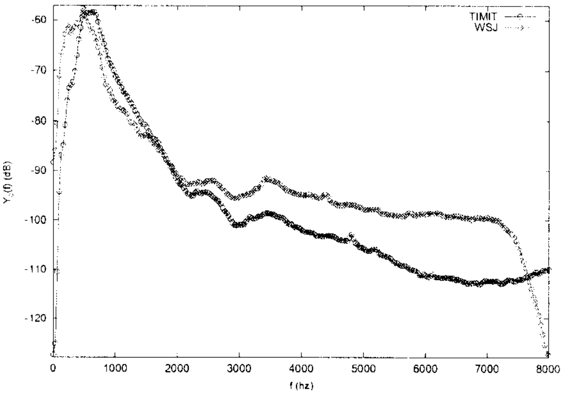

Figure 1: Average short-term spectral energy calculated from a subset of data using the TIMIT[8] and Wall Street Journal[9] corpora.

Task	# Speaker	# File	# Minutes
Recognizer Training	462	2772	143
Recognizer Testing	168	1344	69
Channel Estimator	462	924	46
Broad Classifier Training	462	924	46

Table 1: TIMIT corpus statistics.

The data from both corpora was partitioned into separate sets for phoneme recognizer training and testing, where no speakers were in common between the sets (though for TIMIT, certain sentences did occur in both). The same data was used for channel estimation and training the broad phonetic class classifier; this data was obtained and removed from the phone recognizer training data. The phone recognizer testing data was used to verify the accuracy of the broad phonetic class classifier. The broad phonetic class classifier used six classes: consonants, fricatives, glides, nasals, stops and vowels. Silence was an additional class, but was ignored in the mapping filter calculations.

The TIMIT corpus partitioning statistics are shown in Table 1, while those for the WSJ corpus are given in Table 2.

A number of experiments were run using different channel normalization techniques. The accuracy results are shown in Table 3, which used TIMIT trained acoustic models, and in

Task	# Speaker	# File	# Minutes
Recognizer Training	77	2404	275
Recognizer Testing	24	914	103
Channel Estimator	77	707	88
Broad Classifier Training	77	707	88

Table 2: WSJ corpus breakup statistics.

System Type	Testing Data	
	TIMIT	WSJ
PR	56.80	
BPCC	74.92	
NO NORM		45.30
CMS		51.42
AVG		49.89
CMS + AVG		52.90
SEGRAT		50.51
SEGLS		49.37
COMB		60.48

Table 3: Recognition accuracy results using models trained with TIMIT data.

System Type	Testing Data	
	TIMIT	WSJ
PR		62.03
BPCC		71.69
NO NORM	52.29	
CMS	55.23	
AVG	56.65	
CMS + AVG	56.50	
SEGRAT	56.88	
SEGLS	51.71	
COMB	61.92	

Table 4: Recognition accuracy results using models trained with WSJ data.

Table 4, which contains results for WSJ acoustic models. The system type codes given in Tables 3 and 4 are as follows;

- **PR** - Phoneme Recognizer acoustic models trained and tested using channel-specific data - i.e TIMIT only or WSJ data only.
- **BPCC** - Broad Phonetic Class Classifier, which was trained using unique channel-specific data. PR testing data was used to obtain the accuracy results.
- **NO NORM** - The channel-specific phoneme recognizer was used to decode the unseen channel testing data, e.g. TIMIT trained models decoding WSJ testing data.
- **CMS** - Cepstral Mean Subtraction used by HTK to remove a mean cepstral vector from a set of cepstral vectors extracted from one speech file. The process was applied to both the training and testing data.
- **AVG** - The average short-term spectral energy from each channel was used to derive the mapping filter. The estimation was calculated using unique channel estimation data.
- **SEGRAT** - The channel estimation data was segmented using the BPCC system, which was then used to generate six class-specific average short-term spectral energy estimates. The mapping filter was derived from the average estimates
- **SEGLS** - Same as SEGRAT, except that the mapping filter was derived using a least squares fit between the six average estimates.
- **COMB** - PR acoustic models were trained using data from both channels. No channel normalization methods were used.

4. Discussion

A 5% difference in the corpus-specific phoneme recognizer (PR) results can be explained by the greater number of speakers found in the TIMIT corpus and a larger amount of speech data per speaker in the WSJ corpus. However, the TIMIT phoneme recognition accuracy did improve when the WSJ training data was added to the acoustic model training phase. This improvement was not observed when these acoustic models were tested with the WSJ data. This may indicate that the PR TIMIT acoustic models require much more data to approach the stability of the PR WSJ acoustic models. Considering the channel normalization experiments, the channel-specific acoustic model (PR) results gave an upper bound with which to compare the results obtained from the varying channel normalization tests. The COMB experiment accuracies gave an upper bound for the complete system, and could be considered as an upper bound that can be achieved when both channel normalization and normalization for other factors discussed in Section 1 are employed.

When no channel normalization techniques were used, the phoneme recognizers drop in performance by 10%, which was to be expected. With TIMIT training, the HTK CMS method reduced the drop in accuracies by 5%; that is, about half of the loss is recovered. For WSJ training, only about 30% of the cross-channel loss is recovered with CMS. The average short-term spectral energy filtering method (AVG) gave similar improvements to CMS, being somewhat better for WSJ and somewhat worse for TIMIT. When the CMS and AVG methods were combined (CMS+AVG) and applied to the testing dataset, an improvement in performance was observed compared to the CMS results; now, about 60% of the cross-channel loss is recovered for TIMIT training, and 45% for WSJ training. The AVG and CMS methods can be seen to perform approximately the same task, where AVG modifies the speech waveform and CMS transforms the cepstral coefficients.

The more elaborate BPCC segmentation system gave only small improvements compared to the basic AVG method. Our least-squares approach was clearly not successful, but the SEGRAT was slightly better on both corpora. The statistical significance of the SEGRAT results, compared to the AVG results, were measured using McNemar's test with a chi-squared statistic and the McNemar table of values setup found in Gillick and Cox [11]. A large statistical significance ($P < 0.000001$) was found for the TIMIT trained acoustic models, while the gain obtained for the WSJ trained acoustic models was insignificant ($P < 0.61$). However, many other sensible ways to combine the filters obtained for the different broad phonetic classes remain to be explored. We are therefore confident that the small observed improvement points the way towards even more successful methods.

5. Conclusion

The adverse effect of recording speech data on different channels was demonstrated using the TIMIT and WSJ corpora. A channel normalization technique, which derives a mapping filter from the average short-term spectral energy estimates was shown to give results comparable to the cepstral mean subtraction method. The benefit provided by the new technique is that it is applied to the speech waveform and is therefore independent of the chosen speech parametrization calculations. The broad phonetic class classifier approach provided a small boost to the performance, but the additional work required to implement this method is not justified. However, the marginal improvement was surprising, which indicates that further experimentation must be done to determine how the channel estimation and filtering should be combined to increase the system's performance. The best experimental results that were obtained, came from the case where the acoustic models were trained with speech data from both channel datasets. During the training process, the means and variances of phonetic models are updated to better represent the observed data, therefore a channel normalization process that can translate a model space transform to the speech signal domain should theoretically provide performance enhancements comparable to updated phonetic models. This approach will be further investigated.

6. References

[1] S.J. Wenndt and A.J. Noga, "Blind channel estimation for audio signals", in Proceedings of the Aerospace Conference, March 2004, vol. 5, pp. 3144-3150.

[2] D. Kim and D. Yook, "Feature transform in linear spectral domain for fast channel adaptation", IEE Electronics Letters, vol. 40, no. 20, pp. 1313-1314, September 2004.

[3] Y-F Liao, J-S Lin and S-H Chen, "A Mismatch-Aware Stochastic Matching Algorithm For Robust Speech Recognition", in Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 2003, vol. 2, pp. 101-104.

[4] B. Theobald, S. Cox, G. Cawley and B. Milner, "A Fast Method of Channel Equalisation for Speech Signals and its Implementation on a DSP", IEE Electronics Letters, vol. 35, no. 16, pp. 1309-1311, August 1999.

[5] J. de Veth and L. Boves, "Channel normalization techniques for automatic speech recognition over the telephone", Speech Communication, vol. 25, no. 1-3, pp. 149-164, August 1998.

[6] H. Bourlard, H. Hermansky and H. Morgan,"Towards increasing speech recognition error rates", Speech Communication, vol. 18 , no. 3, pp. 234-235, May 1996.

[7] H. Hermansky and N. Morgan, "RASTA processing of speech", IEEE Trans. Speech Audio Process., vol. 2, no. 4, pp. 578-589, October 1994.

[8] "The DARPA TIMIT Acoustic-Phonetic Continuous speaker space Speech Corpus" (CD-ROM), Available: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1 (Last Accessed: 3 October 2008)

[9] "The DARPA Continuous Speech Recognition Corpus II: Wall Street Journal Sentences" (CD-ROM), Available: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC94S13A (Last Accessed: 3 October 2008)

[10] S. Young, "Large Vocabulary Continuous Speech Recognition.", IEEE Signal Process. Mag., vol. 13, no. 5, pp. 45-57, April 1996.

[11] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms", in Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 1989, vol. 1, pp. 532-535.