

BLIND CHANNEL ESTIMATION BASED ON SPEECH CORRELATION STRUCTURE

Younes Souilmi, Luca Rigazio, Patrick Nguyen, David Kryze, Jean-Claude Junqua

Panasonic Speech Technology Laboratory
3888 State Street, Santa Barbara, CA 93105
{ysouilmi, rigazio, nguyen, kryze, jcj}@research.panasonic.com

ABSTRACT

Cepstral mean normalization is the standard technique for channel robustness. Despite its good performance, the effectiveness of cepstral mean normalization (CMN) for short sentences is argued. CMN underlying hypothesis that the speech cepstral mean is constant is not valid for short processing windows. This implies the removal of some phonetic information. In this paper we show that the speech correlation structure may be used to estimate the communication channel and we propose an efficient algorithm to compute this estimate. We argue that the resulting channel estimate is more accurate because the underlying hypothesis is better verified than the original CMN hypothesis. Results for the Kai-Fu Lee phone recognition task on NTIMIT, with acoustic models trained on TIMIT (mismatch conditions), show that our method provides an 8% relative error rate reduction as compared to CMN.

1. INTRODUCTION

CMN is effective to deal with communication channels [1]. However, CMN is not only removing the channel but also the speech mean. This is a problem especially in the case of short processing windows where the speech mean may carry phonetic information. CMN is based on the assumption that the speech mean does not carry phonetic information or is constant within the processing window. However, it well known that the processing window has to be very long for CMN to work effectively [2]. This may be a problem because one would like to use small windows to deal with non-stationary channels. However, one is also constrained to use long windows to assure that phonetic information is preserved. Our goal is to develop an algorithm that compensates for the communication channel preserving as much information as possible. Our algorithm is not based on the common CMN assumptions, instead it relies on the structure of speech correlation.

2. SPEECH CORRELATION STRUCTURE

The problem of estimating the communication channel affecting a speech signal falls into the category of blind system identification [3]. When only one version of the signal is available (single microphone case) this is a particularly difficult problem that has no general solution. Oversampling may be used to obtain the necessary information to estimate the channel [4]. However, if only one version of the signal is available and no oversampling is possible, assumptions about the signal source need to be made in order to solve each particular instance of the problem. The case of channel estimation for telephone speech recognition, when the recognizer does not have access to the digitizer, falls into this set of problems.

In the blind system identification literature methods that exploit higher order statistics are common. We focus on second order methods since they have been proven very effective and computationally less expensive than higher order methods [5].

We will exploit the structure of speech time correlation. Let $S(t)$ be a speech signal represented in the cepstral (or log spectral) domain. The basic assumption is that the time (inter frame) correlation of clean speech is a decreasing function of τ :

$$E[S(t)S^T(t + \tau)] = f_\tau(E[S(t)S^T(t)]).$$

In other words, f_τ , that captures the short term speech correlation structure, is a characteristic of the speech production process. The intuition is that the communication channel affects the time correlation and that by measuring the correlation of incoming signals we can compute an estimate of the channel. In the time domain, a simple model of the speech signal is an excitation modulated by the vocal track response:

$$s(t) = e(t) * v(t).$$

The physical meaning of our assumption is that v , in average, varies with a quasi-constant speed. Notice that our assumption on the speech correlation structure is weaker than the assumption made by CMN on the long-term cepstral average. It simply captures the fact that the articulatory system moves slowly, without constraining the direction of the movement. We approximate f_τ by a time-invariant linear filter:

$$E[S(t)S^T(t + \tau)] = A(\tau)E[S(t)S^T(t)]. \quad (1)$$

The matrix $A(\tau)$ may be estimated from clean speech as:

$$\hat{A}(\tau) = E[A(\tau)] \approx \frac{1}{T} \int_0^T A(t, \tau) dt, \quad (2)$$

$$A(t, \tau) = E[S(t)S^T(t + \tau)]E[S(t)S^T(t)]^{-1}, \quad (3)$$

$$E[S(t)S^T(t + \tau)] \approx \frac{1}{N} \int_0^N S(t + w)S^T(t + \tau + w)dw, \quad (4)$$

where the integral in equation 2 is carried out over the whole training database, and the integral in equation 4 is carried out over the N samples of the processing window. We estimated the matrices $\hat{A}(\tau)$, $\tau = 1 \dots m$ on the TIMIT database [6]. As we suspected the N_2 norm¹ of the matrices $\|A(\tau)\|^2$ is a decreasing function of the time-lag τ . Also, we measured the relative error introduced by the speech correlation structure assumption. The relative error is an increasing function of the time-lag, ranging from about 4%

¹The N_2 norm for a matrix A is defined as $\|A\|^2 = \text{tr}(A^T A)$.

for time-lags smaller than 50 ms to more than 10% for time-lags larger than 100 ms [7]. We suspect this is due to the fact that for small time-lags we measure intra-phone correlation which is more stable than inter-phone correlation and therefore better predicted by the linear filter model. For channel estimation we should use small time-lags for which the hypothesis is well verified (small relative error). However, the time-lag should not be too small, to ensure that the speech signal correlation does not dominate the communication channel correlation. In this way the measure of the correlation due to the channel, used by the channel estimation algorithm, will be more accurate.

3. THE COMPENSATION ALGORITHM

Consider a speech signal corrupted by a communication channel, observed in cepstral domain (or log-spectral domain):

$$Y(t) = S(t) + H(t).$$

The algorithm is based on the following assumptions:

Hypothesis 1 (Independence) $S(t)$ and $H(t)$ are two independent stochastic processes.

Hypothesis 2 (Short-term stationarity) First order stationarity of $S(t)$: $E[S(t + \tau)] = E[S(t)]$.

Hypothesis 3 (Short-term invariance) The channel $H(t)$ is constant within the processing window: $H(t) = H$.

Hypothesis 4 (Short-term linear correlation structure) The correlation structure of the speech source satisfy the time-invariant linear filter model: $E[S(t)S^T(t + \tau)] = A(\tau)E[S(t)S^T(t)]$.

Our assumptions are to be considered valid for small time-lags (short term structure). The algorithm is based on first and second order statistics. Let's call the correlation at time t with time-lag τ of the signal X , $C_X(\tau) = E[X(t)X^T(t + \tau)]$. Using hypotheses 1 through 3 we obtain the correlation of the observed signal:

$$C_Y(\tau) = C_S(\tau) + \mu_s H^T + H \mu_s^T + H H^T,$$

where $\mu_s = E[S(t)]$. Using the hypothesis 4 we can derive the following terms:

$$\begin{aligned} A &= (I - A(\tau))^{-1} (C_Y(\tau) - A(\tau)C_Y(0)), \\ b &= E[Y(t)], \end{aligned}$$

and the following system:

$$\mu_s \mu_s^T = b b^T - A = B, \quad (5)$$

$$\mu_s + H = b. \quad (6)$$

Since the system is over determined, the solution should be computed by minimizing:

$$\min_{\mu_s} \| \mu_s \mu_s^T - B \|^2. \quad (7)$$

3.1. An efficient solution for the minimization problem

Consider the following minimization problem in the N_2 norm:

$$\min_X \| X X^T - B \|^2,$$

with $X = [x_1 x_2 \dots x_n]^T$ and $B = (b_{ij})_{i,j \in 1, \dots, n}$. Provided that B is diagonalizable, we can write $B = P \Lambda P^*$ where $\Lambda = \text{diag}\{\lambda_1 \dots \lambda_n\}$ is a diagonal matrix and $P = \{p_1, \dots, p_n\}$ is a unitary matrix. Consider the eigenvalues $\lambda_1 \dots \lambda_n$ to be sorted in increasing order $\lambda_1 \geq \dots \geq \lambda_n$. It can be shown that:

$$\min_X \| X X^T - B \|^2 \sim \min_Y \| Y Y^T - \Lambda \|^2,$$

with $Y = P^T X$. Also we can write:

$$\| Y Y^T - \Lambda \|^2 = \sum_i (y_i^2 - \lambda_i)^2 + \sum_i \sum_{j \neq i} (y_i y_j)^2.$$

By taking partial derivatives, we have:

$$\frac{\partial \| Y Y^T - \Lambda \|^2}{\partial y_k} = 4 y_k (\sum_i y_i^2 - \lambda_k).$$

By setting the derivatives to zero we obtain:

$$4 y_k (\sum_i y_i^2 - \lambda_k) = 0, \forall k = 1 \dots n.$$

Since we assumed that $\lambda_1 > \dots > \lambda_n$, from the previous equation, it follows that at most one coefficient among $y_1 \dots y_n$ is non zero. Indeed by contradiction assume that $\exists i_1 \neq i_2 : y_{i_1} \neq 0, y_{i_2} \neq 0$, then we would obtain:

$$\sum_i y_i^2 = \lambda_{i_1},$$

$$\sum_i y_i^2 = \lambda_{i_2},$$

and $\lambda_{i_1} \neq \lambda_{i_2}$, which is impossible. Moreover, given that Y is a non-zero vector we have:

$$\begin{cases} y_{i_0} = \pm \lambda_{i_0} \\ y_i = 0 \quad \forall i \neq i_0 \end{cases}$$

Therefore we conclude that $\| Y Y^T - \Lambda \|^2 = \sum_{i \neq i_0} \lambda_i^2$ and the solution that minimizes $\| Y Y^T - \Lambda \|^2$ is indeed $i_0 = 1$. This also implies that the minimization problem has two solutions $X = \pm \lambda_1 p_1$, where λ_1 is the largest eigenvalue of B and p_1 is the corresponding eigenvector².

3.2. Implementation details

Some implementation details need to be addressed in order to respect our basic hypotheses. The hypothesis of stationarity is not strictly satisfied when using the usual expectation estimator:

$$E[S(t)S^T(t + \tau)] = \frac{1}{N - \tau} \sum_{i=1}^{N-\tau} S(i)S^T(i + \tau).$$

²Note that the solution to the minimization problem is only known up to its sign.

To insure that the hypothesis is satisfied we can use a circular processing window:

$$E[S(t)S^T(t + \tau)] = \frac{1}{N - \tau} \sum_{i=1}^{N-\tau} S(i)S^T(i + \tau) + \frac{1}{\tau} \sum_{i=1}^{\tau} S(N - i)S^T(i).$$

Also, to respect the assumption on the speech correlation structure, only speech frames should be considered when computing the correlation. This requires the ability to distinguish between silence and speech frames. Finally, the short term invariance hypothesis is better satisfied with short processing windows.

Regarding the feature space, the algorithm can be used in both cepstral or log-spectral domain. However, one should assure that the feature dynamic ranges are not too different because that would affect the precision of the diagonalisation algorithm used to solve the mean square error problem. Cepstral coefficients should be normalized by subtracting the long term mean and the covariance matrix should be whitened. In our experiments we used cepstral coefficients, but log-spectral channel removal may be more effective because it is local in frequency and should also be explored [8]. Finally, we used a time-lag of four frames (forty milliseconds) to compute the incoming signal correlation as a compromise between low speech correlation and low intrinsic hypothesis error.

The solution of the minimization problem in equation 7 is obtained by searching for the eigenvector corresponding to the largest eigenvalue (in absolute value). This is a sub case of the diagonalisation problem for non-symmetrical real matrices. Algorithms are known for solving this type of problems, but their precision is bounded by the ratio between the largest and the smallest eigenvalue (i.e. the largest the difference in the eigenvalues the more numerically stable algorithms are) [9]. From our experiments the ratio between the largest and the second largest eigenvalue is between one and two orders of magnitude. Incidentally, this is not only good for the numerical stability of the diagonalisation algorithm, but also means that our assumptions are well verified in practice. It means that it exists one eigenvector that minimizes the cost function much better than any others. This eigenvector provides us with an estimate of the average clean speech μ_s over the processing window. Using this speech estimate we can compute the channel estimate from equation 6. However, the speech estimate is obtained in modulus, therefore an heuristic need to be designed to obtain the correct sign. Hereafter we propose two different methods to obtain the sign:

- Maximum likelihood: the acoustic models may be used to obtain the maximum-likelihood solution for the sign. This may be done with two decoding passes, or with speech and silence GMMs.
- Minimum channel norm: here we make the assumption that, in average, the norm of the channel cepstrum is smaller than the norm of the speech cepstrum. Therefore we compute the sign that minimize the norm of the channel cepstrum.

Also, we use an oracle method to obtain a performance upper bound. The method simply selects the channel that guarantees the smallest error rate. Notice that this method is not usable in practice because it requires perfect knowledge of the phone transcription, and it is only introduced to evaluate the effectiveness of other sign estimation techniques.

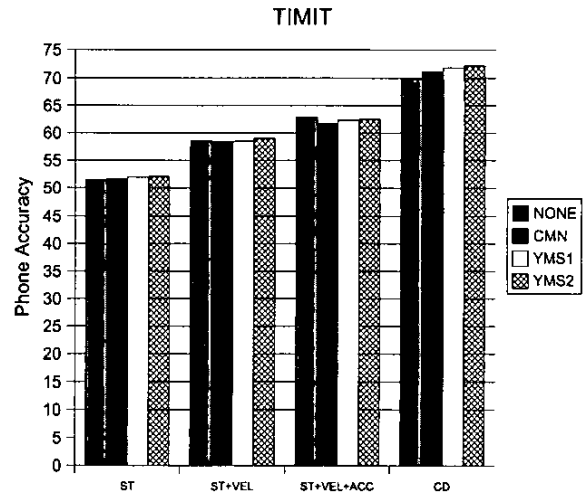


Figure 1: Comparison of CMN and the new adaptation technique on matched channel conditions. Results are reported for context independent models for static coefficients (ST), static plus first derivatives (ST+VEL), static plus first plus second derivatives (ST+VEL+ACC) and context dependent models (CD). Results are reported for no channel compensation (NONE), cepstral mean normalization (CMN), YMS with minimum channel norm sign estimate (YMS1), YMS with oracle sign estimate (YMS2).

4. EXPERIMENTS

Training is performed on the TIMIT database down-sampled at 8 kHz (TIMIT8) and tests are performed on both TIMIT8 and NTIMIT. Thirteen MFCC parameters are computed with window size of 20 ms and frame rate of 100 Hz. Standard Kai-Fu Lee phone recognition tests are performed, with 48 phones used for training and recognition and 39 phones used for scoring. A phone pair language model estimated from the training set phone strings is employed. Context independent phone models (CI) are trained with up to four gaussians per state. Also context dependent tri-phone models (CD) are trained, based on decision tree clustering, with about one thousand states and up to two gaussians per state. Channel compensation is done on the whole utterance and no voice detection is used for both CMN and the proposed algorithm (YMS). This means that both speech and silence are used to estimate the channel. Channel compensation is done at both training and test time. For YMS, the correlation structure matrices $\hat{A}(\tau)$ are estimated on the training set. Experiments for CI models are carried out with three different front-ends: static coefficients only, static plus first derivatives, static plus first and second derivatives. For CD models we run experiments only with the static plus first and second derivatives front-end.

Figure 1 reports results on matched channel conditions. In this case we see that no significant gain in recognition accuracy is obtained from channel compensation methods. This was to be expected since the TIMIT database was recorded in carefully controlled conditions.

Figure 2 reports results on mismatched channel conditions. We remark a significant improvement obtained with channel compensation. The relative improvement tends to decrease when dy-

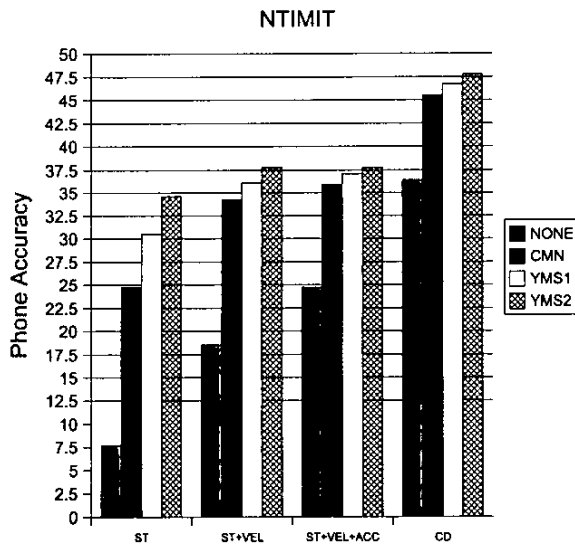


Figure 2: Comparison of CMN and the new adaptation technique on mismatched channel conditions. Results are reported for context independent models for static coefficients (ST), static plus first derivatives (ST+VEL), static plus first plus second derivatives (ST+VEL+ACC) and context dependent models (CD). Results are reported for no channel compensation (NONE), cepstral mean normalization (CMN), YMS with minimum channel norm sign estimate (YMS1), YMS with oracle sign estimate (YMS2).

dynamic features are added. This is because dynamic features are inherently robust to stationary channels. YMS always provides better results than CMN. The improvement obtained with YMS is especially evident when only static features are used. In this case YMS1 (YMS with minimum channel norm sign estimate) obtains 8% relative error rate reduction compared to CMN. This indicates that the channel obtained with YMS is useful in estimating the static component of the speech cepstra. Notice that YMS2 provides only an upper bound of the achievable performance of YMS when the sign is correctly identified, and should not be considered per se as a significant result.

Results for YMS based on the maximum likelihood sign estimate are not reported on the plots. Both GMMs trained on the training database and two pass decoding were tried but yielded poor results. Results for the maximum likelihood sign estimate were always worse than results for the minimum channel norm estimate. We suspect that the initialization of the acoustic models may play a role in this. In principle one should iterate over the training set and estimate the channel sign during the training of the acoustic models. However, since the training database has very little internal channel variations, we chose not to do that and we used a fixed channel sign for training the acoustic models.

The minimum channel norm estimate is the most effective. Both fixed channel signs (plus or minus) performed worse than the minimum channel norm. Also, we tried a maximum channel norm estimate (which provide the opposite sign than the minimum channel norm) and we obtained significantly worse results. However, figure 2 shows that the upper bound accuracy provided by the oracle procedure (YMS2) is still significantly better than the results

obtained with the minimum channel norm sign estimate (YMS1). This means that the sign of the channel estimate carries a lot of residual information and that there still is margin for improvement.

5. CONCLUSIONS

We proposed a new method to estimate the communication channel corrupting a speech signal that is based on the structure of the speech correlation. The method is more effective than the standard cepstral mean normalization because the underlying assumptions are better verified. Our assumptions about the speech correlation structure provide an over determined system that need to be solved via least-squares minimization in order to estimate the communication channel. We proved that an efficient algorithm exists because a least square solution (up to its sign) can be found via diagonalisation. Also, we proposed two methods to estimate the sign of the channel (maximum-likelihood and minimum channel norm), and we compared against CMN when a communication channel mismatch is present.

Static cepstral features, compensated by our channel with minimum norm sign estimate, provided a significant improvement compared to CMN (8% relative error rate reduction). Maximum likelihood sign estimate did not yield an improvement, and should be further investigated. Specifically, one should consider the channel sign as a hidden variable and optimize for it during the EM algorithm, while jointly estimating the acoustic models.

6. REFERENCES

- [1] S. F. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification," *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. ASSP-29, pp. 254-272, April 1981.
- [2] S. Kanthak, S. Molau, A. Sixtus, R. Schlüter, and H. Ney, "The RWTH Large Vocabulary Speech Recognition System for Spontaneous Speech," in *Proc. of Konvens*, Ilmenau, Germany, October 2000, pp. 249-254.
- [3] L. Tong, "Multichannel blind identification: From subspace to maximum likelihood methods," *IEEE*, vol. 86, no. 10, pp. 1951-1968, 1998.
- [4] K. Lee, B. Lee, and S. Ann, "Adaptive filtering for speech enhancement in colored noise," *IEEE Signal Processing Letters*, vol. 4, no. 10, pp. 277-279, 1997.
- [5] H. Zeng and L. Tong, "Blind Channel Estimation Using the Second-Order Statistics: Asymptotic Performance and Limitations," *IEEE Trans. Signal Processing*, vol. 44, no. 8, pp. 2060-2071, 1997.
- [6] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "The DARPA TIMIT acoustic-phonetic continuous speech corpus," in *NIST*, 1990.
- [7] Y. Souilmi, "Robust speech recognition," M.S. thesis, Institut Eurecom, Sophia Antipolis, France, August 2001.
- [8] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 578-589, 1994.
- [9] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes in C*, Cambridge University Press, 1992.