Correlation analysis for the derivation of speech recognition features based on an auditory model

Franck Giron

Sony Deutschland GmbH – Stuttgart Technology Center
EuTEC – Speech and Sound Group
Hedelfinger Strasse 61
giron@sony.de

Abstract

In the framework of TC-STAR, a work package is dedicated to the study of robust methods for speech recognition in home and office environment. One of the approaches considered in this work, is to use features derived from an auditory model. However, as such, auditory models cannot yet be integrated directly in an ASR system without further processing. This paper presents the results of the correlation analysis of different speech recognition features derived from this auditory model. The original goal of the experiments performed for this study was to analyse different post-processing strategies for the computation of robust speech features, which could be included in an HMM-based speech recognition system, using Gaussian Mixture distributions for the representation of the output probabilities of these features. The focus of the current study is placed on the decorrelation properties of existing methods described in the literature, like frequency filtering, lateral inhibition or discrete cosine transform . These methods have been applied to the outputs of the auditory model. These methods are compared to the decorrelation properties of a classical MFCC front-end.

1. Introduction

Most of the current up-to-date Automatic Speech Recognition (ASR) systems are based on a Hidden-Markov-Model (HMM) structure using Gaussian Mixture distributions as representation for the output probabilities of some features, which are assumed to be an adequate parametric representation of the speech signal. In this HMM framework, it is a considerable computational advantage if the covariance matrices included in the modeling of these distributions can be diagonalized. This reduces considerably the number of parameters to be stored, since the dimensionality of the covariance matrix for each Gaussian is then reduced to a single vector of size N, instead of $N(N+1)/2$, for a feature vector of the same size, but this offers also a considerable speed advantage during the computation of the output probabilities.

In order to create an ASR system for the TC-STAR EPPS06 English task, which should be robust for home and office environment, we have mainly concentrated our activities to the development of an auditory-based front-end. The study has taken into account original recordings performed in one of our meeting rooms at STC, called the Clubroom, which contains recording of spoken commands in a home-environment like scenario, where the speaker was located at some distance (2 to 3 m) from the recording devices. We have also used some part of the utterances pronounced by politicians from the EPPS06 English development database. These recordings contain also some reverberation, since the microphones are located at around 50 cm to 1 m distance from the speakers in the very reverberant parliamentary hall, mostly due to its big size.

One of the difficulties encountered during this preliminary study was the confrontation with the problem of determining adequate speech features, which could be integrated in the HMM framework and satisfy the previous condition of diagonalization.

In the following, we will describe in more details the auditory model, which was used as a fundamental step for the computation of the other speech features. A description of the post-processing algorithms, chosen for this study, will be also given. Next the methodology, which was used for the analysis of correlation, will be discussed with some important points of the statistical analysis and some results based on this analysis will be presented. We will finally conclude with some perspective about the integration of an auditory system in an ASR system.

2. Algorithm used for the computation of auditory features

The coding of speech sounds on the auditory nerve involves many spatial and temporal cues. Many different studies have been performed and are still conducted in the hope of obtaining a better understanding on how the auditory system works. The derivation of the speech features, we have been using here for this study, is based on the original auditory model developed by (Seneff, 1986, 1988), which was chosen for its encouraging results in many single words or phonemes speech recognition applications; see for example (Dobrin et al., 1995) and corresponding references for an overview of the authors which have worked in this direction. Some other authors, like (Jankowski et al., 1995), have had a more reserved opinion on the use of auditory models for ASR although they finally concluded that more work is necessary to obtain improved ways of incorporating features from an auditory model into a speech recognizer. The original model, as developed by Seneff, cannot be used directly for a traditional ASR system based on a HMM structure, which is traditionally limited to a frame-based rate of a few milliseconds (around 5 to 10 ms). In fact the coefficients of the Seneff model are traditionally computed for every speech samples at a sampling rate of 16 kHz with 40 channels corresponding to critical band filters separated on a 0.5 Bark scale. The parameters of this model have been adjusted to match existing experimental results of the physiology of the auditory periphery and these outputs correspond to the probability of firing of the auditory nerve as a function of time for an ensemble of similar fibers acting as a group.

The following flow chart on figure 1 describes the structure of the auditory model and following post-processing steps, as envisaged in this paper.

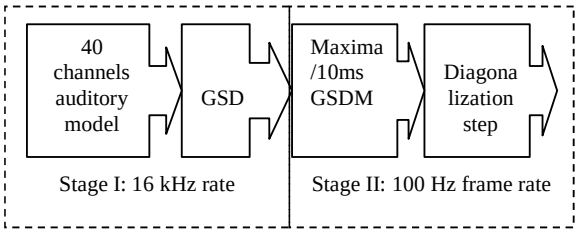

Figure 1 Flow chart of the ASR front -end

Different kind of post-processing steps have been envisaged after this stage by different authors. One of the most famous one, also originally developed by Seneff, is the so-called Generalized Synchrony Detection (GSD). This method has the advantage of producing a clean spectral representation, which preserves prominent peaks at the formant resonances, but also preserves the time discriminative properties of the speech signal. However some authors like (Hunt and Lefèbvre, 1987) and (Abdelatty et al., 2002), have discovered some weak points of the original description and have developed alternative methods. Another approach is the mean-rate model, which is a smoothing of the original output with some predefined time constant and a following down sampling.

We have chosen to develop a similar method, as these authors, based on a combination of the original stage I output of this model with an improved version of the GSD. In stage II, the maximum of each corresponding 40 channels is finally taken on successive windows of 10 ms, to fulfill the frame-by-frame based condition of the ASR. This basic front-end will be named GSDM in the rest of this paper.

An example of output of the original 16 kHz output and the corresponding GSDM is displayed in the following figure 2.

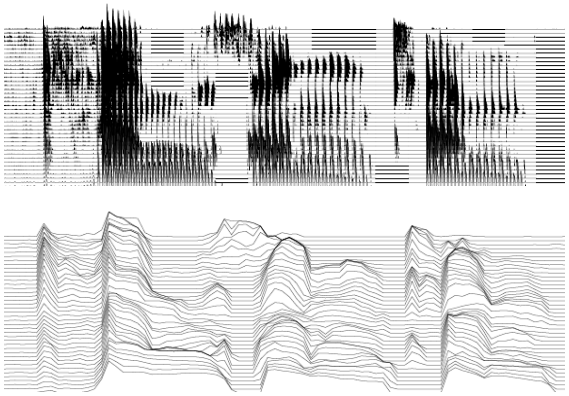

Figure 2 Auditory model output at 16 kHz (top) and derived basic speech features representation GSDM at 100 Hz for the word “Kensington”.

The GSDM preserves the original time-frequency properties but suppresses the glottal excitation present in the speech signal.

The advantage of this representation is its relatively good preservation of time-frequency properties independently of the environmental condition. The figure 3 describes an example of a single command uttered in a reverberant and noisy environment (air-conditioning) recorded with 2 different microphones at 2 different distances from the speaker. The top channel is recorded with a close-talk microphone and corresponds to the best quality achievable in such an environment. The lower one corresponds to a studio microphone placed around 3 m away from the speaker. On the left side are the features resulting from a classical Mel-Frequency bands spectral analysis, before the Discrete Cosine Transform is applied to create the MFCC features. The right column corresponds to the GSDM representation. As can be clearly observed on this figure, the time-frequency structure is better preserved by the GSDM transformation, although the extra noise of the air-conditioning is still present in the low frequency bands.

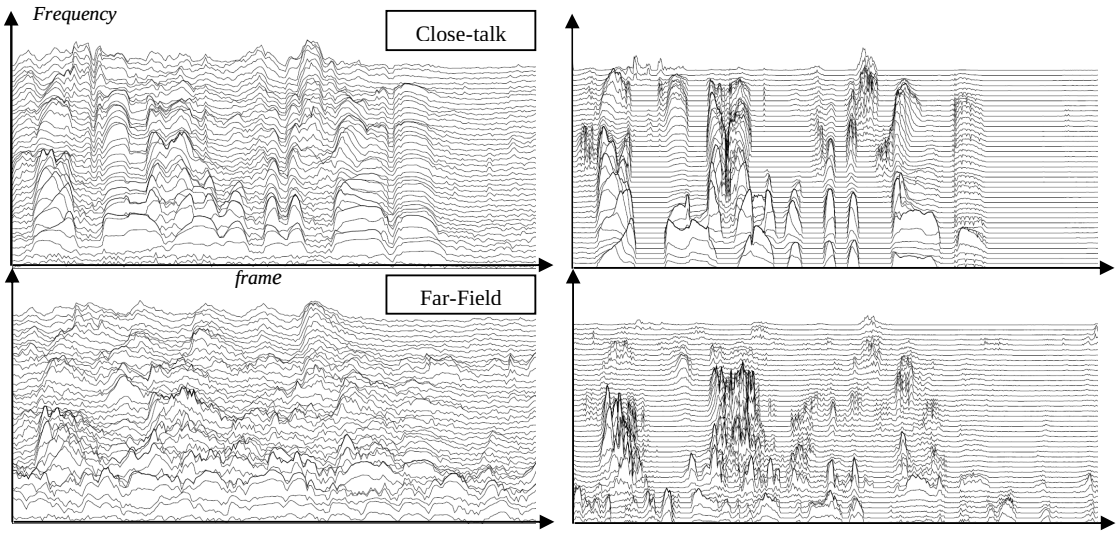

Figure 3 Mel-Frequency spectral representation (left) compared to the GSDM representation (right) for a close-talk microphone (top) and a studio microphone placed at 3 m of the speaker. The speech is uttered in a reverberant room with an air-conditioning low frequency background noise.

3. Post-processing algorithms

Although the GSDM presents very good "visible" spectral properties for the human eye, its adequacy for the integration in an ASR statistical frame-by-frame framework is not guaranteed. The GSDM representation itself is dependent on the original speech level, which means that some normalization issues have to be solved. Moreover, it will be clear later on that all channels are very much correlated and consequently cannot be integrated without any further processing steps.

Different approaches have been envisaged in the following to test their abilities to de-correlate these features. These approaches have been also tested by other authors independently or in combination with an auditory model.

3.1. Frequency filtering

Logarithmic filter-bank energies (FBE) are typical spectral measurements in most current speech recognition systems. The discrete cosine transform is applied to compute, from the set of energies, a set of uncorrelated features, the so-called mel-frequency cepstral coefficients MFCC, which is probably the most widely used spectral representation in speech recognition.

The Frequency-Filtering (FF) features have generally shown an equal or better recognition performance than the MFCCs, and, unlike them, the FF features show a frequency meaning (Nadeu et al., 2001). The FF technique consists of a filtering operation on the frequency sequence of log FBEs, typically with the following second order filter:

$$H(z) = z - z^{-1}$$

In matrix notation,

$$C_F = H \log(FBE)$$

where C_F is the vector of the frequency-filtered parameters, S is the vector of (linear) FBEs and H is the matrix (for 4 dimensions):

$$H = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{bmatrix}$$

Basically, computing the frequency filtered coefficients consists in replacing each original channel by the difference of its adjacent channels. Since this is done in the log domain, this corresponds to a ratio of the respective frequency band energies. In this experiment, the logarithm of the (FBE) has been replaced by the GSDM coefficients.

3.2. Lateral inhibition network

The lateral inhibition network (LIN) has been described in (Shamma, 1985) as a spatial pattern processor, capable of extracting a particular spatial cue in its input pattern. Its simplest and fastest form is the so-called non-recurrent single-layer linear LIN. The following equation is used.

$$X_i = E_i - \sum_j w(i, j) E_j$$

Where X_i is the output coefficient at index i, E_i is the input at i, and $w(i,j)$ are the inhibitory coefficients of the network. For the experiment described in this paper, the input coefficients were also replaced by the GSDM coefficients and the coefficients w were limited to a left-right set, where all values were defined to be equivalent for each channel i.

$$w(i, i-1) = a$$
$$w(i, i) = 0.0$$
$$w(i, i+1) = a$$

The value of a was fixed to 0.5 after experimenting on the results of the correlation coefficient with different utterances of the development database. Basically, in this simple case, the operation can also be interpreted as a kind of frequency filtering with the following filter:

$$H(z) = -a \cdot z^{1} + 1 - a \cdot z^{-1}$$

Or in the other way round, the frequency filtering operation can be interpreted as a lateral inhibition network with weight w:

$$w(i, i-1) = -1$$
$$w(i, i) = 1$$
$$w(i, i+1) = 1$$

In our particular case each channel was then replaced by its difference with the average of the neighboring channels.

4. Correlation analysis

The analysis of correlation needs some careful considerations in the case of the features derived from the auditory model. The most commonly used measure of correlation is called the Pearson product-moment correlation coefficient, which is a measure of how well a linear equation describes the relation between two variables X and Y. However, this coefficient can only be correctly interpreted in term of knowledge from one variable from the other in the case where the analyzed variables are considered to follow the normality condition. This is a relatively good approximation in the case of the MFCC coefficients, but not in the case of the features derived from the auditory features.

In our case, we have studied the following combinations:

- MFCC: 40 traditional Mel frequency cepstral coefficients computed with a 10 ms frame shift on a 16 ms frame length.
- GSD(M): 40 maxima computed on successive frames every 10 ms on our version of the generalized synchrony detection.
- GSD(M)C: 40 coefficients resulting from the discrete cosine transform of the previous GSDM features, corresponding to a kind of MFCC transform where the logarithmic energy of the Mel frequency bands has been replaced by a Bark scala and the corresponding properties of the auditory system.
- G(SDM)FF: 38 Frequency filtering coefficients as defined previously for the GSDM features. The lowest and highest frequencies were excluded, since they don't have neighbors on both sides.
- G(SDM)LH: 38 lateral inhibition of the GSDM features. Here also 2 coefficients less for the same reason.

We have to mention here that it is of course not usual in ASR to use as many cepstral coefficients as the number of frequency bands. These are normally restricted to a smaller number of 13 coefficients to preserve mainly the spectral envelope of the speech signal. However, we have used more coefficients, since the goal here is to compare the decorrelation properties of each method.

For all these features, only the speech part was taken into account, based on the speech detection algorithm scheme used for the computation of the LDA matrix on the MFCC coefficients. By suppressing the silence part of the signal, the normality of all these features is increased, because silence is a very important part of the speech signal, which creates an extra bias of the eventually Gaussian distribution of these features around the means of the silence values.

4.1. Jarque-Bera test

The next figure plots the mean results of the Jarque-Bera test *t* (Bera and Jarque, 1980) and the corresponding *p*-values for all the previous features for an extract of 20 utterances taken for a female speaker of the TC-STAR EPPS06 English development database.

This test is a goodness-of-fit measure of departure from normality, based on the sample kurtosis and skewness. The test statistic is defined as

$$JB = \frac{n}{6}\left(S^2 + \frac{K^2}{4}\right)$$

Where *S* is the skewness, *K* is the kurtosis, and *n* the number of observations. The statistic has an asymptotic chi-squared distribution with two degrees of freedom and can be used to test the null hypothesis that the data are from a normal distribution; since samples from a normal distribution have an expected skewness and kutosis of 0. As the equation shows, any deviation from this increases the JB statistic.

The result of the test is *t* where a value of 1 rejects completely the hypothesis of normality, a value of 0 cannot reject this hypothesis. The hypothesis is rejected if the test is significant at the 5% level. A *p*-value as depicted below less than 0.05 corresponds to a confidence of 95% that this hypothesis is true.

As can be seen here, the original GSDM coefficients (black line with point markers), but also the lateral inhibition and frequency filtering coefficients have completely no normal distribution at all (t=1, p=0), where as some of the corresponding cepstral representation (e.g. GSD-C_i, i = [3,5,25-27]) and almost all MFCC coefficients are normally distributed. ($p > 0.05$) for this speaker.

These results demonstrate that the classical way of computing the correlation is not applicable to the case of auditory (or at least GSDM) based features. The next section gives some indication about a way to cope with this kind of problem.

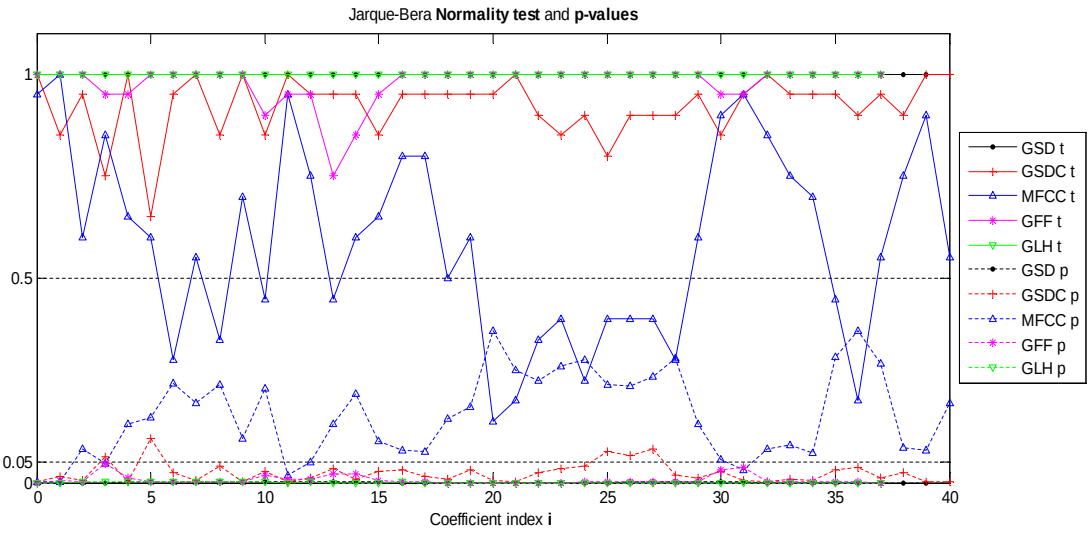

Figure 4 Jarque-Bera Normality test and corresponding p-values (95%) for GSD, GSDC, MFCC , lateral inhibition GLH and frequency filtering GFF features.

4.2. Correlation methods

Alternative ways of analyzing the correlation between variables not satisfying the normality condition are the so-called rank correlations, like the Spearman's rank correlation coefficient r or the Kendall t.

The Spearman's rank r is a non-parametric measure of correlation, which assesses how well an arbitrary monotonic function could describe the relationship between two variables, without making any assumption about the frequency distribution of the variables. Unlike the Pearson product-moment coefficient, it does not require the assumption that the relationship between the variables is linear, nor does it require the variables to be measured on interval scales.

In principle, r is simply a case of the Pearson coefficient in which the data are converted to ranks before calculating the coefficient.

The figure 5 depicts the correlation values of the first coefficient with each other ones for both the FF and the MFCC features. These values have been computed with 2 different methods of correlation computation: The Pearson correlation coefficient and the Spearman's rank correlation. It can be seen that for the MFCC data, which have a normal distribution for almost every coefficients, the difference between each correlation method is very small. This is however not the case for the GFF data, where all coefficients do not have a normal distribution, as we have seen previously.

Correspondingly, we decided to use the Spearman's method for further analysis of the speech features.

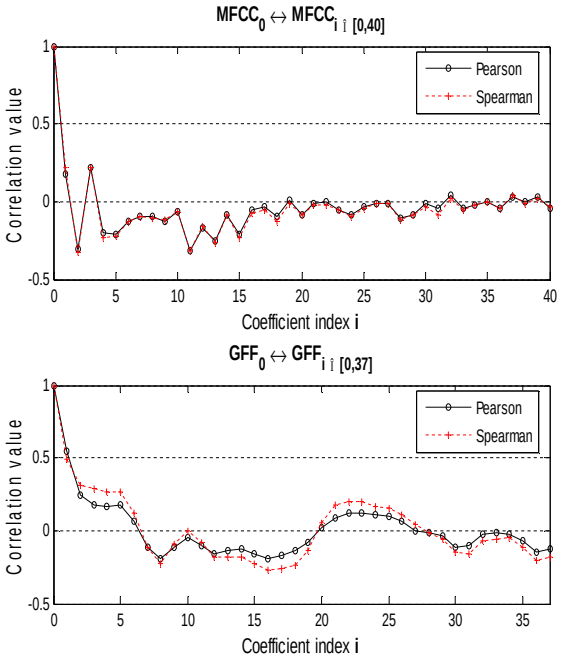

Figure 5 Comparison of Pearson and Spearman's correlation values for the MFCC and G(SDM)FF front end

4.3. Correlation results

The correlation matrices and their corresponding p-values have been computed for the original features MFCC, GSDM, GSDMC, GFF and GLH. The same data as used for the analysis of the normality have been used. For each utterance, the values have been computed separately and the means of all coefficients and p-values have been built.

The figure 6 on next page depicts these results in the form of a square matrix where the values above the diagonal corresponds to the mean of these correlation coefficients; the values depicted in black below the diagonal corresponds to the p-values, which satisfies the condition p-value < 0.05. This means that for correlation values around 0.0, there is a high confidence of 95%, that the corresponding coefficients are not correlated.

For the original auditory coefficients GSDM, it is clear that they are well correlated apart from the comparison between low and high frequencies bands. The corresponding DCT transformation (GSDMC) as well as the lateral inhibition show a clear evidence of their capability to decorrelate the corresponding coefficients, as well as the MFCC is performing for the Mel frequency filter banks. The Frequency Filtering (GFF) method, however, seems to have more area where the coefficients are still correlated. The last picture finally shows a comparison summary between MFCC (lower triangular matrix) versus GSDMC p-values (upper triangular matrix), limited to the first 13 coefficients, as they are usually used in ASR system. This demonstrates that from the point of view of decorrelation, these coefficients have similar properties.

5. Conclusion

This paper has studied the correlation relationship between the features of different post-processing methods applied to an auditory model. It was shown that correlation methods have to be chosen carefully depending on the normality of their distributions.

Among the different post-processing methods used in this paper, no clear winner could be identified in its ability of decorrelating the auditory features. The MFCC and DCT transform of the auditory features have been shown to have relatively similar properties with respect to this aspect.

Consequently the GSDMC features have been used during the EPPS06EN evaluation and compared with the MFCC baseline. No advantage could be identified and lead us to the conclusion that time-frequency approaches should be taken into consideration to profit from the good discriminative properties of this model.

6. Acknowledgments

This work was co-funded by the European Commission within the Sixth Framework Program (2002-2006) for the project TC-STAR (FP6-506738).

7. References

A. M. Abdelatty, J. Van der Spiegel, and P. Mueller. (2002). Robust auditory-based speech processing using the averaged localised synchrony detection.

IEEE Transactions on Speech and Audio Processing, Vol.10 (5), July 2002. pp. 279-292.

A. K. Bera, C. M. Jarque. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economic Letters 6(3)*. pp. 255-259.

C. Dobrin, P. Haavisto, K. Laurila, and J. Astola. (1995). Speech recognition experiments in a noisy environment using auditory system modelling. *EUROSPEECH'95*. pp.131-134.

M. J. Hunt and C. Lefèbvre. 1987. Speech recognition using an auditory model with pitch-synchronous analysis. *Proceedings ICASSP, 1987*, pp. 813-816.

C. R. Jankowski Jr., H-D H. Vo and R. P. Lippmann. 1995. A comparison of signal processing front ends for automatic word recognition. *IEEE Transactions on Speech and Audio Processing, Vol.3 (4), July 1995*.pp.286-293.

C. Nadeu, D. Macho and J. Hernando. 2001. Frequency and Time Filtering of Filter-Bank Energies for Robust HMM Speech Recognition. *Speech Communication, Vol. 34*, pp. 93-114.

S. A. Shamma. 1985. Speech processing in the auditory system II: Lateral inhibition and the central processing of speech evoked activity in the auditory nere. In *J.Acous.Soc.Am.78(5), November 1985*, pp. 1622-1632.

S. Seneff. 1986. A computational model for the peripheral auditory system: application to speech recognition research. *Proceedings ICASSP, 1986*, pp.1983-1986.

S. Seneff. 1988. A joint synchrony/mean-rate model of auditory processing. *J.Phonet., vol.16*, pp.55-76.

Figure 6 Absolute value of correlation matrices (upper triangular part) and corresponding p-values < 5% for the GSDM, its cepstral transformation (GSDM_C), frequency filtering (GFF) and lateral inhibition (GLH), as well as MFCC. The last figure displays the p-values < 0.05% only for the first 13 coefficients of the MFCC and GSDM_C transforms.