interspeech 2008 incorporating SST 08

Pitch adaptive features for LVCSR

Giulia Garau and Steve Renals

Centre for Speech Technology Research, University of Edinburgh, Edinburgh, EH8 9LW, UK

g.garau@ed.ac.uk, s.renals@ed.ac.uk

Abstract

We have investigated the use of a pitch adaptive spectral representation on large vocabulary speech recognition, in conjunction with speaker normalisation techniques. We have compared the effect of a smoothed spectrogram to the pitch adaptive spectral analysis by decoupling these two components of STRAIGHT. Experiments performed on a large vocabulary meeting speech recognition task highlight the importance of combining a pitch adaptive spectral representation with a conventional fixed window spectral analysis. We found evidence that STRAIGHT pitch adaptive features are more speaker independent than conventional MFCCs without pitch adaptation, thus they also provide better performances when combined using feature combination techniques such as Heteroscedastic Linear Discriminant Analysis.

Index Terms: pitch adaptive, speaker normalisation, LVCSR, VTLN, HLDA

1. Introduction

The application of pitch synchronous features to speech recognition has been mainly investigated on small vocabulary recognition tasks. The pitch may be modeled explicitly as a variable [1] or used to extract pitch synchronous features. For example Bozkurt et al. [2] used group delay features extracted using a window centered at the glottal closure instant while Holmes [3] adopted an excitation synchronous fixed length analysis window to extract conventional MFCCs.

In [4] we used a pitch adaptive spectral representation, STRAIGHT [5], to perform experiments on three Large Vocabulary Continuous Speech Recognition (LVCSR) tasks: WSJCAM0, conversational telephone speech and multiparty meeting data. STRAIGHT derived features provided substantial improvements in all the tasks when combined with conventional MFCCs, suggesting that they are complementary to the latter. In this paper we analyse the individual contribution of each representation in two ways. First, we decouple the pitch adaptive and smoothing aspects of STRAIGHT. Experiments performed on the meeting speech recognition task highlight the importance of using a pitch adaptive spectral analysis and the benefit of combining it with a conventional fixed window spectral analysis. Second, a speaker independence metric was used to compare pitch adaptive features with conventional features: it was found that the pitch adaptive component of STRAIGHT provides improved speaker independence. Reduced inter-speaker variability is particularly beneficial when feature combination techniques such as Heteroscedastic Linear Discriminant Analysis (HLDA) are employed.

2. Pitch adaptive features

Previously we adopted a pitch adaptive spectral representation, STRAIGHT [5], to extract MFCCs, yielding consistent improvements in three large vocabulary tasks (WSJCAM0, CTS and meeting data) in combination with conventional features [4]. The spectral analysis of STRAIGHT uses an F0-adaptive window which gives equivalent resolution in both time and frequency domains. An interpolation is then performed on the partial information given by the adaptive windowing. This results in a smoothed time-frequency representation which is not affected by the interference due to the signal periodicity. In our STRAIGHT-based MFCCs we substitute the classic STFT, which uses a Hamming window, with the STRAIGHT spectral analysis where the shape of the window depends on the fundamental frequency and is two fundamental periods long. STRAIGHT based MFCCs were extracted by processing the STRAIGHT (power) spectrogram through a mel scaled filterbank and decorrelating using the discrete cosine transform (DCT).

VTL variability is taken into account by scaling the frequency axis of the observed spectrum with a warping function g_α parameterised by a warping factor α estimated using Maximum Likelihood (ML) [6]. The speaker-specific warp factor α is obtained by maximising the likelihood of the normalised acoustic observation, given a transcription and an acoustic model. In practice the centres of the filters of the mel scaling filterbank are moved according to a piecewise linear frequency warping function.

The STRAIGHT spectral analysis has two concurrent effects: on one side a pitch adaptive window is used for spectral analysis; on the other side smoothing is performed interpolating the partial information provided by the pitch adaptive spectral analysis itself. In our previous experiments we observed that conventional MFCCs outperformed the STRAIGHT based MFCC systems. Therefore in this paper experiments are performed to decouple the two STRAIGHT effects on the close talking meeting task. Figure 1 shows a plot of the spectral contour for one frame of voiced speech for the short time Fourier transform (STFT), and for STRAIGHT, while figure 2 compares the STRAIGHT spectral envelope with that of STRAIGHT using only the smoothing and STRAIGHT using only the pitch adaptive component. It can be noticed that when the pitch adaptive module of STRAIGHT is used with no smoothing some harmonics are still present, while using the smoothing part alone on the other hand seems to yield a very smooth spectral envelope.

3. Experimental Setup

We performed experiments in the meeting domain. Our training set (the same used in our systems for the NIST RT05 and RT06 evaluations) consists of a total of over 100 hours of conversational meetings speech recorded in different sites: 70 hours from the ICSI, 13h from the NIST, 10h from the ISL and finally

Accepted after peer review of full paper

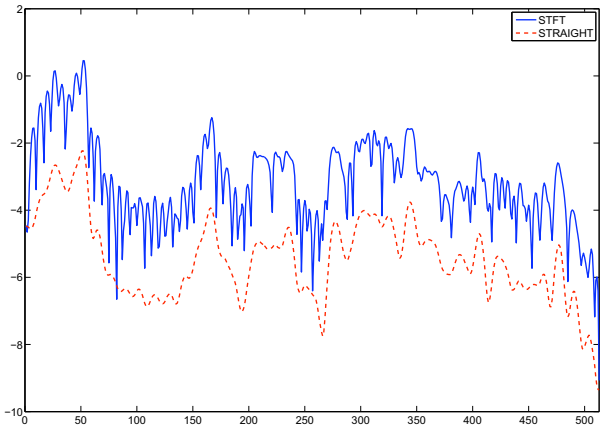

Figure 1: A comparison of the STFT and STRAIGHT spectral analysis

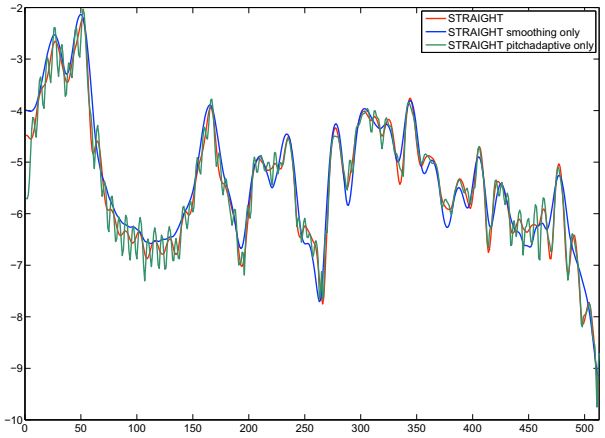

Figure 2: A comparison of the STRAIGHT spectral analysis with pitch adaptive only and smoothing only

16h from the AMI corpus [7]. The testing set consists of the NIST Rich Transcription Spring 2004 evaluation data and contains 11 minutes excerpts from 8 meetings recorded using headset microphones in 4 different data collection sites (2 for each site: CMU, ICSI, LDC and NIST) [1].

Our ASR experiments were performed using an HMM-based speech recognition system with Gaussian mixture model (GMM) output distributions, using the Hidden Markov Model ToolKit (HTK) software [8]. The overall training and decoding structure was that developed for the AMI-ASR system [7]. The baseline acoustic models were trained on conventional MFCCs (computed using a 25ms window with a 10ms shift); we also trained models using STRAIGHT derived MFCCs, and MFCCs derived from STRAIGHT both using the smoothing part only and using the pitch adaptive analysis only. For each representation 12 cepstral coefficients plus the zeroth cepstral coefficient (C0) were estimated, and first and second derivatives were also computed, resulting in a 39-element feature vector (13 coefficients + 13 Δ + 13 $\Delta\Delta$). The acoustic models were state clustered cross-word triphones with 16 mixture components per state. VTLN was also performed during both training and testing and it was applied with cepstral mean and variance normalisation (CMN, CVN) both to the standard MFCC system and to the STRAIGHT derived MFCC systems.

[1] NIST RT04s website: www.nist.gov/speech/tests/rt/rt2004/spring/

		TOT	F	M
MFCC	M1	38.4	38.5	38.3
STRAIGHT MFCC	S1	39.3	38.3	39.7
STRAIGHT MFCC pitch adapt. only	S2	38.2	38.2	38.3
STRAIGHT MFCC smoothing only	S3	40.1	39.9	40.1
HLDA 78 to 39	M1 + S1	36.6	36.3	36.7
HLDA 78 to 39	M1 + S2	36.9	36.1	37.3
HLDA 78 to 39	M1 + S3	37.3	36.6	37.6
HLDA 39 to 39	M1	37.6	37.7	37.5
HLDA 39 to 39	S1	37.4	37.2	37.5
HLDA 39 to 39	S2	37.1	36.0	37.7
HLDA 39 to 39	S3	39.6	38.8	40.1

Table 1: Experiment on RT04seval testing set. From top to bottom: conventional MFCCs (M1); STRAIGHT MFCCs (S1); STRAIGHT MFCCs with pitch adaptive analysis only (no smoothing) (S2); STRAIGHT MFCCs with smoothing only (no pitch adaptative analysis) (S3)(where M1, S1, S2, S3 are all 39 dimensions); HLDA combination of M1 and S1, M1 and S2, M1 and S3 all reducing from 78 to 39 dimensions; HLDA 39 to 39 dimension projection of M1, S1, S2 and S3.

Conventional and STRAIGHT derived MFCCs systems were combined at a feature level using HLDA [9]. This is a generalisation of LDA, which assumes a different covariance matrix for each class. In our experiments we employed HLDA because this has given better results than LDA when a sufficient amount of data is available to estimate the statistics. In the experiments presented in this paper we used mixture components of monophone models as classes to estimate the HLDA transform.

4. Decoupling the pitch adaptive and the smoothing effect of STRAIGHT

As mentioned in section 2 the aim of the experiments described in the first part of this paper is to decouple the two effects of the STRAIGHT spectral analysis: the use of a pitch adaptive window on one side and the smoothing obtained through an interpolation of the partial information given by the pitch adaptive spectral analysis on the other side. The results of these experiments have been reported in table 1. First we observe that the pitch adaptive analysis without smoothing (S2) gives a small but not significant improvement over conventional MFCCs (M1) and an even bigger improvement on S1 (STRAIGHT derived MFCCs). This is particularly evident for female speakers while for male speakers there is a big improvement especially when compared to purely STRAIGHT derived MFCCs (S1). Smoothing is particularly bad for male speakers and this is also confirmed by the experiment on the use of the smoothing part only of STRAIGHT without pitch adaptive analysis (S3). The MFCCs extracted using the smoothing component only of STRAIGHT performed consistently worse than conventional MFCCs.

We also combined conventional MFCCs with the pitch adaptive only (M1+S2) and smoothing only (M1+S3)

STRAIGHT derived MFCCs using HLDA feature combination with monophone mixture components as classes reducing from 78 to 39 dimensions. While none of this combinations outperformed the combination of conventional and STRAIGHT derived MFCCs (M1+S1) overall, the combination with pitch adaptive only STRAIGHT derived MFCCs (M1+S2) gave better performances for female speakers (for which pitch adaptive analysis is more important). The combination with smoothing only STRAIGHT derived MFCCs (M1+S3) on the other hand gave a smaller improvement. This is further evidence that the complementarity between conventional and STRAIGHT derived MFCCs is arisen from the use of a pitch adaptive window by the latter.

5. Measuring the speaker independence of STRAIGHT derived features

One of the aims of using a pitch adaptive spectral representation for feature extraction is to obtain features which have increased speaker independence. Ideally we would like to have features which only vary across different classes and which have as little as possible variation across different speakers within the same class used for speech recognition.

The relationship between speaker normalisation techniques and LDA have been studied both in [10] where it was proposed to use an LDA based metric to measure the effectiveness of CMN and CVN and VTLN and in [11] where the importance of applying LDA on top of speaker normalised features (to achieve better class separability) has been demonstrated. Suppose each acoustic feature vector x_i is labelled according to the class j and the speaker s to which it belongs (the association of a particular frame to a class j can be done automatically by forced alignment). We can define the corresponding total number of feature vectors $x_i \in (j,s)$ as $N^{(j,s)}$, and $\hat{\mu}^{(j,s)}$ and $\hat{\Sigma}^{(j,s)}$ are the mean vector and covariance matrix respectively corresponding to class j and speaker s. We can also define the class specific total number of feature vectors $N^{(j)}$, the mean vector $\hat{\mu}^{(j)}$ and the class specific covariance matrix as:

$$\hat{\Sigma}^{(j)} = \frac{1}{N^{(j)}} \sum_{s \in S} N^{(j,s)} \hat{\Sigma}^{(j,s)} + \underbrace{\frac{1}{N^{(j)}} \sum_{s \in S} N^{(j,s)} (\hat{\mu}^{(j,s)} - \hat{\mu}^{(j)})(\hat{\mu}^{(j,s)} - \hat{\mu}^{(j)})^T}_{\hat{\Sigma}_{B^S_{(j)}}} \quad (1)$$

where $\hat{\Sigma}_{B^S_{(j)}}$ is the between speaker covariance for class j. The total within-class covariance is therefore due to two distinct components: the variance due to the classes themselves and the between speaker covariance:

$$\hat{\Sigma}_{wc} = \frac{1}{N} \sum_{j \in J} N^{(j)} \hat{\Sigma}^{(j)} = \underbrace{\frac{1}{N} \sum_{j \in J} \sum_{s \in S} N^{(j,s)} \hat{\Sigma}^{(j,s)}}_{\hat{\Sigma}^N_{wc}} + \underbrace{\frac{1}{N} \sum_{j \in J} N^{(j)} \hat{\Sigma}_{B^S_{(j)}}}_{\hat{B}^S} \quad (2)$$

where $\hat{B}^S$ is the total between speaker covariance and $\hat{\Sigma}^N_{wc}$ is the within class covariance matrix we would have if the features were ideally speaker independent. Therefore the total covariance $\hat{\Sigma}$ has a component dependent on inter-speaker variability and another one which would occur if the features were completely speaker independent too: $\hat{\Sigma} = \hat{\Sigma}_{bc} + \hat{\Sigma}^N_{wc} + \hat{B}^S$ (where $\hat{\Sigma}_{bc}$ is the between class covariance matrix). The goal of LDA is finding the projection θ which maximises the between class covariance and minimises the within class covariance in the projected space. The trace (the eigenvalues sum) of $\hat{\Sigma}^{-1}_{wc}\hat{\Sigma}_{bc} = (\hat{\Sigma}^N_{wc} + \hat{B}^S)^{-1}\hat{\Sigma}_{bc}$ can be considered as the LDA objective function. Saon et al. [11] argued that, since ideally the between-speaker covariance $\hat{B}^S$ should be zero for speaker normalised features, the LDA objective function for normalised features should always be higher than that of non normalised features. Unfortunately even using speaker normalisation techniques, the between-speaker covariance is not completely zero (for example coarticulation differences are not normalised by VTLN) and the LDA objective function can be used as a measure of speaker independence of the features.

We adopted the inter-speaker metric introduced in [10] where the trace of $\hat{\Sigma}_{bc}/\hat{B}^S$ is used to measure speaker normalisation effectiveness. Gaussian components of monophone models have been used as classes in order to maintain the same type of classes used in our HLDA combination experiments. We compared conventional MFCCs, STRAIGHT derived MFCCs without the smoothing, STRAIGHT derived MFCCs without the use of the pitch adaptive window and STRAIGHT derived MFCCs with both smoothing and the pitch adaptive window usage. We used the entire meeting training corpus described in section 3 which contains a total of 115 male and 49 female speakers. The results of these experiments, using 39 dimensional feature vectors (12 cepstral coefficients plus C0 plus Δs and $\Delta\Delta$s), are shown in figure 3. The trend of the trace measure shows 3 big humps (left part of figure 3) due to the different nature of cepstral coefficients and their first and second derivatives and to the fact that as lower order cepstral coefficients are more discriminative so are their corresponding first and second derivatives (the gradient is higher for lower order coefficients and their derivatives), while higher order cepstral coefficients are more noisy and therefore less discriminative; thus they have a corresponding eigenvalue which is smaller than that of lower order coefficients.

Looking at the magnified right part of figure 3 (which shows the trace trend for the first 12 cepstral coefficients only) we can observe that STRAIGHT derived MFCCs using the pitch adaptive windowing but without smoothing shows the higher inter-speaker independence. Pitch adaptive features are significantly more speaker independent than both conventional MFCCs and smoothing only STRAIGHT derived MFCCs. STRAIGHT derived features using the pitch adaptive component only are the most speaker invariant.

Saon at al. [11] argued that LDA gives better performance on features that are more speaker independent. HLDA transforms are estimated by maximising the likelihood of the original data given the estimated statistics with the objective function:

$$\log L(\mathbf{x};\mathbf{A}) = -\frac{nN}{2} + \sum_{j=1}^{J} \frac{N_j}{2} \log \left(\frac{(\det \mathbf{A})^2}{(2\pi)^n \prod_{k=1}^{p} a_k \hat{\Sigma}^{(j)} a_k^T \prod_{k=p+1}^{n} a_k \hat{\Sigma} a_k^T} \right), \quad (3)$$

where the transformation is from n to p dimensions. We have shown that the total covariance matrix $\hat{\Sigma} = \hat{\Sigma}_{bc} + \hat{\Sigma}_{wc}$ can be further decomposed into two parts: the covariance that would be obtained if the features were perfectly speaker normalised, and the between speaker covariance (equation 2). The per class covariance matrix $\hat{\Sigma}^{(j)}$ (equation 1) can be also split into a class-specific covariance and the between speaker covariance matrix for the class. Ideally, if the features were completely speaker

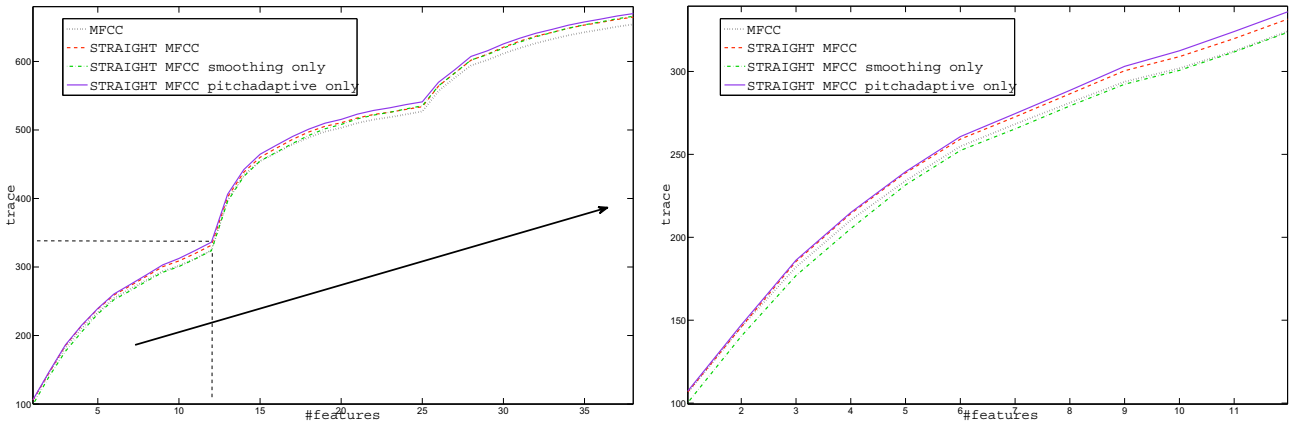

Figure 3: Trace measure as a function of the feature dimension measured using the whole meeting training data on 39 dimensions (on the left) and the magnified area representing the first 12 features (on the right)

normalised, the between speaker covariance would be null and therefore the likelihood of equation 3 would be greater for normalised features compared to features with some speaker dependence.

We applied HLDA directly on the 39 dimensional conventional MFCCs (M1) and STRAIGHT derived MFCCs (S1) with pitch adaptation only (S2) and smoothing only (S3) projecting to 39 dimensions. The results of this experiment are shown in the last 4 rows of table 1. The improvement obtained by the use of HLDA is bigger for pitch adaptive STRAIGHT derived MFCCs than for conventional MFCCs and smoothing only STRAIGHT MFCCs, We hypothesise that this is due to the better speaker independence of pitch adaptive features as shown similarly by Saon et al. for LDA applied on VTLN features [11].

6. Conclusions

A STRAIGHT based pitch adaptive spectral representation has been successfully applied to extract acoustic features for a challenging LVCSR task, multiparty conversational speech in the meeting domain. The combination with conventional MFCCs using HLDA was particularly beneficial yielding consistent improvements over conventional features alone. In this paper the two key components of STRAIGHT, pitch adaptive analysis and smoothing through interpolation, have been studied independently. Experimental results showed that adopting pitch adaptive features can significantly improve speech recognition performances. Non smoothed pitch adaptive features outperformed smoothed non pitch adaptive features, when combined with conventional MFCCs. This improvement is principally due to the adoption of a pitch adaptive representation. The use of a pitch adaptive representation is particularly beneficial for female speakers, because for high pitched speakers the Mel filters are not broad enough to remove the horizontal spectral lines due to the pitch interference.

We have also measured the speaker independence of all the features adopted in this study. Using an LDA based metric we found evidence that the pitch adaptive features are more speaker independent than conventional MFCCs. We observed that the improved speaker independence has the desirable effect of making HLDA more effective and making STRAIGHT derived features more suitable for this technique than conventional features.

7. Acknowledgments

This work is supported by the EU 6th FWP IST Integrated Project AMIDA (Augmented Multiparty Interaction with Distant Access IST FP6-033812, publication AMIDA–100). This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein. Moreover the authors would like to thank Prof. Hideki Kawahara for giving us the opportunity to use the STRAIGHT code, and the members of the AMI-ASR team.

8. References

[1] M. Magimai-Doss, T. A. Stephenson, S. Ikbal, and H. Bourlard, "Modelling auxiliary features in tandem systems," in *Proc. ICSLP*, 2004.

[2] B. Bozkurt and L. Couvreur, "On the use of phase information for speech recognition," in *Proc. EUSIPCO*, 2005.

[3] W. J. Holmes, "Improving the representation of time structure in front-ends for automatic speech recognition," in *Proc. ICSLP*, 2000.

[4] G. Garau and S. Renals, "Combining Spectral Representations for Large Vocabulary Continuous Speech Recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, pp. 508–518, March 2008.

[5] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using pitch adaptive time-frequency smoothing and instantaneous-frequency-based F0 extraction: possible role of repetitive structure in sounds," *Speech Communication*, 1999.

[6] T. Hain, P.C. Woodland, T.R. Niesler, and E.W.D. Whittaker, "The 1998 HTK system for transcription of conversational telephone speech," *Proc. IEEE ICASSP*, 1999.

[7] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, J. Vepa, and V. Wan, "The AMI system for the transcription of speech in meetings," in *Proc. IEEE ICASSP*, 2007.

[8] S. Young et al., *The HTK book (v3.4)*, Cambridge University Engineering Department, December 2006.

[9] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved recognition," *Speech Communication*, vol. 26, pp. 283–297, 1998.

[10] R. Haeb-Umbach, "Investigations on Inter-Speaker Variability in the Feature Space," in *Proc. ICASSP*, 1999, vol. 1, pp. 397–400.

[11] G. Saon, M. Padmanabhan, and R. Gopinath, "Eliminating Inter-Speaker Variability Prior to Discriminant Transforms," in *Proc. ICASSP*, 2002, vol. 1, pp. 73–76.