# Combining Spectral Representations for Large Vocabulary Continuous Speech Recognition

Giulia Garau*, and Steve Renals, *Member, IEEE*

***Abstract*—In this paper we investigate the combination of complementary acoustic feature streams in large vocabulary continuous speech recognition (LVCSR). We have explored the use of acoustic features obtained using a pitch-synchronous analysis, STRAIGHT, in combination with conventional features such as mel frequency cepstral coefficients. Pitch-synchronous acoustic features are of particular interest when used with vocal tract length normalisation (VTLN) which is known to be affected by the fundamental frequency. We have combined these spectral representations directly at the acoustic feature level using heteroscedastic linear discriminant analysis (HLDA) and at the system level using ROVER.**

**We evaluated this approach on three LVCSR tasks: dictated newspaper text (WSJCAM0), conversational telephone speech (CTS), and multiparty meeting transcription. The CTS and meeting transcription experiments were both evaluated using standard NIST test sets and evaluation protocols. Our results indicate that combining conventional and pitch-synchronous acoustic feature sets using HLDA results in a consistent, significant decrease in word error rate across all three tasks. Combining at the system level using ROVER resulted in a further significant decrease in word error rate.**

***Index Terms*—LVCSR, VTLN, pitch-synchronous, feature combination, HLDA, ROVER, STRAIGHT.**

**EDICS Category: SPE-RECO**

## I. INTRODUCTION

THE combination of multiple acoustic feature streams has the potential to improve the accuracy of automatic speech recognition (ASR) [1]–[5]. Different acoustic representations have different strengths, and thus will tend to result in ASR systems that make different errors. The combination of acoustic feature representations is a way to exploit complementary information and to take advantage of the strengths of particular representations. In this paper we investigate the combination of conventional acoustic features, such as mel frequency cepstral coefficients (MFCCs), in combination with features obtained using a pitch-synchronous analysis for large vocabulary continuous speech recognition (LVCSR).

LVCSR systems typically include a speaker normalization component, such as vocal tract length normalization (VTLN) [6]–[9], in which a transform is inferred to make the feature vectors for a target speaker appear close to those of an "average" speaker. In the case of VTLN, this transformation often takes the form of a piecewise linear warping of the frequency axis parameterised by a warping factor. Such a frequency warping factor is known to be affected by the fundamental frequency [10], [11] as well as vocal tract size. It is therefore of interest to explore the use of a pitch-synchronous analysis. As discussed in section II, pitch-synchronous representations have been investigated in the context of speaker recognition and for small vocabulary ASR. However, investigation of pitch-synchronous representations for LVCSR has been very limited.

We have explored the use of spectral representations derived from STRAIGHT, a pitch-synchronous analysis developed by Kawahara [12], reviewed in section III. This analysis results in a smoothed time-frequency representation from which it is possible to extract MFCCs and mel frequency perceptual linear prediction (MF-PLP) cepstral coefficients. We have combined these pitch-synchronous acoustic representations with conventional representations both at the feature level using heteroscedastic linear discriminant analysis (HLDA) and at the decoding level using the ROVER technique to combine the outputs of multiple decodings (section IV).

In section V we report on experiments using these combined spectral representations on three LVCSR tasks: transcription of dictated newspaper text (WSJCAM0); conversational telephone speech (CTS) recognition; and transcription of multiparty meetings using both close-talking and distant microphones. This set of experiments has allowed us to test the approach in a range of speaking styles and channel conditions. Although, the WSJCAM0 task consists of read speech using a close-talking microphone in a quiet environment, the other two tasks are more challenging. Both are concerned with spontaneous conversational speech. Moreover, CTS involves telephone speech which is subject to a bandpass filter that partly obscures the pitch, while the multiparty meetings were recorded in reverberant conditions with overlapping speakers. The situation is further complicated for the meeting task when multiple distant microphones are used to record the speech, and beamforming algorithms are applied to the recorded signals.

The results of our experiments indicate that combining conventional and STRAIGHT-based acoustic features using HLDA results in a consistent relative decrease in the word error rate of 3–9% across all three domains, with the largest relative reductions observed for the telephone speech and distant microphone tasks. A further 8% relative reduction in word error rate was observed when ROVER combination was applied to the meeting transcription task.

## II. PITCH-SYNCHRONOUS ANALYSIS

The short time Fourier transform (STFT) involves the computation of a separate Fourier transform for each frame of

This work was partly supported by the EU 6th FWP IST Integrated Project AMIDA (Augmented Multiparty Interaction with Distance Access IST FP6-033812, publication AMIDA–34).

The authors are with the Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9LW United Kingdom (email: {g.garau,s.renals}@ed.ac.uk).

the signal waveform under a sliding window. This process is affected by the uncertainty principle, which states that it is impossible to have an arbitrary resolution in both time and frequency [13]. The effect of this physical law is that the use of a long window in time (longer than two fundamental periods of the signal) leads to a good resolution in frequency and poorer time resolution, whereas a short window in time leads to the converse, good time resolution at the cost of frequency resolution. For speech the fundamental frequency of the signal varies over time, and if a fixed size window is applied, then its effect will be evident on the spectrum, particularly for high pitch speakers. This effect will be apparent even after the application of a mel-scaled filterbank, in which the standard filter bandwidth in the lower frequency region is usually around 200–300 Hz. This is not broad enough to remove the harmonic structures for high pitched speakers, usually females, although it is able to provide a smooth representation for males [14]. It is therefore of interest to investigate the use of a pitch-synchronous window that adapts according to the current estimate of the fundamental frequency.

In speech synthesis and speech coding, where it is important to generate the correct fundamental frequency, pitch-synchronous analyses have been well studied (e.g., [15]). The use of pitch-synchronous features has also been investigated for speaker recognition. Voice source information, as manifest in the pitch, is a speaker-specific characteristic, and source features derived from a pitch-synchronous analysis have been proposed as features for speaker recognition [16], [17]. Zilca et al [18] proposed a pitch-adaptive analysis, referred to as "depitching", which attempts to filter out pitch information from the speech signal. Although depitched features alone resulted in lower accuracy for speaker recognition, combining systems using conventional and depitched MFCCs resulted in a significant improvement, with a more uniform error distribution across speakers.

The fundamental frequency provides prosodic information and information about the speaker but, for non-tonal languages, pitch is not used to encode words and phonemes. Therefore, factoring out the pitch information in speech recognition should result in a system with greater speaker independence. Two basic approaches have been reported in the literature: the use of pitch-synchronous acoustic features, and acoustic models in which the pitch is explicitly modelled as a variable. An example of the latter approach [19] uses dynamic Bayesian networks (DBNs) in which the variables corresponding to the MFCCs are conditioned on the pitch, although this did not result in a significant improvement in accuracy.

Bozkurt et al. [20] investigated a pitch-synchronous analysis based on group delay features (the negative of the differential phase spectrum) extracted using a window centered at the glottal closure instant, from which a phase spectrum was computed. Applying these features to ASR, in combination with MFCCs, resulted in a significant increase in accuracy over a baseline MFCC system on the AURORA–2 corpus. Holmes [21] proposed the use of "excitation synchronous" windows for the extraction of MFCCs. In comparison with features extracted using "fixed interval" windows, a significant improvement was observed on a digit recognition task. An alternative pitch-adaptive representation, pitch synchronous zero crossing peak-amplitude (PS-ZCPA), has also shown some promise in reducing errors on noisy speech (the AURORA–2J corpus) [22].

Irino et al. [23] employed the pitch-synchronous STRAIGHT representation, discussed in the next section, using it as the underlying spectral representation for the extraction of MFCCs. STRAIGHT-based MFCCs were compared with conventional MFCCs in HMM-based speech recognition on a small database, but no significant improvement in accuracy was observed. In this work, we explore the use of STRAIGHT-based acoustic features, in conjunction with speaker normalisation using VTLN, and in combination with conventional MFCC and MF-PLP features.

## III. STRAIGHT-BASED FEATURES

STRAIGHT [12] is a vocoder consisting of analysis and synthesis parts. The spectral analysis of STRAIGHT uses a pitch-adaptive window which gives equivalent resolution both in 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 interference arising from signal periodicity.

We derived STRAIGHT-based MFCCs by replacing the classic STFT, which typically uses a Hamming window, with the STRAIGHT spectral analysis using a window that is Gaussian both in time and frequency:

$$w(t) = \frac{1}{\tau_0} \exp\left(-\pi(t/\tau_0)^2\right) \tag{1}$$

$$W(\omega) = \frac{\tau_0}{\sqrt{2\pi}} \exp\left(-\pi(\omega/\omega_0)^2\right) . \tag{2}$$

This window was chosen by Kawahara et al. [12] because of its isometric properties (it is the only smooth non-zero function which transforms to itself) and its unique property of minimum time-bandwidth product. The shape of the window depends on the estimated fundamental frequency $f_0 = 1/\tau_0 = 2\pi/\omega_0$. If we compare it with a 25 ms Hamming window: for $f_0 \cong 80$ Hz they are almost equivalent; while for $f_0 < 80$ Hz the pitch synchronous window gives a better frequency resolution and lower temporal resolution; and for $f_0 > 80$ Hz it provides a better temporal resolution and lower frequency resolution.

The value of $f_0$ used for the window computation can be estimated using various algorithms. TEMPO, the algorithm for pitch tracking provided in the STRAIGHT framework [12], is based on the use of the so-called *fundamentalness* measure, obtained using a wavelet Gabor filter designed to highlight the fundamental frequency (maximal filter output) and to reject harmonic replicas. However, other pitch trackers may be used and most of our experiments employed the RAPT pitch tracking algorithm [24][1] which is based on cross-correlation in the time domain. As discussed further in section V, although no significant difference between the use of the two pitch trackers was found when using clean read speech, RAPT

[1]Implemented as ESPS get_f0, available from: www.speech.kth.se/snack/

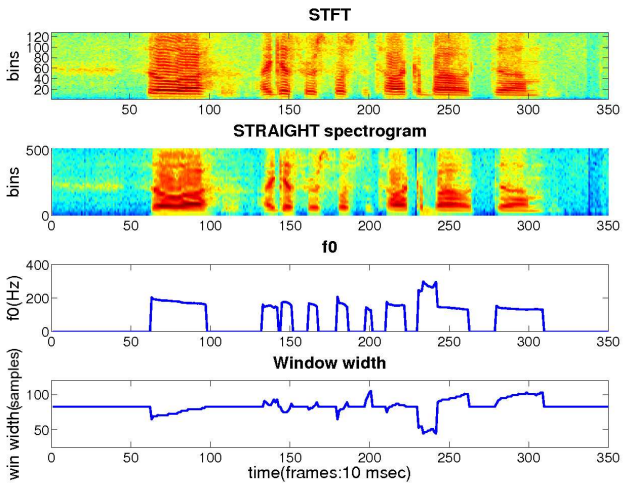


Fig. 1. Example of STFT spectrogram, STRAIGHT spectrogram, f0 and specral analysis window width in the time domain for a telephone speech signal, with a sample rate of 8 kHz.

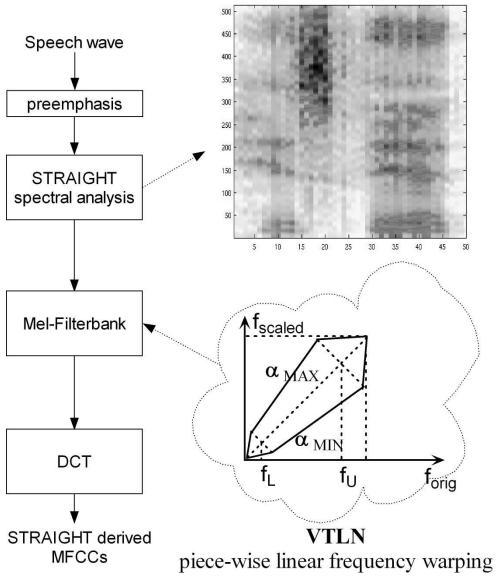


Fig. 2. A block diagram of STRAIGHT MFCCs extraction with VTLN frequency warping

proved to be more reliable for conversational telephone speech, as well as being more computationally efficient.

The STRAIGHT pitch spectrogram of a telephone speech signal is compared with a conventional STFT spectrogram in figure 1. The harmonic structure, visible in the STFT, is not present in the smoother STRAIGHT spectrogram. The lower part of the figure shows the pitch value plotted along with the width of the analysis window in the time domain (measured at $1/3$ of the height of the window in number of samples), illustrating how the spectrogram resolution follows the value of the fundamental frequency of the signal. A reliable pitch estimate is important, since pitch tracking errors such as pitch doubling can lead to a very wide window in the frequency domain and poor spectral resolution. For unvoiced speech a default value of about 10 ms was used for the window width (measured at $1/3$ of the maximum window amplitude), corresponding to a fundamental frequency of 160 Hz.

Figure 2 shows a block diagram of the extraction procedure for STRAIGHT derived MFCCs. The log STRAIGHT (power) spectrogram is processed through a mel scaled filterbank and decorrelated using the discrete cosine transform. This is similar to the feature extraction process presented in [23] but here we perform a normal DCT instead of a warped DCT because we do not require feature inversion. MF-PLPs have also been extracted from the log STRAIGHT spectrogram, by mel scaling, followed by equal loudness pre-emphasis, cube root compression and linear predictive cepstral analysis.

In addition, we have employed a VTLN frequency warping procedure, shown in figure 2. The centres of the filters of the mel scaled filterbank are moved according to a piecewise linear frequency warping function where different warping factors $\alpha$ are defined for different frequency bandwidths (depicted in the VTLN box in figure 2). This takes into account the inverse proportionality between formant positions and the length of the vocal tract, such that a change of scale by a factor of $\alpha^{-1}$ results in a scaling of the frequency axis by a factor $\alpha$. The warping factors are estimated using maximum likelihood in the acoustic model training process [8], the speaker-specific warp factor $\alpha$ being set to maximise the likelihood of the normalised acoustic observation feature vectors $X^{\alpha}$, given a transcription $W$ and an acoustic model $\lambda$ [8], [9]. This approach is consistent with the overall optimization of the acoustic models, and has proven to be very effective in LVCSR.

An exhaustive search for the optimal warping factor for a speaker would be computationally expensive; however, it has been experimentally observed that the log-likelihood $\log p(X^{\alpha} \mid \lambda, W)$ has a parabolic behavior with respect to $\alpha$. Therefore a one-dimensional Brent search was used to find the maximum of this curve.

## IV. FEATURE COMBINATION

Different acoustic representations have different strengths and weaknesses for ASR. Approaches to combine representations, at the feature, model and system level, have proven to be effective in reducing the word error rate. Feature combination may be carried out directly at the feature vector level by concatenating feature vectors, followed by a dimension reducing transform such as linear discriminant analysis (LDA) or heteroscedastic LDA (HLDA) [25], indirectly at the model level [1], [3], or as a postprocessing procedure applied to the outputs of multiple recognizers [26]. As mentioned in section II the combination of pitch-synchronous and conventional features at the decoding level has been shown to be effective for speaker and speech recognition [16], [18], [20].

The simplest form of direct feature combination involves the concatenation of the acoustic feature vectors. This approach has a number of drawbacks including a substantial increase in the dimensionality of the feature space to be modelled, and the introduction of strong correlations between components in the concatenated vector, which can cause problems for acoustic models based on diagonal covariance Gaussians. Both

these problems are addressed through the use of dimension reducing, decorrelating transforms such as LDA, HLDA and principal components analysis (PCA). PCA estimates a global transform, and has been found to be much less well-suited to the task compared with LDA and HLDA, which allow the decorrelating transforms to be estimated on a per-class (or per-state) basis.

Zolnay et al [3] have demonstrated that discriminant feature-level combination may be nested successfully inside a model-based combination approach, and this has resulted in reduced word error rates for two LVCSR tasks, VerbMobil-II and the European Parliamentary Plenary Sessions corpus. More recent work by this group [4], involving the investigation of auditory-inspired features from a gammatone filterbank, have indicated that a system level combination using ROVER [26] results in a significant reduction in word error rate.

### A. HLDA

In our experiments, we have performed feature-level combination using HLDA, a generalisation of LDA. HLDA enables the derivation of a linear projection that decorrelates concatenated feature vectors, and performs a dimensionality reduction. In both HLDA and LDA, each feature vector that is used to derive the transformation is assigned to a class. Since one of the goals of these techniques is to improve the discrimination between the classes used during decoding, HLDA and LDA classes are typically HMM states or mixture components, obtained using Viterbi alignment.

Hunt [27] proposed the use of LDA to improve discrimination between syllables, and in later work used LDA to combine feature streams from an auditory model front end [28]. Given an $n$ dimensional feature vector $\mathbf{x}$ the goal of LDA is to find a linear transformation $\theta^T : \Re^n \rightarrow \Re^p$ with $p \leq n$ such as to project $\mathbf{x}$ in a $p$ dimensional space according to $\mathbf{y} = \theta^T \mathbf{x}$. The transform is chosen to maximise the between class covariance $\Sigma_b$ and to minimise the within class covariance $\Sigma_w$, using the eigenvectors corresponding to the $p$ largest eigenvalues of $\Sigma_b \Sigma_w^{-1}$.

LDA makes two assumptions: first, all the classes follow a multivariate Gaussian distribution; second, they share the same within-class covariance matrix. HLDA (introduced by Kumar and Andreou [29]) relaxes the second assumption and may be considered as a generalisation of LDA. In HLDA, the optimal transformation matrix $\mathbf{A}$ is found by maximising the likelihood of the original data $\mathbf{x}$

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

where $\hat{\Sigma}$ and $\hat{\Sigma}^{(j)}$ are the global and per class covariance matrix estimates respectively, and $N$ and $N_j$ are the total and per class number of training vectors. Since the maximisation of (3) has no closed-form solution, an iterative algorithm is employed. We have used a method implemented by Burget [25], [30], inspired by the approach proposed by Gales [31] in which the transform matrix $\mathbf{A}$ is computed by periodically reestimating individual rows $\mathbf{a}_k$ as follows:

$$\hat{\mathbf{a}_k} = \mathbf{c}_k \mathbf{G}^{(k)-1} \sqrt{\frac{N}{\mathbf{c}_k \mathbf{G}^{(k)-1} \mathbf{c}_k^T}}. \quad (4)$$

$\mathbf{c}_i$ is the $i$th row vector of co-factor matrix $C = |\mathbf{A}|\mathbf{A}^{-1}$ for the current estimate of $\mathbf{A}$ and

$$\mathbf{G}^{(k)} = \begin{cases} \sum_{j=1}^{J} \frac{\gamma_j}{\mathbf{a}_k \hat{\Sigma}^{(j)} \mathbf{a}_k^T} \hat{\Sigma}^{(j)} & k \leq p \\ \frac{N}{\mathbf{a}_k \hat{\Sigma} \mathbf{a}_k^T} \hat{\Sigma} & k > p. \end{cases} \quad (5)$$

$\gamma_j$ is the number of training feature vectors belonging to the $j$th class.

The main characteristic which sets apart HLDA from LDA is the requirement to estimate a different covariance matrix for each class. In LDA the within class covariance matrix is approximately the weighted sum of the individual HLDA class covariance matrices. A minimum amount of in–class data is necessary to find reliable estimates for the individual HLDA covariance matrices. Therefore, in order to avoid data sparsity, the type of classes used to estimate the HLDA transformation matrices should be carefully considered. We experimented with two possible choices of classes (section V): (1) classes corresponding to the HMM triphone states of our models; (2) Gaussian mixture components of monophone models.[2]

### B. System-level combination

In addition to feature-level combination, we also explored the use of system-level combination using ROVER [26], a technique to combine the output of multiple speech recognition systems. In ROVER, the transcriptions are first compared by aligning them using dynamic programming to minimise the number of substitutions, deletions and insertions. This alignment depends on the word sequence chosen as the reference.

The multiple alignments are then combined using a voting approach, performed either by choosing the most frequently recognised hypothesis (majority voting) or by selecting the hypothesis with the highest confidence score (maximum confidence score voting). The choice of the voting criteria is not limited to these two techniques and any approach able to disambiguate between multiple transcriptions can be adapted [5]. It is also possible to obtain a lower bound on the word error rate achievable by ROVER, by using an oracle combination in which the closest available word sequence to the correct transcription is selected. A disadvantage of ROVER is the need to train and decode each component system separately, in contrast to HLDA which requires a single decoding pass.

## V. EXPERIMENTS

VTLN attempts to normalise for the variation of the vocal tract length across different speakers, which is approximately constant over time. In a previous study about the use of VTLN

[2]Monophone models are estimated as part of the triphone training process.

for multiparty meetings, we found that the VTLN warping factors estimated using ML exhibited significant variability over time [32]. This variation was partly explained by the fact that warping factor estimates were correlated with pitch. It is therefore of interest to investigate the use of a spectral representation which is less pitch-dependent, in conjunction with VTLN.

We expect VTLN to benefit from the smoother pitch independent spectral representation provided by STRAIGHT. The main goal of the experiments described below, is the exploitation of this representation in a range of LVCSR tasks. In particular we hypothesise that female speakers, with a higher fundamental frequency, will benefit the most from a pitch synchronous representation, since for these speakers the Mel filter bandwidths are not sufficiently wide to smooth the harmonic lines due to pitch interference. However, there are disadvantages to the STRAIGHT representation. STRAIGHT provides a smoother pitch synchronous spectral representation, that is sensitive to pitch tracking errors and may be less informative than the conventional STFT, owing to over-smoothing.

Given these advantages and disadvantages, we have performed extensive experiments in which the feature streams are combined. Recent experience in LVCSR has indicated that while it is rarely straightforward to obtain significant and consistent speech recognition accuracy gains from novel features, it may be possible to obtain consistent improvements by combining conventional and novel features. This has been the case for gammatone features [4] and for features based on posterior probability estimates [2], as well as for pitch synchronous features [16], [18], [20].

We have used HLDA to combine feature streams. Schlüter et al. [33] have argued that numerical problems can arise when strongly correlated features are combined using LDA. Such problems did not arise in our experiments, since the feature streams are not highly correlated due to the different analysis windows employed. In addition to HLDA, system level combination experiments were performed using majority voting ROVER.

### A. Experimental setup

Our ASR experiments have been performed using an HMM-based speech recognition system with Gaussian mixture model (GMM) output distributions, using the Hidden Markov Model ToolKit (HTK) software [34]. The overall training and decoding structure was that developed for the AMI-ASR system [35]. The baseline acoustic models were trained on conventional MFCCs (computed using a 25ms window with a 10ms shift); for each domain we also trained models using STRAIGHT derived MFCCs. 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 $\Delta$ + 13 $\Delta\Delta$). The acoustic models were state clustered cross-word triphones with 16 mixture components per state. We also performed VTLN during both training and testing, using an iterative method which alternated the estimation of warping factors and the estimation of acoustic model parameters, described in detail in [32]. VTLN was applied both to the standard MFCC system and to the STRAIGHT derived MFCC system.

We carried out a number of experiments to determine the sensitivity of the STRAIGHT-based features to the pitch tracking algorithm that was used. An initial set of experiments employed the Keele pitch extraction reference corpus [36]. This corpus features ten British English speakers reading a phonetically-balanced story, for which the fundamental frequency ground truth was obtained from a laryngograph signal. The corpus is not large enough to re-estimate the acoustic models, and it is from a different domain to any of the domains studied here. Since it consists of British English read speech, we used WSJCAM0 acoustic and language models (described in detail in section V-B) to automatically transcribe it. The use of these models, which were not well-matched to the domain of the Keele corpus, resulted in rather high word error rates (over 40%): there was no available development data to adapt the models to this domain. We extracted STRAIGHT derived MFCCs both using the reference pitch, and the TEMPO and the RAPT pitch trackers, observing less than 1% difference in word error rate between features using the ground truth pitch track (43.6%), versus features using the TEMPO or RAPT algorithms (both 44.7%). Although there is a small, but significant, improvement in using the reference pitch tracks, we conclude that both of the automatic pitch tracking algorithms offer acceptable accuracy. Although training with reference pitch tracks might result in further improvements, a database suitable for speech recognition with laryngograph signals is not available.

For this data, and for WSJCAM0, the ASR performance for systems using TEMPO and RAPT was almost identical. For the CTS domain we observed that RAPT resulted in significantly lower word error rates compared with TEMPO (see table III). Since RAPT also has lower computational demands, we used this pitch tracker for all our experiments (except where stated otherwise).

### B. WSJCAM0

Our first set of experiments was performed on the WSJCAM0 corpus [37], recorded at Cambridge University, and consisting of native British English read speech, using text from the Wall Street Journal (WSJ0) corpus. WSJCAM0 was recorded in an acoustically isolated room with head-mounted microphones, and has a training set (si_tr) consisting of 7 861 utterances, corresponding to around 15 hours of speech, spoken by 39 female and 53 male speakers. We tested on the 20 000 words "open vocabulary" task development set (si_dt20a) which has 10 female and 10 male speakers (consisting of about 41 minutes of speech). We used the standard MIT Lincoln Labs 20k Wall Street Journal trigram language model.

Table I shows our baseline results for this corpus. The top four lines show the word error rates for the conventional and STRAIGHT-based MFCC systems, with and without VTLN. The conventional system has a lower word error rate than the STRAIGHT-based system, with the difference between the

TABLE I
WORD ERROR RATES ON THE WSJCAM0 SI_DT20A DATASET, COMPARING CONVENTIONAL AND STRAIGHT-BASED MFCCS, WITH AND WITHOUT VTLN. THE COMBINED SYSTEM (BOTTOM LINE) USED CONCATENATED FEATURE VECTORS WITH NO DIMENSION REDUCTION.

| | Dimension | Total | Female | Male |
|---|---|---|---|---|
| STD MFCCs | 39 | 13.2 | 12.8 | 13.5 |
| STRAIGHT MFCCs | 39 | 14.4 | 13.7 | 15.2 |
| STD MFCCs + VTLN | 39 | 12.5 | 12.0 | 13.0 |
| STRAIGHT MFCCs + VTLN | 39 | 13.0 | 12.5 | 13.5 |
| STRAIGHT + STD MFCCs + VTLN | 78 | 15.4 | 15.2 | 15.7 |

TABLE II
ERROR RATES AFTER COMBINING CONVENTIONAL AND STRAIGHT DERIVED MFCCS USING HLDA, TESTING ON WSJCAM0 SI_DT20A. THE XWRD/STATES CONDITION INDICATES THAT THE STATES OF CROSS-WORD TRIPHONE MODELS ARE USED AS HLDA CLASSES; THE MONO/COMPONENTS CONDITION INDICATES THAT GAUSSIAN COMPONENTS OF MONOPHONE MODELS ARE USED AS HLDA CLASSES.

| Dimension | HLDA content/classes | Total | Female | Male |
|---|---|---|---|---|
| 52 | xwrd/states | 12.3 | 11.9 | 12.8 |
| 39 | xwrd/states | 12.4 | 12.1 | 12.7 |
| 52 | mono/components | 12.3 | 11.9 | 12.8 |
| 39 | mono/components | 12.1 | 11.4 | 12.8 |

two reduced by half in the case of VTLN. The final row of the table shows the baseline feature combination experiment, in which the two feature vectors are simply concatenated at each frame, ending up with a 78-element feature vector. This resulted in a considerable increase to the word error rate, as might be expected. To minimise the correlations within the combined feature vector, and to reduce the overall dimensionality, we applied HLDA to the concatenated features. Table II summarises the main results of these experiments, in terms of the word error rates with respect to the reduced dimensionality and the choice of class in the HLDA.

The upper part of table II (*xwrd*) shows the results obtained when the HLDA statistics were estimated using the states of the cross-word triphone HMMs, a total of 1 927 classes. The lower part (*mono*) shows the results obtained using monophone mixture components as classes — 2 208 in total (46 phones, 3 states/phone, 16 gaussians/state). The *xwrd* condition is more focused on discriminating between triphone states, allowing consistency between the HLDA classes and the acoustic triphone models (used during recognition). On the other hand the *mono* condition, using mixture components as classes, ensures that the distribution of the feature vectors corresponding to each class are more gaussian. Once the 78 dimension features were projected and decorrelated in the HLDA feature space, a complete training from scratch—following exactly the same procedure used for the single feature stream systems—was performed, obtaining state clustered cross-word triphone models. For each HLDA class type, we experimented with different dimension reductions, with the best results being obtained with a reduction from 78 to 39 dimensions. For comparison we also show results using 52 dimensions. The best results were achieved using monophone state mixture components as classes, yielding 3.2% relative improvement compared with the baseline standard MFCC system. We also performed experiments using LDA and smoothed HLDA [25], with HLDA consistently performing at least as accurately as the other approaches.

### C. Conversational Telephone Speech

The next set of experiments used CTS data, based on a 72 hour training set containing 57 hours from Switchboard–1, 8 hours from Switchboard–2, and 7 hours from the Call Home English corpus. This training set, a subset of a training set we have previously used [38], was prepared such that each of the three parts had equal numbers of male and female speakers. Our test set was the NIST Hub5 Eval01 evaluation set[3] consisting of around 6 hours of speech in total, equally distributed between Switchboard–1 (SW1), Switchboard–2 (S23) and Switchboard-cellular (Cell), comprising 60 male and 60 female speakers.

We used clustered cross-word triphone acoustic models with about 5 400 tied states. For this task we conducted several experiments in which we compared the accuracies of systems using conventional and STRAIGHT derived MFCCs, with and without cepstral mean and variance normalisation (CMN/CVN), and with and without VTLN. We also compared the use of the TEMPO and RAPT pitch trackers for STRAIGHT, in this case on systems without normalisation (no CMN/CVN and no VTLN). We used the same trigram language model in all cases, with a vocabulary of 50 000 words, trained on various additional sources including web data, broadcast news transcripts and email text [38].

Word error rates for various configurations are shown in table III. The first three rows show results in the case of no normalisation, including a comparison between TEMPO and RAPT pitch trackers for STRAIGHT. Conventional MFCCs result in the best performance, and RAPT gives a significant decrease in word error rate of 4% relative compared with TEMPO. We note that pitch tracking telephone speech is significantly more challenging owing to the bandpass filtering and other channel effects [39]. Applying CMN/CVN and VTLN results in a decrease in word error rate by over 10% for both conventional and STRAIGHT-based systems. As in the WSJCAM0 task, the gap between conventional and STRAIGHT-based systems is considerably reduced when VTLN is applied: indeed, there is no significant difference in error rate between the normalised conventional and STRAIGHT-based systems on CTS. This is evidence that the smoother spectral representation offered by STRAIGHT is well-matched to VTLN, which uses frequency warping to normalise speech to increase speaker independence.

We combined the two normalised systems using HLDA both using triphone states and monophone mixtures as classes. Each combination yielded an 8% relative improvement compared to the baseline, a conventional MFCC system with VTLN and CMN/CVN. The improvements are consistent for both female and male speakers and for all the testing subsets. This is a significant result, since the baseline system is strong, given the training set of 72 hours, and the fact that additional techniques such as maximum likelihood linear transforms and discriminative training are not applied.

[3]http://www.nist.gov/speech/tests/ctr/h5_2001/index.htm

TABLE III
WORD ERROR RATES ON THE CTS NIST HUB5 EVAL01 DATA FOR CONVENTIONAL AND STRAIGHT DERIVED MFCCS, AND THEIR COMBINATION USING HLDA. TEMPO AND RAPT PITCH TRACKERS ARE COMPARED FOR STRAIGHT FEATURES (LINES 2–3). BOTH TRIPHONE STATES AND MONOPHONE MIXTURE COMPONENTS ARE USED AS HLDA CLASSES FOR A FEATURE REDUCTION FROM 78 TO 39 DIMENSIONS (LINES 6–7). CMN AND CVN ARE CEPSTRAL MEAN AND VARIANCE NORMALISATIONS.

| | TOTAL | Female | Male | SW1 | S23 | Cell |
|---|---|---|---|---|---|---|
| MFCC (no CMN/CVN) | 42.7 | 41.8 | 43.6 | 36.5 | 43.3 | 47.9 |
| STRAIGHT (TEMPO no CMN/CVN) | 47.6 | 46.0 | 49.1 | 40.7 | 49.0 | 52.8 |
| STRAIGHT (RAPT no CMN/CVN) | 45.7 | 44.5 | 46.9 | 40.0 | 46.6 | 50.3 |
| MFCC+CMN/CVN+VTLN | 37.6 | 37.0 | 38.3 | 31.8 | 37.1 | 43.5 |
| STRAIGHT (RAPT) +CMN/CVN+VTLN | 39.2 | 38.2 | 40.1 | 33.6 | 39.0 | 44.5 |
| MFCC + STRAIGHT (RAPT) +CMN/CVN+VTLN+HLDA(xwrd) | 34.6 | 33.6 | 35.6 | 28.3 | 34.5 | 40.5 |
| MFCC + STRAIGHT (RAPT) +CMN/CVN+VTLN+HLDA(mono) | 34.7 | 33.8 | 35.6 | 28.6 | 34.7 | 40.5 |

### D. Multiparty meetings

Our final, and most extensive, set of experiments was in the domain of multiparty meetings. For this task the training set, which was the same used for the AMI-ASR systems [40] in the NIST RT05 and RT06 evaluations [41], consisted of a total of over 100 hours of conversational meeting speech from four corpora of multiparty meeting recordings: 70 hours from the ICSI corpus, 13 hours from the NIST corpus, 10 hours from the CMU-ISL corpus and 16 hours from the AMI corpus, with 115 male and 49 female speakers. The testing set consisted of the NIST Rich Transcription Spring 2004 evaluation set[4] and is composed of about 100 minutes excerpted from 8 meetings recorded in four different data collection sites (CMU, ICSI, LDC and NIST).

The NIST meeting recognition evaluation has two principal testing conditions, individual headset microphone (IHM) and multiple distant microphones (MDM). We conducted experiments using both conditions, training separate acoustic models for the each condition. For the MDM task, the speech is recorded using a number of microphones placed in the meeting room. The microphone positions, which were not provided, varied depending on the site where the data were collected. The additional processing in the MDM system included Wiener filtering of each distant channel, estimation of the energy scaling factor and of the delay of each channel by generalised cross correlation with respect to a given reference channel, and the use of these parameters to perform delay and sum beamforming [35].

We used clustered cross-word triphone acoustic models with 16 mixture components per state and around 6 600 tied states in total, and trained a set of models for each condition using VTLN. We used a vocabulary of 50 000 words and a trigram language model trained on web collected data, meeting data and CTS data [38]. As before, we constructed baseline systems using the conventional and STRAIGHT-based systems independently, then produced a combined feature stream by concatenation and dimension reduction using HLDA (using both monophone Gaussian components and cross-word triphone states as classes). The resulting systems corresponded to a sub-system (denoted VTLN enhanced P1) of the AMI-ASR meeting transcription system [40] which participated in the NIST RT evaluation 2006, with the difference that MFCC features were used rather than MF-PLP features.

[4]http://www.nist.gov/speech/tests/rt/rt2004/spring/

TABLE IV
WORD ERROR RATES FOR MEETING TRANSCRIPTION (IHM CONDITION) USING THE RT04SEVAL TESTING SET. RESULTS ARE GIVEN FOR BASELINE SYSTEMS USING CONVENTIONAL AND STRAIGHT-DERIVED MFCCS, AND FOR COMBINED FEATURE VECTORS OBTAINED USING HLDA.

| | TOTAL | Female | Male | CMU | ICSI | LDC | NIST |
|---|---|---|---|---|---|---|---|
| MFCC+VTLN (A) | 38.4 | 38.5 | 38.3 | 42.7 | 23.9 | 52.1 | 30.9 |
| STRAIGHT+VTLN (B) | 39.3 | 38.3 | 39.7 | 44.7 | 24.8 | 53.1 | 31.2 |
| MFCC+STRAIGHT +VTLN | 42.1 | 44.4 | 41.0 | 45.6 | 28.5 | 55.4 | 37.0 |
| MFCC+STRAIGHT VTLN+HLDA xwrd (E) | 37.3 | 37.6 | 37.2 | 41.4 | 23.8 | 51.9 | 29.4 |
| MFCC+STRAIGHT VTLN+HLDA mono (F) | 36.6 | 36.3 | 36.7 | 41.0 | 22.5 | 51.2 | 28.5 |

The results for the IHM condition are shown in table IV. The STRAIGHT derived MFCCs result in slightly higher word error rates than conventional MFCCs; we note that pitch extraction is also challenging in the meeting domain. Lower error rates are observed for female speakers using STRAIGHT, while for male speakers lower error rates are observed for conventional MFCCs. Combination of the two systems using HLDA with monophone Gaussian component classes results in a significant absolute reduction in word error rate of 1.8% (5% relative) compared with the baseline conventional MFCCs.

Word error rates for the MDM condition are shown in table V. In this case there is a 2% absolute difference between the baseline conventional and STRAIGHT systems, which is larger than for the IHM case. Beamformed signals from remote microphones have increased additive and channel noise, compared with the IHM condition, leading to less reliable pitch tracking, and hence less reliable estimates of the pitch-adaptive window in STRAIGHT. However, the combination of the two systems by HLDA using monophone Gaussian classes results in a substantial decrease in word error rate of 3.6% absolute (7.3% relative), which is consistent over the different subsets. It is possible that when conventional and STRAIGHT MFCCs are combined, STRAIGHT mis-estimations are compensated for by conventional MFCCs and conversely conventional MFCCs are enhanced by the smoother and more accurate STRAIGHT spectral representation.

There is also a large difference between word error rates for male and female speakers. Beamforming is known to have less directionality at lower frequencies, while it has some aliasing at higher frequencies. Since, in male voices, information content and the fundamental frequency is concentrated at lower frequencies, it is possible that the higher error rate observed for males results from this limited directionality at low frequencies and therefore less reliable pitch tracking.

TABLE V
WORD ERROR RATES FOR MEETING TRANSCRIPTION (MDM CONDITION) USING THE RT04SEVAL TESTING SET. RESULTS ARE GIVEN FOR BASELINE SYSTEMS USING CONVENTIONAL AND STRAIGHT-DERIVED MFCCS, AND FOR COMBINED FEATURE VECTORS OBTAINED USING HLDA.

| | TOTAL | Female | Male | CMU | ICSI | LDC | NIST |
|---|---|---|---|---|---|---|---|
| MFCC+VTLN | 49.5 | 46.8 | 50.8 | 55.7 | 26.2 | 60.1 | 33.1 |
| STRAIGHT+VTLN | 51.5 | 48.6 | 52.9 | 57.4 | 26.2 | 63.4 | 34.6 |
| MFCC+STRAIGHT VTLN+HLDA xwrd | 46.8 | 42.2 | 49.1 | 52.5 | 24.3 | 58.1 | 29.5 |
| MFCC+STRAIGHT VTLN+HLDA mono | 45.9 | 42.7 | 47.4 | 50.8 | 21.3 | 57.7 | 30.1 |

TABLE VI
EXTENDED DIMENSIONALITY EXPERIMENT ON RT04SEVAL TESTING SET USING VTLN FEATURES FOR THE IHM CONDITION. FROM TOP TO BOTTOM: 39 DIMENSIONS CONVENTIONAL AND STRAIGHT DERIVED MFCCS; 63 DIMENSIONS CONVENTIONAL AND STRAIGHT DERIVED MFCCS.

| | Dimensions | TOTAL | Female | Male | CMU | ICSI | LDC | NIST |
|---|---|---|---|---|---|---|---|---|
| MFCC+VTLN (A) | 39 | 38.4 | 38.5 | 38.3 | 42.7 | 23.9 | 52.1 | 30.9 |
| STRAIGHT+VTLN (B) | 39 | 39.3 | 38.3 | 39.7 | 44.7 | 24.8 | 53.1 | 31.2 |
| MFCC+VTLN (C) | 63 | 37.1 | 38.5 | 36.4 | 41.3 | 22.2 | 51.5 | 31.2 |
| STRAIGHT+VTLN (D) | 63 | 36.7 | 36.4 | 36.8 | 41.0 | 22.3 | 50.8 | 30.0 |

### *E. Further experiments on meetings*

Higher order cepstral coefficients are known to be the most affected by the spectral harmonic components due to the pitch [23], hence systems using conventional MFCCs typically limit their dimensionality to twelve coefficients plus C0 or the log energy. However, using the smoothed STRAIGHT spectral representation, which is not affected by spectral harmonics, we should be able to exploit the information in higher order coefficients. To assess this possibility, we carried out a set of experiments using the first 20 cepstral coefficients (plus C0) and their first and second temporal derivatives, resulting in 63-dimension acoustic feature vectors, in the IHM meeting domain both for the STFT-based MFCCs and our pitch-synchronous MFCCs.

The results of these experiments are shown in table VI, where we repeat the results of the 39-dimension systems to facilitate comparison. It is interesting to observe that the system based on higher order STRAIGHT derived MFCCs has a lower word error rate than both lower and higher order conventional MFCC based systems. In particular the higher order MFCC system does not result in fewer errors for female speakers: this is due to the fact that for high pitched speakers the mel filter bandwidths are not sufficiently broad to remove the harmonic structure which affects the higher order coefficients. On the other hand STRAIGHT derived features, which are not influenced by pitch harmonics, are able to exploit the information of higher order coefficients even for female speakers for which they perform significantly better than STFT based features.

We also performed some experiments on the use of STRAIGHT for MF-PLP extraction. Here a PLP implementation based on that of HTK [34] was used, where the mel frequency scaling is performed on the STRAIGHT spectrogram. Similarly to MFCCs, twelve cepstral coefficients plus C0 were extracted along with their first and second derivatives. Word error rates of systems based on STRAIGHT derived MF-PLPs have been compared with those of conventional MF-PLPs extracted by HTK and these two feature streams have been concatenated and reduced through HLDA from 78 to 39 dimensions using monophone mixture components as classes. Results are shown in table VII. Word error rates were somewhat lower both for the individual feature systems and for the combination through HLDA, compared with the MFCC experiments. The combination by HLDA yields a word error rate reduction of 1.2% absolute (3.2% relative) compared with conventional PLPs.

TABLE VII
MF-PLP EXPERIMENT ON RT04SEVAL TESTING SET USING VTLN FEATURES FOR THE IHM CONDITION. FROM TOP TO BOTTOM: CONVENTIONAL MF-PLPS 39 DIMENSIONS; STRAIGHT MF-PLPS 39 DIMENSIONS; HLDA COMBINATION FROM 78 TO 39 DIMENSIONS USING MONOPHONE MIXTURES AS CLASSES.

| | TOTAL | Female | Male | CMU | ICSI | LDC | NIST |
|---|---|---|---|---|---|---|---|
| MF-PLP+VTLN (G) | 37.4 | 35.8 | 38.3 | 42.5 | 23.3 | 50.8 | 30.4 |
| STRAIGHT MF-PLP +VTLN (H) | 38.4 | 37.4 | 38.9 | 43.7 | 24.4 | 51.9 | 30.3 |
| MF-PLP+STRAIGHT MF-PLP VTLN+HLDA mono (I) | 36.2 | 36.0 | 36.3 | 40.0 | 22.4 | 51.0 | 28.5 |

### *F. ROVER combination experiments on meetings*

To fully exploit the complementarity of conventional and pitch synchronous representations, we performed combination experiments at the system level using majority voting ROVER for the IHM condition of the meeting domain. We considered all the different IHM systems discussed in the previous subsections, with the exception of the simple feature combination with no dimension reduction. Results are reported in table VIII, where we also present WERs for the ROVER oracle to provide a lower bound on the achievable word error rates for each combination. Results for each individual system are reported in tables IV, VI and VII, and each of the nine systems is identified by a letter. *A* and *B* denote the conventional and STRAIGHT derived systems for lower order MFCCs, while *C* and *D* are the same but for higher order MFCCs; *E* and *F* are the HLDA combinations of *A* and *B* with monophone Gaussian classes and triphone state classes respectively; finally *G* and *H* are the MF-PLP systems from conventional and STRAIGHT derived spectral representations, while *I* is their combination using HLDA and monophone Gaussian classes.

First of all comparing the combinations *ACG* (STFT spectral representations) and *BDH* (STRAIGHT representations), we observe that while they have similar accuracies overall, STRAIGHT representations seem to favour female speakers while male speakers are recognised better by the conventional STFT based features. When they are merged together in *ABCDGH* the greatest improvement is still maintained for females.

TABLE VIII
SYSTEM LEVEL COMBINATION IN THE MEETING DOMAIN (IHM CONDITION) ON RT04SEVAL IHM, USING ROVER. THE LEFT HAND TABLES SHOW MAJORITY VOTING ROVER RESULTS AND THE RIGHT SHOWS ROVER ORACLE RESULTS FOR COMPARISON. NINE SYSTEMS ARE COMBINED, LABELLED A–I, AND RESULTS FOR THE INDIVIDUAL SYSTEMS ARE SHOWN IN TABLES IV, VI AND VII.

| | ROVER voting | | | | | | | ROVER oracle | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TOT | F | M | CMU | ICSI | LDC | NIST | TOT | F | M | CMU | ICSI | LDC | NIST |
| A C G | 36.0 | 35.8 | 36.1 | 40.6 | 22.0 | 49.8 | 29.0 | 27.9 | 26.8 | 28.4 | 31.8 | 15.9 | 40.1 | 21.0 |
| B D H | 36.4 | 35.2 | 37.0 | 41.7 | 22.2 | 49.8 | 28.8 | 29.6 | 28.4 | 30.2 | 34.6 | 17.0 | 41.4 | 22.6 |
| A B C D G H | 34.9 | 33.5 | 35.6 | 39.8 | 21.0 | 48.5 | 27.2 | 23.9 | 22.6 | 24.6 | 28.1 | 13.2 | 34.5 | 17.3 |
| A B C D E F | 34.1 | 33.3 | 34.5 | 38.7 | 20.5 | 47.6 | 26.8 | 22.4 | 21.3 | 23.0 | 26.3 | 12.3 | 33.0 | 15.6 |
| A B C D | 34.9 | 33.4 | 35.6 | 39.8 | 21.0 | 48.5 | 27.1 | 26.2 | 25.2 | 26.7 | 30.6 | 14.5 | 37.6 | 19.4 |
| G H I | 35.4 | 34.3 | 35.9 | 40.0 | 21.5 | 49.3 | 27.8 | 27.3 | 25.8 | 28.1 | 31.5 | 15.7 | 38.9 | 20.6 |
| A B E F | 35.1 | 34.4 | 35.5 | 39.8 | 21.3 | 49.2 | 27.2 | 25.9 | 24.5 | 26.6 | 30.0 | 14.6 | 37.7 | 18.5 |
| A B G H | 36.5 | 35.1 | 37.2 | 41.8 | 22.6 | 49.7 | 28.8 | 28.0 | 26.3 | 28.9 | 32.8 | 16.1 | 39.7 | 20.9 |
| A B E F G H I | 34.9 | 33.8 | 35.4 | 39.7 | 21.1 | 48.8 | 26.8 | 23.0 | 21.3 | 23.9 | 27.0 | 12.9 | 33.5 | 16.2 |
| A B C D E F G H I | 33.8 | 32.6 | 34.4 | 38.4 | 20.1 | 47.2 | 26.6 | 20.9 | 19.5 | 21.6 | 24.7 | 11.4 | 30.6 | 14.5 |

ROVERing the HLDA system outputs with those of the original ones used for the combination gives a substantial improvement with respect to the HLDA feature combinations: *ABEF* gives a 1.5% improvement compared to *E* alone, while *ABCDEF* is 0.8% better than *ABCD*; similarly for PLPs, *GHI* improves the HLDA combination system *I* by 0.8% also. This is of interest because it indicates that ROVER acts in a complementary way to HLDA, being able to further improve the already combined systems.

Complementarity between MFCC- and PLP-based systems is more difficult to exploit than that between conventional and STRAIGHT-based systems. When we consider the combination of all the MFCC based systems *ABCD* with the PLP-based systems *GH*, we observe that *ABCDGH* has a similar error rate to *ABCD* for the majority voting experiment, although there was a substantial improvement in the oracle case. On the other hand, the contribution of the higher order representations (*CD*) is significant (around 1% absolute), and occurs consistently when comparing *ABCDEF* with *ABEF*, *ABCDGH* with *ABGH*, and *ABCDEFGHI* with *ABEFGHI*.

Finally the best result is obtained by combining all the available systems *ABCDEFGHI*, consistent with Schlüter et al. [4]. This yields a substantial decrease in word error rate of 2.4% absolute (6.6% relative) compared with the best HLDA system *I* (HLDA combination of PLPs), and 2.9% absolute (7.9% relative) compared with the best single stream system *D* (higher order STRAIGHT derived MFCCs). Overall, by combining HLDA and ROVER we were able to reduce the word error rate by 4.6% absolute (12% relative) compared with the baseline normalised lower order MFCC system. The oracle results indicate that it is possible to further exploit complementarity between representations and thus reduce word error rates more.

### G. Discussion

STRAIGHT derived features offer the most benefit in conjunction with VTLN, as expected. However, they mostly did not result in an overall improvement in accuracy, compared with conventional features, although improvements were observed for female speakers. The elimination of pitch interference effects also proved to be important when higher order coefficients were used.

Combining conventional and STRAIGHT-based features using HLDA reduced the word error rate in all cases. Conventional MFCCs are affected by pitch interference but they are extracted from a sharper representation, while STRAIGHT features are affected by pitch tracking errors, but are smoother and devoid of pitch interference. The two spectral representations are thus complementary and their combination provides consistent improvements. Pitch tracking errors occur in telephone speech because of the band-pass filtering channel effect, in the meeting domain because of the presence of cross-talk, and in case of beamformed signals because of the decreased directionality at lower frequencies. Nevertheless the combination using HLDA is able to yield consistent improvements even in more challenging domains (CTS and MDM meetings), where the relative improvement is, in fact, greater.

In order to analyse our experiments, and to better exploit the complementarity of the pitch synchronous spectral representation, we investigated system combination using ROVER. These experiments confirmed that STRAIGHT is well-matched to female speakers, the importance of the information contained in higher order coefficients (which can be exploited thanks to the pitch synchronicity of STRAIGHT), and the complementarity of HLDA and ROVER techniques.

## VI. CONCLUSIONS

We have investigated a pitch synchronous acoustic parameterisation for speech recognition, derived from the STRAIGHT approach to time-frequency analysis, with a particular focus on speaker normalisation (VTLN) and combination with conventional features using HLDA. We performed experiments on three large vocabulary domains, using standard data sets and evaluation protocols: WSJCAM0, conversational telephone speech and multiparty meeting transcription, considering both close-talking and microphone array conditions in the latter domain.

In each domain we observed significant reductions in word error rate through the combination of conventional and STRAIGHT-based features using HLDA. The resulting systems based on these combined representations were able to achieve relative reductions in word error rate of 3.2% on WSJCAM0, 8% on conversational telephone speech, and for the meeting

domain 4.7% for the IHM condition and 7.3% for the MDM condition. In both the WSJCAM0 and CTS domains, we found that STRAIGHT derived features benefit the most from VTLN (because of their smoother representation). Experiments on the CTS domain showed that the influence of the pitch tracker is of importance for STRAIGHT derived feature extraction.

Experiments on the use of pitch synchronous MF-PLPs for the meeting IHM task showed a 3.2% relative WER improvement when combined with conventional MF-PLPs using HLDA. On the same task the use of higher order coefficients (20 MFCCs plus C0) was evaluated both for standard and pitch-synchronous features, finding that STRAIGHT-based features performed better than conventional features, particularly for female speakers. In fact, for STFT derived features, higher order coefficients are strongly affected by pitch interference which is more evident in high-pitched speakers. Finally ROVER system level combination was applied on top of HLDA feature level combination finding that further improvements can be achieved merging the output of the baseline systems with the correspondent HLDA combined system; therefore showing that ROVER is complementary to HLDA.

In the future we will further investigate the individual contributions from each representation, decoupling the pitch synchronicity from the smoothing effect of STRAIGHT to help us understand whether speech recognition errors are due to pitch misestimations or to an excessive smoothing.

## ACKNOWLEDGEMENTS

We thank Prof. Hideki Kawahara of the Faculty of Systems Engineering, Wakayama University for giving us the opportunity to use the STRAIGHT code, the University of Cambridge Engineering Department Speech Group for the use of h5train03, and the members of the AMI-ASR team, in particular Ing. Martin Karafiat (Brno University of Technology) and Dr. Thomas Hain (University of Sheffield).

## REFERENCES

[1] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Conversational speech recognition using acoustic and articulatory input," in *Proc. IEEE ICASSP*, 2000.

[2] Q. Zhu, B. Chen, N. Morgan, and A. Stolcke, "On using MLP features for LVCSR," in *Proc. Eurospeech*, 2004.

[3] A. Zolnay, D. Kocharov, R. Schlüter, and H. Ney, "Using multiple acoustic feature sets for speech recognition," *Speech Communication*, vol. 49, pp. 514–525, 2007.

[4] R. Schlüter, I. Bezrukov, H. Wagner, and H. Ney, "Gammatone features and feature combination for large vocabulary speech recognition," in *Proc. IEEE ICASSP*, 2007.

[5] D. Hillard, B. Hoffmeister, M. Ostendorf, R. Schlüter, and H. Ney, "iROVER: Improving system combination with classification," in *Proc. NAACL-HLT*, 2007.

[6] J. Cohen, T. Kamm, and A. Andreou, "Vocal tract normalization in speech recognition: compensating for systematic speaker variability," *J. Acoust. Soc. Am.*, vol. 97, no. 5, Pt. 2, pp. 3246–3247, 1995.

[7] L. Lee and R. Rose, "Speaker normalisation using efficient frequency warping procedures," *Proc. IEEE ICASSP*, pp. 353–356, 1996.

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

[9] L. Welling, H. Ney, and S. Kanthak, "Speaker adaptive modeling by vocal tract normalisation," *IEEE Transactions on Speech and Audio Processing*, vol. 10, pp. 415–426, Sept. 2002.

[10] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalisation on conversational telephone speech," in *Proc. IEEE ICASSP*, 1996.

[11] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. IEEE ICASSP*, 1996, pp. 346–348.

[12] 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*, vol. 27, no. 3-4, pp. 187–207, 1999.

[13] T. F. Quatieri, *Discrete Time Speech Signal Processing: Principles and Practice*. Prentice Hall, 2001.

[14] L. Gu and K. Rose, "Perceptual harmonic cepstral coefficients for speech recognition in noisy environments," in *Proc. IEEE ICASSP*, 2001.

[15] A. V. Rao, S. Ahmadi, J. Linden, A. Gersho, V. Cuperman, and R. Heidari, "Pitch adaptive windows for improved exitation coding in low-rate CELP coders," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 648–659, 2003.

[16] H. Ezzaidi and J. Rouat, "Comparison of MFCC and pitch synchronous AM, FM parameters for speaker identification," in *Proc. ICSLP*, 2000.

[17] S. Kim, T. Eriksson, H. Kang, and D. H. Youn, "A pitch synchronous feature extraction method for speaker recognition," in *Proc. IEEE ICASSP*, 2004.

[18] R. Zilca, J. Navratil, and G. N. Ramaswamy, "Depitch and the role of fundamental frequency in speaker recognition," in *Proc. IEEE ICASSP*, 2003.

[19] T. A. Stephenson, J. Escofet, M. Magimai-Doss, and H. Bourlard, "Dynamic Bayesian network based speech recognition with pitch and energy as auxiliary variables," in *Proc. IEEE Workshop in Neural Networks for Signal Processing*, 2002.

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

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

[22] M. Ghulam, T. Fukuda, J. Horikawa, and T. Nitta, "A noise-robust feature extraction method based on pitch-synchronous ZCPA for ASR," in *Proc. ICSLP*, 2004.

[23] T. Irino, Y. Minami, T. Nakatani, M. Tsuzaki, and H. Tagawa, "Evaluation of a speech recognition/generation method based on HMM and STRAIGHT," in *Proc. ICSLP*, 2002.

[24] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier, 1995, pp. 495–518.

[25] L. Burget, "Combination of speech features using smoothed heteroscedastic linear discriminant analysis," in *Proc. ICSLP*, 2004.

[26] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognition Output Voting Error Reduction (ROVER)," in *Proc. IEEE Workshop on ASRU*, 1997.

[27] M. J. Hunt, "A statistical approach to metrics for word and syllable recognition," *J. Acoustical Society of America*, vol. 66, pp. S535–536, 1979.

[28] M. J. Hunt and C. Lefebvre, "Speaker dependent and independent speech recognition experiments with an auditory model," in *Proc. IEEE ICASSP*, vol. 1, 1988, pp. 215–218.

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

[30] L. Burget, "Complementarity of speech recognition systems and system combination," Ph.D. dissertation, Brno University of Technology, 2004.

[31] M. Gales., "Semi-tied covariance matrices for Hidden Markov Models," *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 272–281, 1999.

[32] G. Garau, S. Renals, and T. Hain, "Applying vocal tract length normalisation to meeting recordings," in *Proc. Eurospeech*, 2005.

[33] R. Schlüter, A. Zolnay, and H. Ney, "Feature combination using linear discriminant analysis and its pitfalls," in *Proc. Interspeech*, 2006.

[34] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book (v3.4)*. Cambridge University Engineering Department, December 2006.

[35] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, I. McCowan, D. Moore, V. Wan, R. Ordelman, and S. Renals, "The 2005 AMI system for the transcription of speech in meetings," in *Machine Learning for Multimodal Interaction: Proceedings of MLMI '05*, ser. Lecture Notes in Computer Science. Springer, 2006, no. 3869, pp. 450–462.

[36] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Proc. Eurospeech*, 1995.

[37] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition," in *Proc. IEEE ICASSP*, Detroit, MI, 1995, pp. 81–84.

[38] T. Hain, J. Dines, G. Garau, M. Karafiat, D. Moore, V. Wan, R. Ordelman, I.Mc.Cowan, J.Vepa, and S.Renals, "An investigation into transcription of conference room meetings," in *Proc. Eurospeech*, 2005.

[39] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-24, no. 5, October 1976.

[40] 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.

[41] J. G. Fiscus, J. Ajot, M. MIchel, and J. S. Garofolo, "The Rich Transcription 2006 spring meeting recognition evauation," in *Machine Learning for Multimodal Interaction: Proceedings of MLMI '06*, ser. Lecture Notes in Computer Science. Springer, 2006, no. 4299, pp. 309–322.