# Estimating acoustic speech features in low signal-to-noise ratios using a statistical framework

Philip Harding, Ben Milner*

*School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK*

## Abstract

Accurate estimation of acoustic speech features from noisy speech and from different speakers is an ongoing problem in speech processing. Many methods have been proposed to estimate acoustic features but errors increase as signal-to-noise ratios fall. This work proposes a robust statistical framework to estimate an acoustic speech vector (comprising voicing, fundamental frequency and spectral envelope) from an intermediate feature that is extracted from a noisy time-domain speech signal. The initial approach is accurate in clean conditions but deteriorates in noise and with changing speaker. Adaptation methods are then developed to adjust the acoustic models to the noise conditions and speaker. Evaluations are carried out in stationary and nonstationary noises and at SNRs from -5dB to clean conditions. Comparison with conventional methods of estimating fundamental frequency, voicing and spectral envelope reveals the proposed framework to have lowest errors in all conditions tested.

*Keywords:* Voicing, Fundamental frequency, Spectral envelope, Noise

*Corresponding author
*Email address:* b.milner@uea.ac.uk (Ben Milner)

adaptation, Speaker adaptation

## 1. INTRODUCTION

Acoustic speech features take many forms and include parameters such as voicing, fundamental frequency, spectral envelope, formant frequencies and voice activity. Excitation features, such as voicing and fundamental frequency, are used in many speech processing applications and include, for example, speech coding, enhancement, noise estimation, automatic speech recognition in noisy conditions and tonal language speech recognition (Kaewtip et al., 2013; Kawahara et al., 2001; Lei et al., 2006; Ma et al., 2007; McAulay and Champion, 1990; Morales-Cordovilla et al., 2011a,b). Similarly, spectral envelope and formant features are used in a range of applications such as speech coding, synthesis, recognition and voice conversion (Hermansky, 1990; Kawahara et al., 2001, 2009; Koriyama et al., 2014). Consequently, many methods have been developed to estimate acoustic speech features and these operate in both clean and noisy conditions. In this work a single statistical framework is proposed for estimating a set of acoustic speech features and is designed to be robust at low signal-to-noise ratios (SNRs). Specifically, from a wide range of acoustic features, this work concentrates on estimating voicing, fundamental frequency and spectral envelope, although could equally be applied to other acoustic features.

Many approaches have been developed to estimate voicing and fundamental frequency ($f_0$) and attain good accuracy in noise free conditions but deteriorate as signal-to-noise ratios (SNRs) reduce.

In fact such methods have been applied to a range of periodic signals, such as music, sonar and heart rate monitoring (Godsill and Davy, 2002; Schäck et al., 2015), although in this work the focus is on speech signals. Methods to estimate fundamental frequency can be broadly considered as being parametric or non-parametric. Common non-parametric methods include RAPT and YIN (de Cheveigné and Kawahara, 2002; Talkin, 1995). RAPT uses peaks in the autocorrelation function (ACF) as candidate fundamental frequencies and then uses dynamic programming to find voiced frames and an $f_0$ contour. YIN takes peaks of the squared difference function as fundamental frequency estimates, which is shown to be more robust than the ACF. Further processing reduces over- and underestimation although no voicing classification is made. However, these methods are generally inaccurate in low noise conditions and when estimating low fundamental frequencies (Nielsen et al., 2016). Noise-robust non-parametric methods include XAFE and PEFAC (ETSI, 2003; Gonzalez and Brookes, 2014). XAFE first employs explicit noise reduction and then searches the power spectrum for spectral peaks, corresponding to harmonics, which form $f_0$ candidates. PEFAC works in the log-frequency spectral domain and achieves robustness by first normalising the speech periodogram to reduce noise and channel effects before using a matched filter to extract a series of $f_0$ candidates. A voiced speech probability is also computed and dynamic programming applied to identify voiced regions and an $f_0$ contour. Parametric methods employ a model of the noisy speech signal with one of its parameters being fundamental frequency (although other parameters such as the amplitudes and phases of the harmonics can also be included in the parameter set). An estimate of the model parameters

is then made from the noisy signal using, for example, maximum likelihood (ML), non-linear least squares (NLS) and weighted least squares (WLS) methods (Christensen and Jakobsson, 2009; Li et al., 2000; Nielsen et al., 2016). Parametric methods are inherently robust to noise and can come close to optimal performance according to the Cramér-Rao lower bound (Christensen and Jakobsson, 2009). A further benefit of parametric methods is that they have been shown to estimate accurately low fundamental frequencies which non-parametric methods tend to be less effective at doing (Christensen, 2013a). Furthermore, including prior information on how the fundamental frequency evolves from frame to frame enabled a maximum a posteriori (MAP) estimator capable of tracking fundamental frequencies through a dynamic programming implementation (Tabrikian et al., 2004). While some methods of fundamental frequency estimation implicitly provide voicing classification, other methods have been developed explicitly for voicing classification (Dhananjaya and Yegnanarayana, 2013; Harding and Milner, 2012).

Methods to estimate spectral envelope seek a smooth contour that joins important spectral peaks and is estimated typically by linear predictive coding (LPC) (Makhoul, 1975), cepstrum processing (Oppenheim and Schafer, 1975) or filterbank analysis (Cappe and Moulines, 1996). Again, the accuracy of these methods is good in clean conditions but deteriorates in noise. Applying noise removal methods, for example (Ephraim and Malah, 1985; Scalart and Vieira-Filho, 1996), before estimation improves the resulting spectral envelopes.

This work proposes a single statistical framework for estimating an acoustic speech vector from an intermediate feature vector that is extracted

from a time-domain speech signal. The acoustic speech vector in this work contains fundamental frequency, voicing and spectral envelope and is estimated in its entirety within the statistical framework rather than requiring individual algorithms for each acoustic feature. The intermediate feature vector can take different forms and could be, for example, an MFCC vector, filterbank vector or other suitable representation (Milner, 2002). Earlier work into statistical estimation of acoustic features developed effective methods of fundamental frequency, voicing and formant estimation in noise-free conditions (Darch, 2008). More recently, noise compensation was included using parallel model combination (PMC) which increased accuracy (Gales and Young, 1996; Milner and Darch, 2011). This work advances earlier work in three areas. First, the previous work operated within the ETSI DSR standard which constrained the intermediate feature to be a 13-D MFCC vector extracted from a 23-D mel filterbank (ETSI, 2003). This constraint is now removed which allows better intermediate features to be found. Second, previous work found that when moving from speaker dependent modelling to speaker independent modelling the increased speaker variability reduced estimation accuracy. Speaker adaptation is now integrated into the estimation procedure to improve acoustic modelling. Third, noise compensation uses a nonlinear noise mismatch function that considers both phase mismatch and spectral mismatch and is applied using the unscented transform (Hu and Huo, 2006).

Similarity can be drawn between the proposed method and the extended invariance principle (EXIP) where a set of intermediate parameters are also first extracted from the signal and then fitted to a more structured model (Stoica and Söderström, 1989). EXIP has been applied to a range of

tasks that include radar target estimation and speech processing applications such as vector quantisation for speech enhancement and separation as well as fundamental frequency estimation (Christensen, 2013b; Li et al., 2000; Swindlehurst and Stoica, 1998). As an example, the Markov-like weighted least squares (WLS) method of fundamental frequency estimation (Li et al., 2000) extracts first from the speech signal a set of unstructured initial parameter estimates that include harmonic frequencies and their amplitudes and phases. These are then fitted to a structured model where a weighting matrix accounts for uncertainty (or accuracy of estimation) and dependencies within the parameter vector. The final estimate of fundamental frequency is a closed-form solution that is a weighted summation of the unstructured harmonic frequencies.

The remainder of this paper is organised as follows. Section II reviews statistical estimation of acoustic speech features and extends previous work by considering forms that the intermediate feature may take. Sections III and IV examine the effect of speaker and noise variability and integrate adaptation methods into the estimation. Experimental results are presented in Section V that analyse fundamental frequency, voicing and spectral envelope estimation using speaker dependent and speaker independent systems.

## 2. FEATURES AND ESTIMATION

This section begins with an overview of statistical estimation of acoustic speech features and then considers choices for the intermediate feature. The effects of noise and speaker variability are then examined which leads to the proposal of a noise and speaker adaptive method of estimating acoustic speech

features.

*2.1. Statistical estimation of acoustic speech features*

Several studies have shown correlation to exist between acoustic speech features (fundamental frequency, voicing, formant frequencies and spectral envelope) and intermediate features such as MFCCs, filterbank, and LPC coefficients (Darch, 2008; Harding, 2013; Hirahara, 1988; Syrdal and Steele, 1985). Previous work used this correlation to enable a statistical method of estimating acoustic speech features from intermediate vectors (Milner and Darch, 2011). A joint feature vector, $\boldsymbol{z}_i$, is defined

$$\boldsymbol{z}_i = [\boldsymbol{x}_i, \boldsymbol{\theta}_i] \tag{1}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{\theta}_i$ are the intermediate vector and acoustic speech feature extracted from the $i$th frame of speech - note, for clarity, the frame index is subsequently omitted. The elements of the acoustic feature vector, $\boldsymbol{\theta}$, depend on the acoustic features being modelled, which in this work are voicing, fundamental frequency and spectral envelope. This gives an acoustic feature

$$\boldsymbol{\theta} = [f_0, \boldsymbol{\chi}] \tag{2}$$

where $f_0$ and $\boldsymbol{\chi}$ are the fundamental frequency and spectral envelope of the frame of speech. To signify unvoiced speech and non-speech $f_0$ is set to zero. Spectral envelope is represented using an $M$-channel filterbank where each element $\chi(m)$ is the log amplitude of the $m$th channel. The framework allows other acoustic features to be included in $\boldsymbol{\theta}$, for example formant frequencies, if required.

From a set of training data, along with reference annotations in terms of voicing class, three vector pools, $\Upsilon^{v}$, $\Upsilon^{uv}$ and $\Upsilon^{ns}$, are created that contain voiced speech, unvoiced speech and non-speech vectors respectively. Applying expectation-maximisation (EM) training to each vector pool creates Gaussian mixture models (GMMs), $\Phi^{z}_{v}$, $\Phi^{z}_{uv}$ and $\Phi^{z}_{ns}$, that model the joint density of intermediate feature and acoustic speech feature for voiced, unvoiced and non-speech respectively. To simplify notation the GMMs are expressed as $\Phi^{z}_{vc}$ where $vc \in \{v, uv, ns\}$. Each GMM comprises a mixture of Gaussian probability density functions, $\phi^{z}_{k,vc}$

$$p(\boldsymbol{z}|\Phi^{z}_{vc}) = \sum_{k=1}^{K} \alpha_{k,vc}\, \phi^{z}_{k,vc}(\boldsymbol{z}) = \sum_{k=1}^{K} \alpha_{k,vc}\, \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}^{z}_{k,vc}, \boldsymbol{\Sigma}^{zz}_{k,vc}) \tag{3}$$

The $k$th mixture component has mean vector $\boldsymbol{\mu}^{z}_{k,vc}$ and covariance matrix $\boldsymbol{\Sigma}^{zz}_{k,vc}$, which are defined

$$\boldsymbol{\mu}^{z}_{k,vc} = \begin{bmatrix} \boldsymbol{\mu}^{x}_{k,vc} \\ \boldsymbol{\mu}^{\theta}_{k,vc} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}^{zz}_{k,vc} = \begin{bmatrix} \boldsymbol{\Sigma}^{xx}_{k,vc} & \boldsymbol{\Sigma}^{x\theta}_{k,vc} \\ \boldsymbol{\Sigma}^{\theta x}_{k,vc} & \boldsymbol{\Sigma}^{\theta\theta}_{k,vc} \end{bmatrix} \tag{4}$$

where the mean vector contains means of the intermediate feature, $\boldsymbol{\mu}^{x}_{k,vc}$, and acoustic speech feature, $\boldsymbol{\mu}^{\theta}_{k,vc}$. The covariance matrix comprises covariances of the intermediate feature, $\boldsymbol{\Sigma}^{xx}_{k,vc}$, the acoustic speech feature, $\boldsymbol{\Sigma}^{\theta\theta}_{k,vc}$, and their cross-covariances, $\boldsymbol{\Sigma}^{x\theta}_{k,vc}$ and $\boldsymbol{\Sigma}^{\theta x}_{k,vc}$. Prior probabilities, $\alpha_{k,vc}$, reflect the proportion of training data in each cluster.

Acoustic feature estimation begins by extracting intermediate feature vectors, $\boldsymbol{x}$, from the speech. The first acoustic feature to be estimated is the voicing as this determines which other acoustic features need to be estimated from that frame of speech - i.e. fundamental frequency if voiced and spectral

envelope if speech. The voicing, $\hat{v}$, is estimated by identifying the GMM with the highest probability for $\boldsymbol{x}$,

$$\hat{v} = \underset{vc \in \{v, uv, ns\}}{\arg\max}\ p(\boldsymbol{x}|\Phi^x_{vc}) \tag{5}$$

where $\Phi^x_{vc}$ is the GMM marginalised to the intermediate vector.

*2.1.1. Fundamental frequency estimation*

For frames classified as voiced (i.e. $\hat{v} = v$) an estimate of fundamental frequency is made from the voiced GMM marginalised to intermediate vector and fundamental frequency components to give $\Phi^{xf_0}_v$. The fundamental frequency estimate, $\hat{f}_0$, given the intermediate vector, $\boldsymbol{x}$, is computed from the conditional mean of each component of the GMM, combined by weighting by the posterior probability,

$$\hat{f}_0 = \sum_{k=1}^{K} h^x_{k,v}(\boldsymbol{x}) \left[\mu^{f_0}_{k,v} + \Sigma^{f_0,x}_{k,v} \left(\Sigma^{xx}_{k,v}\right)^{-1} (\boldsymbol{x} - \mu^x_{k,v})^T\right] \tag{6}$$

where the posterior probability, $h^x_{k,v}(\boldsymbol{x})$, of the intermediate vector, $\boldsymbol{x}$, is computed from marginalised distributions $\phi^x_{k,v}$ taken from each cluster $k$ of the voiced GMM

$$h^x_{k,v}(\boldsymbol{x}) = \frac{\alpha_{k,v}\ p\left(\boldsymbol{x}|\phi^x_{k,v}\right)}{\sum_{j=1}^{J} \alpha_{j,v}\ p\left(\boldsymbol{x}|\phi^x_{j,v}\right)} \tag{7}$$

*2.1.2. Spectral envelope estimation*

For frames classified as speech (i.e. $\hat{v} \in \{v, uv\}$) the estimate of spectral envelope, $\hat{\boldsymbol{\chi}}$, is made from the voiced or unvoiced GMM marginalised to intermediate vector and spectral envelope components, $\Phi^{x\boldsymbol{\chi}}_{\hat{v}}$. Given the intermediate vector, $\boldsymbol{x}$, a weighted conditional mean estimate of spectral

envelope is calculated,

$$\hat{\boldsymbol{\chi}} = \sum_{k=1}^{K} h^{x}_{k,\hat{v}}(\boldsymbol{x}) \left[ \mu^{\boldsymbol{\chi}}_{k,\hat{v}} + \Sigma^{\boldsymbol{\chi},x}_{k,\hat{v}} \left( \Sigma^{\boldsymbol{\chi}\boldsymbol{\chi}}_{k,\hat{v}} \right)^{-1} (\boldsymbol{x} - \mu^{x}_{k,\hat{v}})^{T} \right] \tag{8}$$

where posterior probability, $h^{x}_{k,\hat{v}}(\boldsymbol{x})$, is as defined in Eq. (7).

### *2.2. Intermediate feature extraction*

Many candidates exist for the intermediate feature and include spectral features, filterbank, cepstrum, MFCCs, LPC, line spectral pairs and perceptual linear prediction (Davis and Mermelstein, 1980; Hermansky, 1990; Milner, 2002; Soong, 1984). Studies analysing their correlation to acoustic speech features have found varying levels of correlation to exist (Darch, 2008; Harding, 2013; Hirahara, 1988; Syrdal and Steele, 1985). Of the features investigated, MFCCs were found to have highest correlation to acoustic speech features, particularly when extracted with a large number of channels in the filterbank. This gives fine spectral detail at low frequencies which was found to be important for fundamental frequency estimation (Harding, 2013). Our previous work on acoustic feature estimation (Milner and Darch, 2011) was constrained to the Aurora Distributed Speech Recognition (DSR) standard and limited to 13-D MFCC vectors computed from a 23 channel filterbank (ETSI, 2003). That restriction is now removed which allows larger filterbanks and more MFCCs to be considered as the intermediate feature.

MFCC extraction follows broadly the method proposed in the Aurora DSR standard (ETSI, 2003) and begins by computing the power spectrum of 20ms Hamming windowed frames of audio. These are input into an $M$ channel mel filterbank and a log and discrete cosine transform (DCT) applied. The resulting vectors are not truncated as this was found to give higher accuracy

which results in $M$-dimensional MFCC vectors. Investigations into the size of $M$ are presented in Section V.

### *2.3. Effects of noise and speaker variation*

Estimation of acoustic features is accurate in noise-free conditions and when the speaker matches the speaker used in training. However, in practice the input speech will likely be noisy and speakers not used in training will be encountered. The input speech will now be mismatched to the acoustic models and estimation accuracy will reduce.

This problem is encountered in automatic speech recognition where noise and speaker mismatches increase word error rates (WERs) (Vaseghi and Milner, 1997; Chung and Hansen, 2013). Matched training and testing gives substantial reductions in WER but is not practical in changing conditions. A more effective solution employed in speech recognition is to adapt the acoustic models to the current noise and speaker where significant reductions in WER have been reported (Gales and Young, 1996; Hu and Huo, 2006; Moreno et al., 1996; Gauvain and Lee, 1994; Woodland, 2001). Given the similarity between speech recognition and the proposed acoustic feature estimation, the statistics of the clean trained GMMs, $\Phi_{vc}^{x\theta}$ of Eq. (3), will be adapted to the current speaker and noise to create a matched GMM, $\Phi_{vc}^{\hat{y}\hat{\theta}}$. The next two sections describe the speaker and noise adaptation methods.

## 3. ADAPTATION TO SPEAKER

To adapt the acoustic models to better represent a new speaker, MAP adaptation is used due to its effectiveness in speech recognition applications (Gauvain and Lee, 1994; Woodland, 2001). In speech recognition, adaptation is

applied to the many hundreds or thousands of HMM/GMMs used in acoustic modelling, while in this work adaptation is applied to only the voiced and unvoiced GMMs. Furthermore, in speech recognition the feature is typically an MFCC vector, while in this work both the intermediate and acoustic features require adaptation.

### *3.1. MAP speaker adaptation*

Adaptation is applied to the means, covariances and prior probabilities of the voiced and unvoiced GMMs. From a new speaker a sequence of adaptation vectors, $\boldsymbol{A} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N\}$, is extracted that take the same form as the joint feature vector in Eq. (1). For each vector, $\boldsymbol{a}_i$, the probability, $\gamma_{k,v}(i)$, of each mixture component, $\phi_v^z$, in the GMM, $\Phi_v^z$, is computed (adaptation is shown for the voiced GMM but is identical for the unvoiced GMM)

$$\gamma_{k,v}(i) = \frac{\alpha_{k,v} p(\boldsymbol{a}_i | \phi_{k,v}^z)}{\sum_{j=1}^{K} \alpha_{j,v} p(\boldsymbol{a}_i | \phi_{j,v}^z)} \tag{9}$$

The estimate of the adapted mean, $\boldsymbol{\mu}_{k,v}^{\hat{z}}$, for the $k$th cluster in the GMM is calculated as a weighted combination of the prior mean from the speaker-independent model, $\boldsymbol{\mu}_{k,v}^z$, and the estimated mean from the adaptation data in the $k$th cluster (Gauvain and Lee, 1994)

$$\boldsymbol{\mu}_{k,v}^{\hat{z}} = \frac{\tau \boldsymbol{\mu}_{k,v}^z + \sum_{i=1}^{N} \gamma_{k,v}(i) \boldsymbol{a}_i}{\tau + \sum_{i=1}^{N} \gamma_{k,v}(i)} \tag{10}$$

where $\tau$ determines the bias between the prior mean and mean of the adaptation data. Similarly the covariance, $\boldsymbol{\Sigma}_{k,v}^{\widehat{zz}}$, and mixture weights, $\hat{\alpha}_{k,v}$, are updated by the adaptation vectors

$$\boldsymbol{\Sigma}_{k,v}^{\widehat{zz}} = \frac{\boldsymbol{\Sigma}_{k,v}^{zz} + \sum_{i=1}^{N} \gamma_k(i) (\boldsymbol{a}_i - \boldsymbol{\mu}_k^{\hat{z}})(\boldsymbol{a}_i - \boldsymbol{\mu}_k^{\hat{z}})^T + \boldsymbol{\Psi}}{\omega_k - (M + M_\theta) + \sum_{i=1}^{N} \gamma_k(i)} \tag{11}$$

$$\text{where} \quad \boldsymbol{\Psi} = \tau (\boldsymbol{\mu}_k^z - \boldsymbol{\mu}_k^{\hat{z}})(\boldsymbol{\mu}_k^z - \boldsymbol{\mu}_k^{\hat{z}})^T$$

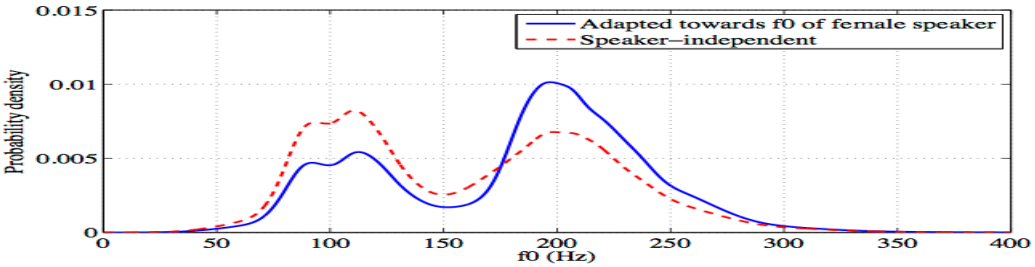
$$\hat{\alpha}_{k,v} = \frac{\alpha_{k,v} - 1 + \sum_{i=1}^{N} \gamma_{k,v}(i)}{\sum_{j=1}^{K} \left( \alpha_{j,v} - 1 + \sum_{i=1}^{N} \gamma_{j,v}(i) \right)} \tag{12}$$

$\omega_{k,v}$ relates to the summed probability of adaptation vectors for the $k$th mixture and is defined in (Gauvain and Lee, 1994). The means and covariances can be calculated using equal or different values of $\tau$, with $2 \leq \tau \leq 20$ commonly used (Woodland, 2001). Experiments in Section 4 use $\tau = 12$ as this was found to give best performance.

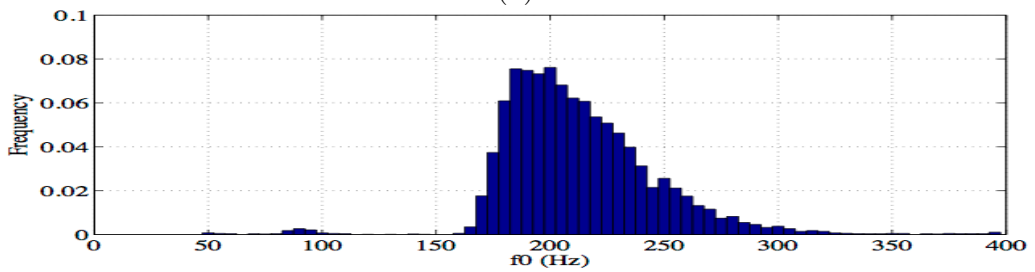
### *3.2. Implementation*

Adaptation should adjust the intermediate feature and acoustic feature components to model the characteristics of the new speaker. Extracting intermediate features from the new speaker is straightforward, however extracting acoustic features can be more erroneous. This leads to two approaches that have been considered: i) adapt both the intermediate and acoustic feature components, and ii) adapt only the intermediate feature component.

To adapt only the intermediate feature, Eq. (9) is marginalised so only the intermediate feature determines the contribution of each mixture component in the GMM, and subsequently Eqs. (10)–(12) adapt only the statistics of the intermediate feature. Tests comparing both approaches found full adaptation gave lower estimation errors and so is the method chosen for subsequent testing. The adaptation process for fundamental frequency estimation is illustrated in Fig. 1(a) which shows (with the dashed line) the original, speaker independent, distribution of $f_0$ obtained by marginalising $\Phi_v^z$. This is learnt from the speaker independent dataset introduced in Section 5 and is bimodal corresponding to male and female speakers. The histogram of a set of adaptation data taken from a single female speaker is shown in Fig.




(a)




(b)

Figure 1: Fundamental frequency adaptation showing: a) speaker independent (dotted) and adapted (solid) distributions, b) histogram of adaptation data.

1(b) and is used to adapt the speaker independent distribution. The adapted distribution is shown in Fig. 1(a) as the solid line and is seen to model more closely the distribution of the female speaker. Varying $\tau$ allows the adaptation data to have more or less influence on the adapted distribution. Interestingly, although fundamental frequency errors are present in the adaptation data (around 90Hz) they have very little effect in the adapted distribution due to the averaging.

Some preliminary tests were carried out to examine the effect that different amounts of adaptation data have on performance. As an example, Fig. 2 shows the accuracy of spectral envelope estimation, as measured using the log likelihood ratio (LLR) (see Section 5.3 for further details), as the

amount of adaptation data is varied. To serve as bounds on performance, the LLR attained with speaker independent models is shown as the dashed line and represents the unadapted case and the starting point for adaptation. These models were trained from a set of 111 different speakers - see Section 5 for details. The LLR attained when using speaker dependent models that were trained solely on the speaker under test is shown as the dotted line and represents best performance and the target for adaptation. In this situation the models are trained on 13 minutes of data from the single speaker. The LLR of the adapted models (solid line) improves rapidly with small amounts of adaptation data and then levels off as more adaptation data becomes available. Specifically, using 5 seconds of adaptation data the performance gain reaches 32% of that obtained with speaker dependent models and with 20 seconds of data reaches 69%. Beyond 20 seconds of adaptation data the rate of gain reduces with gradual convergence towards the speaker-dependent model. A similar trade-off between the amount of adaptation data and performance gain was also observed for the voicing and fundamental frequency acoustic features. Consequently, in this work, 20 seconds of adaptation data is used for each new speaker which is applied statically. If desired, the estimation framework would also allow adaptation to be applied dynamically and the models updated continuously as more data is processed during the utterance (Geiger et al., 2010).

## 4. ADAPTATION TO NOISE

Noise adaptation is a very effective method for improving the noise robustness of HMM-based speech recognisers and adjusts the clean speech

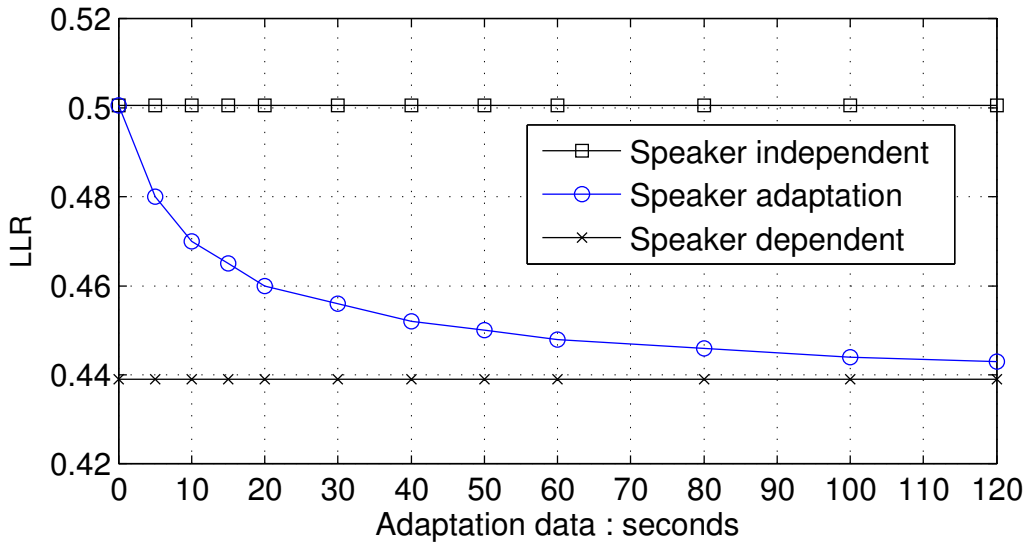



Figure 2: Effect of increasing amount of speaker adaptation data for spectral envelope estimation.

statistics within each state of the HMM to model noisy speech. Methods such as parallel model combination (PMC), vector Taylor series (VTS) and the unscented transform are all effective (Gales and Young, 1996; Hu and Huo, 2006; Moreno et al., 1996). The adaptation required for acoustic feature estimation is slightly different as the joint feature vector, $\boldsymbol{z}$, comprises an intermediate vector component, $\boldsymbol{x}$, which must be adapted to model noisy intermediate features, $\boldsymbol{y}$, and an acoustic feature component, $\boldsymbol{\theta}$, that needs no adaptation – i.e. transforming the GMM $\Phi^{x\theta}$ into $\Phi^{\hat{y}\theta}$. We use the unscented transform for adaptation, based on the results of preliminary investigations, although in practice any of the three should be effective (Harding, 2013). Adaptation is a two stage process that requires first a mismatch function to model the effect of noise on the intermediate feature and secondly application of the unscented transform to adapt the GMM to noise.

*4.1. Derivation of mismatch function*

To adapt a GMM to noise it is necessary to examine how noise affects the intermediate feature, which in this work is an MFCC vector. In the time-domain, speech and noise are additive and remain so in the frequency domain

$$Y(f) = X(f) + D(f) \quad 0 \leq f \leq F-1 \tag{13}$$

where $Y(f)$, $X(f)$ and $D(f)$ are the complex spectra of the noisy speech, clean speech and noise respectively and $f$ is the spectral bin. Transforming to power spectrum gives

$$|Y(f)|^2 = |X(f)|^2 + |D(f)|^2 + 2|X(f)||D(f)|cos(\varphi(f)) \tag{14}$$

where $\varphi(f)$ is the phase difference between the noise and clean speech in the $f$th spectral bin. In many noise compensation methods this phase-related mismatch is assumed zero and ignored. However, recent studies have shown that retaining the phase component improves the modelling of noisy speech (Faubel et al., 2008). Transforming the power spectrum into an M-channel mel filterbank by multiplying by an $M \times F$ matrix, $\boldsymbol{W}$, (where each row is a filterbank basis function) gives filterbank features, $y^{fb}(m)$, $x^{fb}(m)$ and $d^{fb}(m)$, where $m$ denotes the channel

$$y^{fb}(m) = x^{fb}(m) + d^{fb}(m) + 2\beta(m)\sqrt{x^{fb}(m)d^{fb}(m)} \tag{15}$$

$\beta(m)$ relates to the phase difference between the clean speech and noise in the $m$th channel and is defined

$$\beta(m) = \frac{\sum_{f=0}^{F-1} W(m,f)cos(\varphi(f))|X(f)||D(f)|}{\sqrt{x^{fb}(m)d^{fb}(m)}} \tag{16}$$

Taking the log of the phase sensitive mismatch function and defining $\boldsymbol{y}^l$, $\boldsymbol{x}^l$ and $\boldsymbol{d}^l$ as log filterbank vectors, the noisy log mel features can be obtained from a function, $g(.)$, of the clean speech and noise log mel features

$$\begin{aligned} \boldsymbol{y}^l &= g(\boldsymbol{x}^l, \boldsymbol{d}^l, \boldsymbol{\beta}) \\ &= \boldsymbol{x}^l + log(1 + \exp^{\boldsymbol{d}^l - \boldsymbol{x}^l} + 2\boldsymbol{\beta}\sqrt{\exp^{\boldsymbol{d}^l - \boldsymbol{x}^l}}) \end{aligned} \tag{17}$$

Finally, in the MFCC domain, following multiplication by an $M \times M$ DCT matrix, $\boldsymbol{C}$ (where each row is a cosine basis function), the same mismatch function, $g(.)$, gives noisy MFCC vector, $\boldsymbol{y}$, from clean speech and noise MFCC vectors, $\boldsymbol{x}$ and $\boldsymbol{d}$, and phase term, $\boldsymbol{\beta}$,

$$\boldsymbol{y} = \boldsymbol{C}\boldsymbol{y}^l = \boldsymbol{C}g(\boldsymbol{C}^{-1}\boldsymbol{x}, \boldsymbol{C}^{-1}\boldsymbol{d}, \boldsymbol{\beta}) \tag{18}$$

$\beta$ is obtained using the method in (Faubel et al., 2008) where a look-up table is trained offline. Then, for a given $\boldsymbol{x}$ and $\boldsymbol{d}$ the look-up table provides a phase averaged estimate of $\beta$ that is used in Eq. (18).

### *4.2. Updating model parameters*

The voiced, unvoiced and non-speech GMMs that model the joint density of clean intermediate feature and acoustic speech feature are now adapted to model the joint density of noisy intermediate feature and acoustic speech feature. The unscented transform is an effective method for estimating the statistics of a distribution that has undergone a non-linear transformation as is the case of speech and noise addition in the intermediate feature domain (Hu and Huo, 2006). As noise affects only the intermediate feature and not the acoustic feature, the unscented transform adapts only the intermediate feature statistics.

For each GMM and each mixture component, $k$, a set of $2(M + M_\theta)$ sigma points, $\boldsymbol{S}^z_{k,vc} = \{\boldsymbol{s}^z_{1,k,vc}, \boldsymbol{s}^z_{2,k,vc}, \ldots, \boldsymbol{s}^z_{2(M+M_\theta),k,vc}\}$, are sampled from the distribution, $\phi^z_{k,vc}$, and comprise intermediate and acoustic feature components

$$\boldsymbol{s}^z_{i,k,vc} = \left[\boldsymbol{s}^x_{i,k,vc}\ \boldsymbol{s}^\theta_{i,k,vc}\right] \tag{19}$$

where $M_\theta$ is the dimensionality of the acoustic feature vector. The sigma points are sampled so that their mean and covariance equal the mean and covariance of the $k$th mixture component in the GMM, i.e. $\boldsymbol{\mu}^z_{k,vc}$ and $\boldsymbol{\Sigma}^{zz}_{k,vc}$, as described in (Hu and Huo, 2006).

A single Gaussian is also trained on intermediate features extracted from noise-only data and has mean and covariance $\boldsymbol{\mu}^d$ and $\boldsymbol{\Sigma}^{dd}$. An augmented mean and covariance, $\tilde{\boldsymbol{\mu}}^d$ and $\tilde{\boldsymbol{\Sigma}}^{dd}$, are created using zero padding to give the same dimensionality as the joint vector, i.e. $M + M_\theta$

$$\tilde{\boldsymbol{\mu}}^d = \begin{bmatrix} \boldsymbol{\mu}^d_{(M)} \\ \mathbf{0}_{(M_\theta)} \end{bmatrix} \quad \text{and} \quad \tilde{\boldsymbol{\Sigma}}^{dd} = \begin{bmatrix} \boldsymbol{\Sigma}^{dd}_{(M\times M)} & \mathbf{0}_{(M\times M_\theta)} \\ \mathbf{0}_{(M_\theta\times M)} & \mathbf{0}_{(M_\theta\times M_\theta)} \end{bmatrix} \tag{20}$$

A further set of $2(M + M_\theta)$ sigma points, $\boldsymbol{S}^{\tilde{d}} = \{\boldsymbol{s}^{\tilde{d}}_1, \boldsymbol{s}^{\tilde{d}}_2, \ldots, \boldsymbol{s}^{\tilde{d}}_{2(M+M_\theta)}\}$, representing noise are generated from the noise distribution and sampled to have the same mean and covariance as in Eq. (20). These take the form

$$\boldsymbol{s}^{\tilde{d}}_i = \left[\boldsymbol{s}^d_i\ \mathbf{0}_{(M_\theta)}\right] \tag{21}$$

Using the mismatch function, $g(.)$, in Eq. (18) the clean intermediate feature sigma points, $\boldsymbol{s}^x_{i,k,vc}$, and noise sigma points, $\boldsymbol{s}^d_i$, are combined to give noisy sigma points, $\boldsymbol{s}^{\hat{y}}_{i,k,vc}$. Augmenting these with the acoustic feature sigma points, $s^\theta_{i,k,vc}$, gives noise adapted sigma points, $s^{\hat{z}}_{i,k,vc}$

$$\boldsymbol{s}^{\hat{z}}_{i,k,vc} = \left[\boldsymbol{s}^{\hat{y}}_{i,k,vc}\ \boldsymbol{s}^\theta_{i,k,vc}\right] = \left[g(\boldsymbol{s}^x_{i,k,vc}, \boldsymbol{s}^d_i,\ \boldsymbol{\beta})\ \boldsymbol{s}^\theta_{i,k,vc}\right] \tag{22}$$

Finally, for each set of adapted sigma points, $\boldsymbol{S}^{\hat{z}}_{k,vc}$, their mean and covariance are computed and provide the updated statistics for each component, $\phi^{\hat{y}\theta}_{k,vc}$, of the GMM to model the joint density of noisy intermediate feature and acoustic feature.

### *4.3. Implementation*

To investigate the effect that the amount of noise adaptation data has on estimation accuracy, preliminary tests were performed that examined acoustic feature estimation accuracy as the amount of noise adaptation data increased from 0 to 30 seconds. For illustration, Fig. 3 shows fundamental frequency estimation error, $E_{f_0}$ (defined in Eq. (23)), in white, babble and destroyer noises at an SNR of 0dB when using from 0 to 30 seconds of noise. The destroyer noise is taken from the operations room and comprises a fairly constant mechanical noise, people talking and a tannoy sounding at irregular intervals. Adaptation to the stationary white noise is rapid, with just 1 second sufficient to estimate the noise statistics. In babble and destroyer noises, their less stationary nature results in adaptation requiring more data to capture the characteristics, although error is minimised with 5 seconds of data and is still effective with only 1 second. Tests on estimating spectral envelope and voicing features were also made and were found to have similar rates of convergence.

Based on this analysis, the remainder of tests reported in this work use 0.5 seconds of noise adaptation data which is taken from the beginning of each utterance and is before the start of the speech. In a practical scenario the noise statistics would need to be computed using, for example, voice activity detection, minimum statistics or recursive averaging techniques (Martin, 2001;

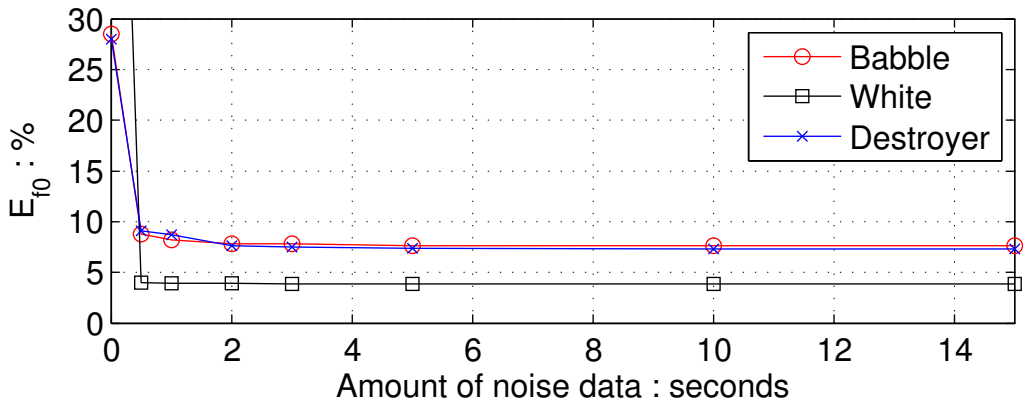



Figure 3: Speaker dependent $E_{f_0}$ for varying amounts of noise adaptation data at an SNR of 0dB in white noise, babble noise and destroyer noise. Note - for clarity the ordinate is limited to 30%.

Rangachari and Loizou, 2006; Taghia et al., 2011). We chose to use a simple method as the noise estimate is reliable which means subsequent analysis of acoustic feature estimation is not subject to variations in noise estimation accuracy. For completeness we did perform some tests using the more practical noise estimation methods and found no significant difference in performance.

## 5. EXPERIMENTAL RESULTS

This section presents experiments that examine acoustic feature estimation in noisy conditions and under speaker variability. Experiments use first a speaker dependent database to examine specifically the effectiveness of noise adaptation. This database contains 579 phonetically balanced sentences for training and 246 for testing (130,000 test vectors) spoken by a female US English speaker, each with a duration of between 5 and 9 seconds. Along with the audio, laryngograph recordings were also made so the reference fundamen-

tal frequency and voicing are very accurate. The reference spectral envelope was estimated using filterbank analysis on the clean speech. Approximately 65% of frames were voiced, 20% unvoiced and 15% non-speech. A second set of experiments uses speaker independent data and now examines both noise and speaker adaptation. These use 22 hours of the WSJCAM0 database for training, taken from 48 female and 63 male speakers, and a further hour for testing, spoken by a different set of 5 female and 5 male speakers containing 360,000 test vectors (Robinson et al., 1995). No laryngograph recordings were available and instead the PRAAT tool was applied to clean speech to give reference fundamental frequency and voicing that were checked manually for voicing errors and halving and doubling errors (Boersma and Weenik, 2001). Approximately 50% of frames were voiced, 30% unvoiced and 20% non-speech.

The audio was sampled at 8kHz and tests carried out in white noise, babble noise and destroyer operations room noise, added at SNRs between -5dB and +15dB as an aim of the work is robustness in high noise levels (Varga and Steeneken, 1993). White noise was chosen because of its stationary character and babble and destroyer noises because of their non-stationary character. Together these give varied test conditions across a range of potential environments and provide useful analysis. This section now considers fundamental frequency estimation, voicing classification and spectral envelope estimation, and within each, first speaker-dependent and then speaker-independent performance.

### 5.1. Fundamental frequency estimation

Fundamental frequency error, $E_{f_0}$, is measured as

$$E_{f_0} = \frac{1}{N_V} \sum_{i=1}^{N} \left( \frac{|\hat{f}_{0i} - f_{0i}|}{f_{0i}} \right) \times 100\% \quad \forall f_{0i} > 0 \tag{23}$$

$\hat{f}_{0i}$ and $f_{0i}$ are the estimated and reference fundamental frequencies for the $i$th voiced frame and $N_V$ is the number of voiced frames. To avoid analysis being affected by voicing errors, $E_{f_0}$ is measured on only frames classified as voiced.

#### 5.1.1. Speaker dependent

Speaker dependent tests first examine the effect of the number of filterbank channels and number of mixture components in the GMM. Filterbanks from 16 to 128 channels and mixture components from 1 to 512 were tested across a range of SNRs from -5dB to clean conditions. As an example, Fig. 4 shows $E_{f_0}$ in white noise at an SNR of 10dB. Increasing the number of mixture components reduces estimation error considerably until approximately 128 after which gains become minimal. Increasing the filterbank from 16 to 32 channels reduces error but beyond that no further reduction is observed. Similar patterns of results were obtained at other SNRs and noises. Taken across all test conditions, lowest errors were with a 32 channel filterbank and 256 mixture components and this configuration is used for the remainder of fundamental frequency estimation.

Noise adaptation is now examined and no speaker adaptation applied as the tests use speaker dependent data. Fig. 5 shows $E_{f_0}$ in white, babble and destroyer noises at SNRs from -5dB to +15dB for clean trained models, models

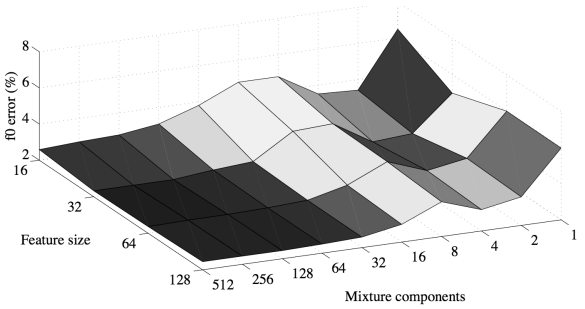



Figure 4: Speaker dependent $E_{f0}$ for 16 to 128 filterbank channels and 1 to 512 mixture components in white noise at an SNR of 10dB.

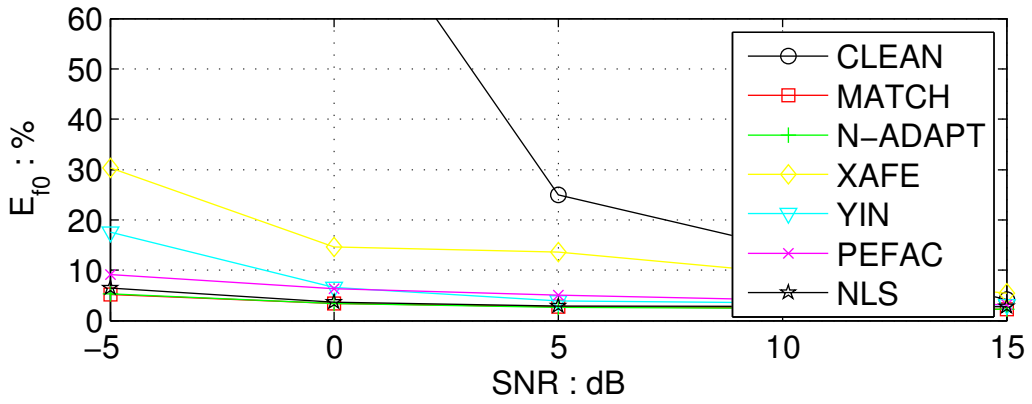
trained and tested under the same noise conditions and models adapted to noise. $E_{f_0}$ was 2.1% in clean conditions. For comparison, the performance of the ETSI XAFE (ETSI, 2003), YIN (de Cheveigné and Kawahara, 2002), PEFAC (Gonzalez and Brookes, 2014) and parametric fast NLS (Nielsen et al., 2016) methods are also shown. These had $E_{f_0}$ of 2.5%, 3.2%, 3.8% and 2.2%, respectively in clean conditions. (Note - the NLS method assumes white noise, so for testing in babble and destroyer noises pre-whitening is applied). Estimates from the clean trained model (CLEAN), whilst accurate in clean conditions, deteriorate rapidly as SNRs fall as the input speech statistics become unmatched to those in training. Matching the statistics of the model (MATCH) to the noisy speech by training and testing in the same noise conditions gives a substantial reduction in error. Using noise adaptation (N-ADAPT) to adjust the clean trained model to the noise characteristics is highly effective with $E_{f_0}$ almost identical to matched models. For example, at an SNR of -5dB in white noise the adapted models had $E_{f_0}$ of 5.3% which is an increase of just 3.2% over that obtained in clean conditions. In comparison, $E_{f_0}$ for YIN, XAFE, PEFAC and NLS are 17.6%, 30.3%, 9.2%

and 6.5%, respectively. In fact, estimates from these comparative methods were consistently worse than the proposed method. We did examine whether accuracy could be improved by applying speech enhancement (both log MMSE and Wiener filtering (Ephraim and Malah, 1985; Scalart and Vieira-Filho, 1996)) prior to $f_0$ estimation but found this to reduce performance which we attribute to distortion introduced, particularly at lower SNRs.

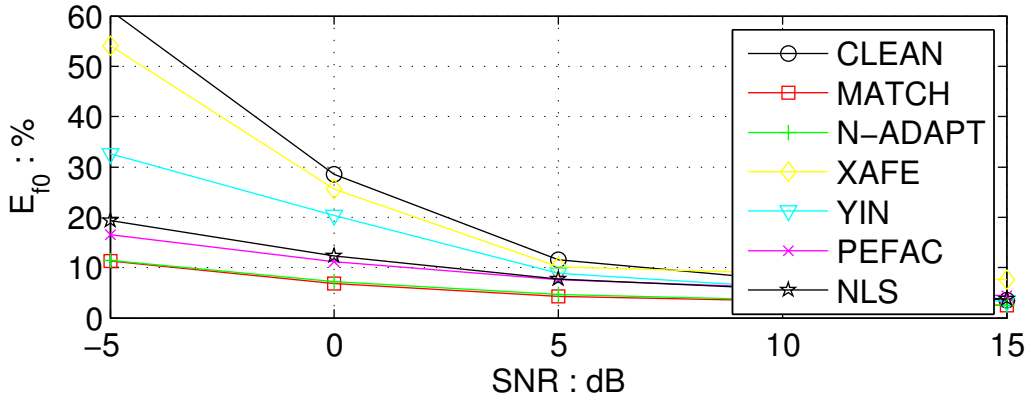
*5.1.2. Speaker independent*

For speaker independent estimation, noise adaptation is identical to that for speaker-dependent models, and speaker adaptation is also used to adjust the statistics of the model to the speaker under test. Fig. 6 shows $E_{f_0}$ for clean trained models, models adapted to noise and then models adapted to speaker and noise. $E_{f0}$ in clean conditions was 3.8%. Accuracy using XAFE, YIN, PEFAC and NLS are also shown and these had $E_{f_0}$ of 8.4%, 3.6%, 3.8% and 3.7%, respectively, in clean conditions. Adapting the speech models to noise (N-ADAPT) again gives substantial improvement over clean trained models. Errors are further reduced when the models are adapted to both speaker and noise (NS-ADAPT) and this gives best performance of all the methods tested and was almost identical to that of matched conditions (not shown to improve clarity of plots).

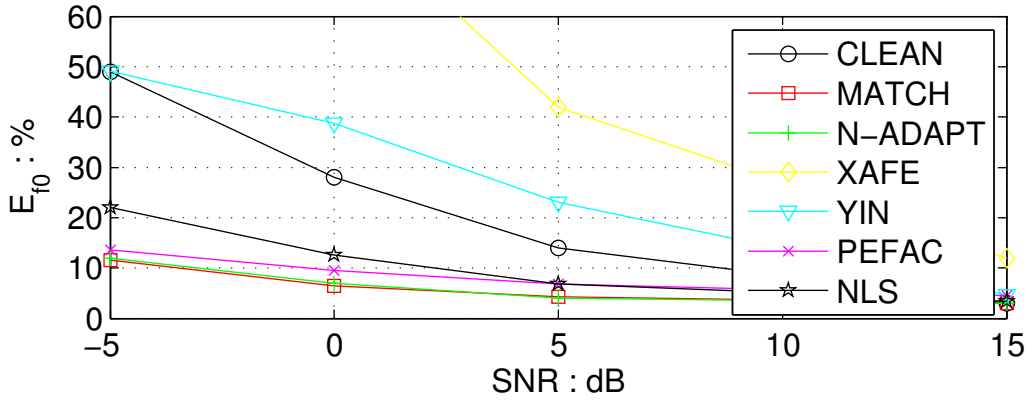
In comparison, YIN and PEFAC are effective in white noise, with errors just higher than adapted models, but less so in the non-stationary noises. NLS is even more effective in white noise, achieving lower errors than YIN and PEFAC, but slightly higher than adapted models, and is also less accurate in babble and destroyer noises. XAFE deteriorates rapidly in both noise types as SNRs fall below 15dB. For example, at an SNR of -5dB in white noise, $E_{f0}$




(a)




(b)




(c)

Figure 5: Speaker dependent $E_{f_0}$ in a) white noise, b) babble noise and c) destroyer operations room noise, at SNRs from -5dB to +15dB for clean trained models (CLEAN), matched models (MATCH), noise adapted models (N-ADAPT), XAFE, YIN, PEFAC and NLS. Note - to improve clarity at lower $E_{f_0}$ values the ordinate is limited to 60%.

using adaptation is 10.2%, while for YIN $E_{f_0}$ is 17.9%, for PEFAC is 14.9% and for NLS is 12.0%. Again, we found that applying speech enhancement increased $E_{f_0}$.

### *5.2. Voicing classification*

Voicing classification error, $E_V$, is measured as

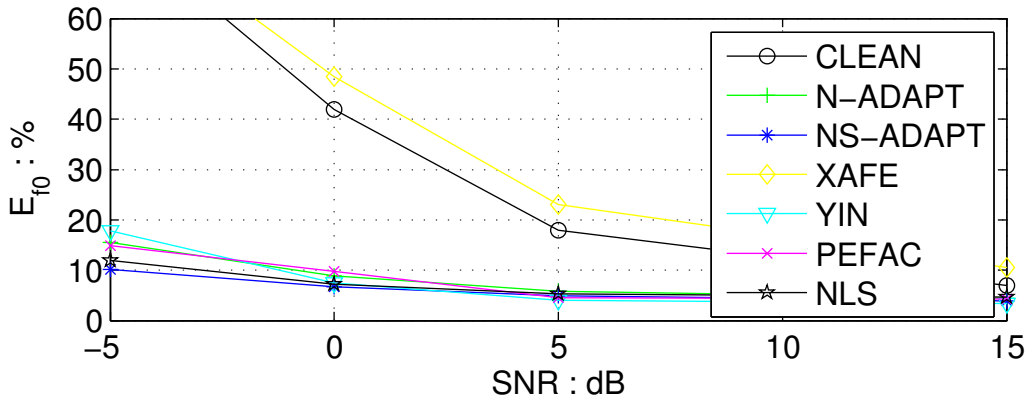
$$E_V = \frac{N_{\overline{V}|V} + N_{\overline{UV}|UV} + N_{\overline{NS}|NS}}{N_T} \times 100\% \tag{24}$$

$N_{\overline{V}|V}$, $N_{\overline{UV}|UV}$ and $N_{\overline{NS}|NS}$ are the number of voiced, unvoiced and non-speech frames classified incorrectly and $N_T$ is the total number of frames under test.

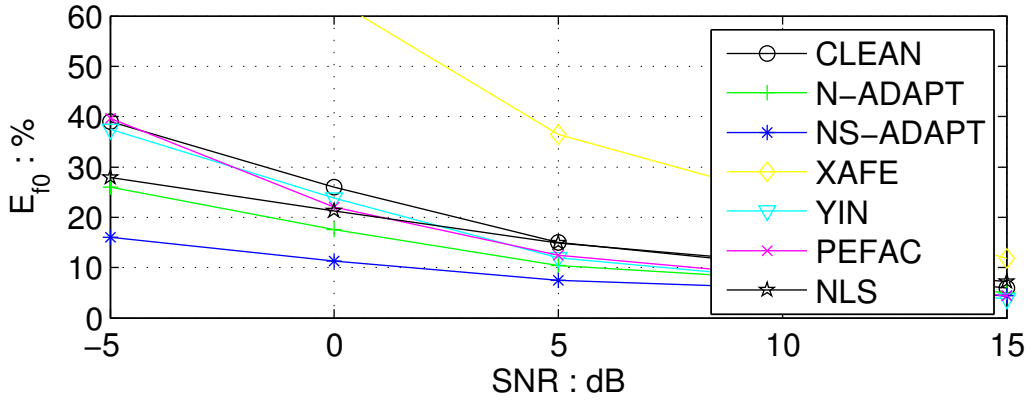
#### *5.2.1. Speaker dependent*

An initial test, similar to that for fundamental frequency estimation in Section 5.1.1, investigated the effect of the size of filterbank and number of mixture components. $E_V$ was minimised with 8 channels and 16 mixture components and beyond these values the error remained unchanged. This is considerably fewer channels than minimised $E_{f0}$ and is attributed to less fine spectral detail needed to discriminate between voicing classes. Similarly, the GMM requires fewer mixture components than needed to minimise $E_{f0}$. Section 5.1 determined that fundamental frequency error was minimised with 32 filterbank channels and 256 mixture components. Given that no reduction in voicing classification error was observed beyond 8 channels and 16 mixture components the same GMMs used for fundamental frequency estimation can also be used for voicing classification which gives a single statistical framework for both acoustic features.

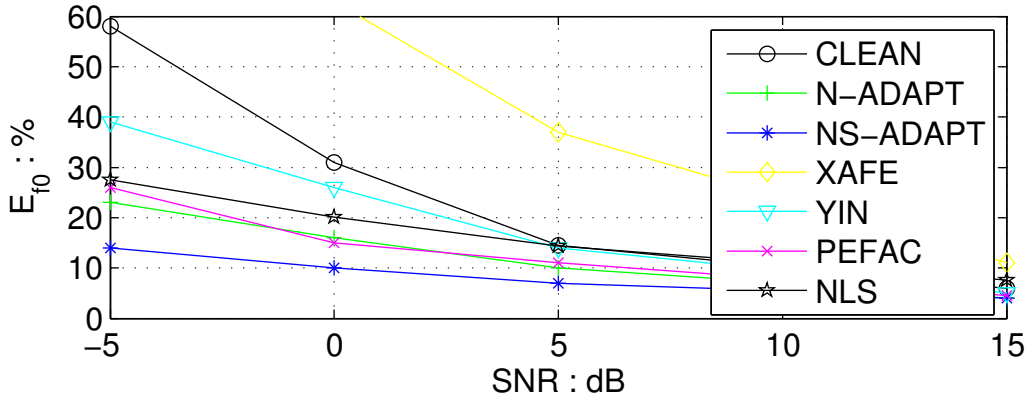
Fig. 7 shows speaker-dependent voicing error in white, babble and destroyer noises at SNRs from -5dB to +15dB for clean trained models,




(a)




(b)



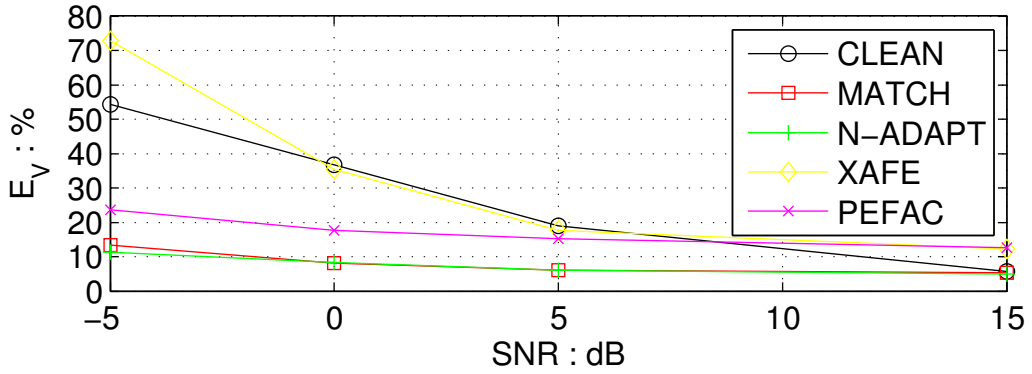

(c)

Figure 6: Speaker independent $E_{f_0}$ in a) white noise, b) babble noise and c) destroyer operations room noise, at SNRs from -5dB to +15dB for clean trained models (CLEAN), noise adapted models (N-ADAPT), noise and speaker adapted models (NS-ADAPT), XAFE, YIN, PEFAC and NLS. Note - to improve clarity at lower $E_{f_0}$ values the ordinate is limited to 60%.

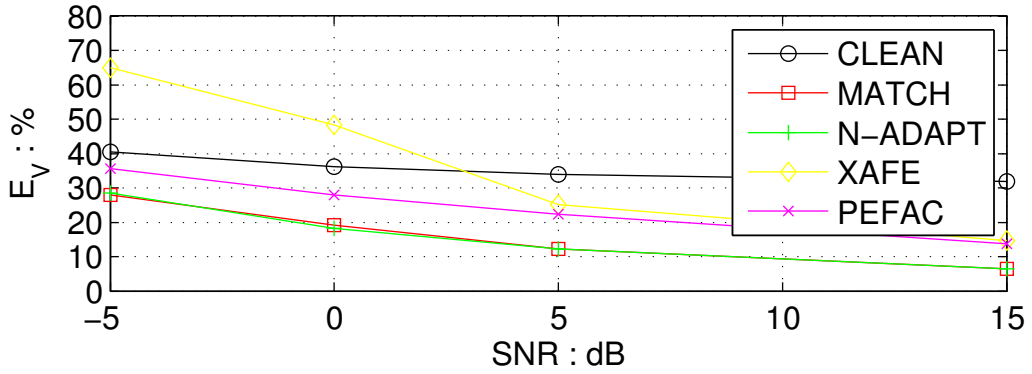
matched models and models adapted to noise. Voicing error in clean conditions is 4.5%. For comparison, voicing error is shown for XAFE and PEFAC which have $E_V$ of 6.1% and 10.8% in clean conditions. Using clean trained models (CLEAN), voicing errors increase rapidly as SNRs reduce. The majority of these errors are non-speech and unvoiced frames being incorrectly classified as speech due to their increased energy matching better to the higher energy of the voiced model. Adapting the clean models to noise (N-ADAPT) reduces error substantially and performance is now indistinguishable from models trained under matched noise conditions (MATCH). $E_V$ deteriorates rapidly for XAFE whilst PEFAC remains much more robust and has $E_V$ about 10% higher than the adapted models.

*5.2.2. Speaker independent*

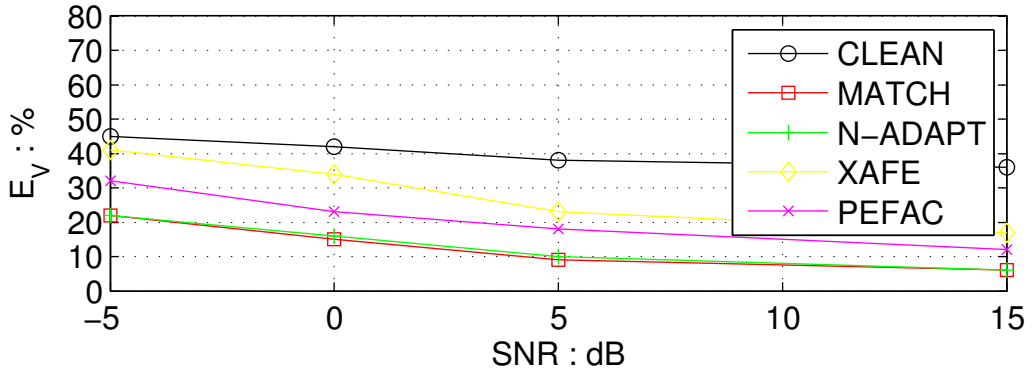
Fig. 8 shows $E_V$ for speaker-independent testing in white, babble and destroyer noises using clean trained models, noise adapted models and noise and speaker adapted models. In clean conditions, $E_V$ is 7.1% with clean speaker independent models which reduces to 6.3% when adapted to the speaker under test. Results using XAFE and PEFAC are also shown which have $E_V$ of 10.8% and 11.0%. Compared to clean models (CLEAN), noise adaptation (N-ADAPT) gives a substantial reduction in error and a further reduction when also applying speaker adaptation (NS-ADAPT). However, the reduction in error with speaker adaptation is much less than observed in fundamental frequency estimation. PEFAC and XAFE introduce significantly more errors as SNRs fall – for example at -5dB in white noise, $E_V$ for PEFAC is 19.2% compared to 12.5% for adapted models.




(a)

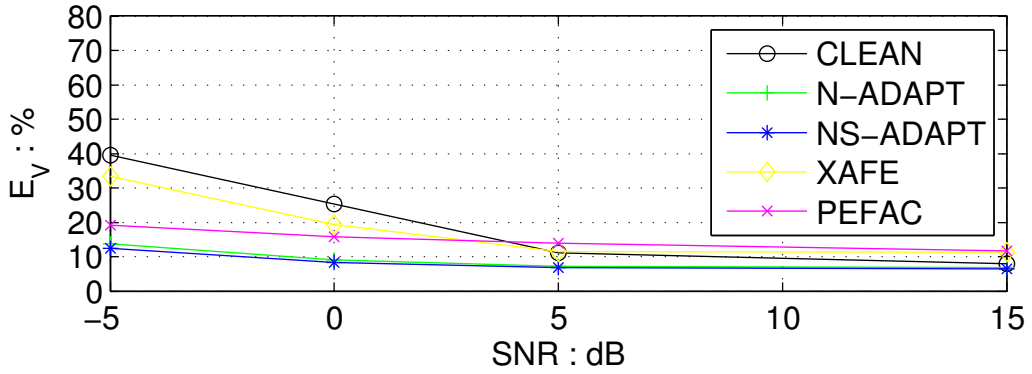



(b)

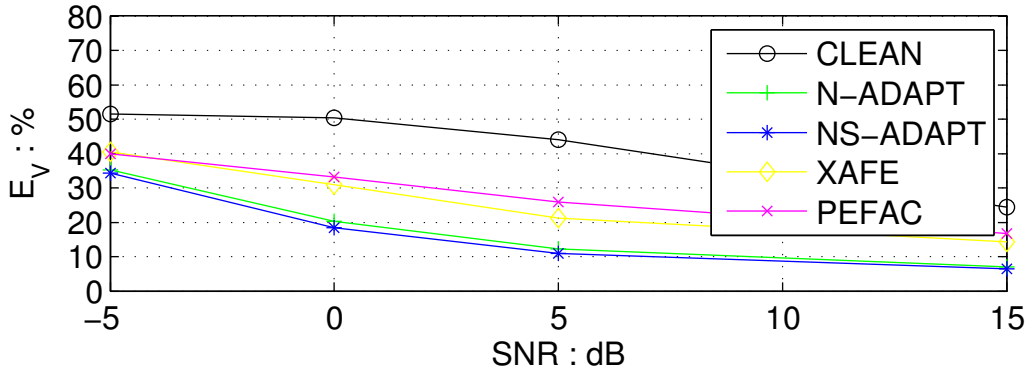



(c)

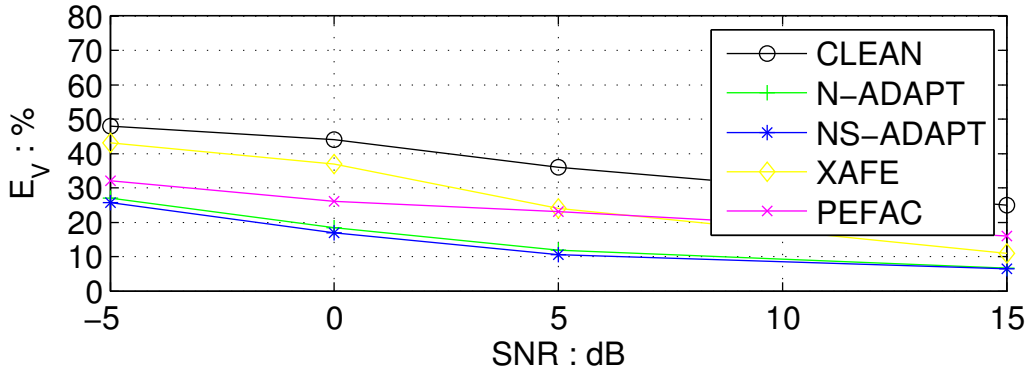
Figure 7: Speaker dependent $E_V$ in a) white noise, b) babble noise and c) destroyer noise at SNRs from -5dB to +15dB for clean trained models, noise adapted models, matched models, XAFE and PEFAC.




(a)




(b)




(c)

Figure 8: Speaker independent $E_V$ in a) white noise, b) babble noise and c) destroyer noise at SNRs from -5dB to +15dB for clean trained models, noise adapted models, noise and speaker adapted models, XAFE and PEFAC

*5.3. Spectral envelope estimation*

The log likelihood ratio (LLR) is used to evaluate spectral envelope estimation as it is relatively insensitive to fine spectral detail from harmonic structure and more sensitive to spectral envelope. A low LLR indicates a closer spectral match and is defined (Loizou, 2007)

$$LLR = \frac{1}{N}\sum_{i=1}^{N} log\left(\frac{\hat{\boldsymbol{b}}_x^T(i)\boldsymbol{R}_{xx}(i)\hat{\boldsymbol{b}}_x(i)}{\boldsymbol{b}_x^T(i)\boldsymbol{R}_{xx}(i)\boldsymbol{b}_x(i)}\right) \tag{25}$$

where $\boldsymbol{R}_{xx}(i)$ and $\boldsymbol{b}_x(i)$ are the autocorrelation matrix and LPC coefficient vector computed from the $i$th frame of the the original clean speech signal. $\hat{\boldsymbol{b}}_x(i)$ is the LPC coefficient vector found by inverting the estimated filterbank vector, $\hat{\chi}_i$, into a smoothed power spectrum using cubic spline interpolation and taking a inverse Fourier transform. Initial tests examined the effect of filterbank and mixture components with best performance obtained using 32 channels and 256 mixture components. This is the same as the best configuration for fundamental frequency estimation and allows all three acoustic features to be estimated from the same filterbank/GMM configuration.

*5.3.1. Speaker dependent*

Fig. 9 shows LLRs for clean trained models, matched models and models adapted to noise in white, babble and destroyer noises at SNRs from -5dB to 15dB. For comparison LLRs were also computed from noisy speech that had been enhanced by spectral subtraction, Wiener filtering and log MMSE methods of speech enhancement (Berouti et al., 1979; Ephraim and Malah, 1985; Loizou, 2007; Scalart and Vieira-Filho, 1996). In each case the noisy signal was input into the enhancement method and LLRs computed

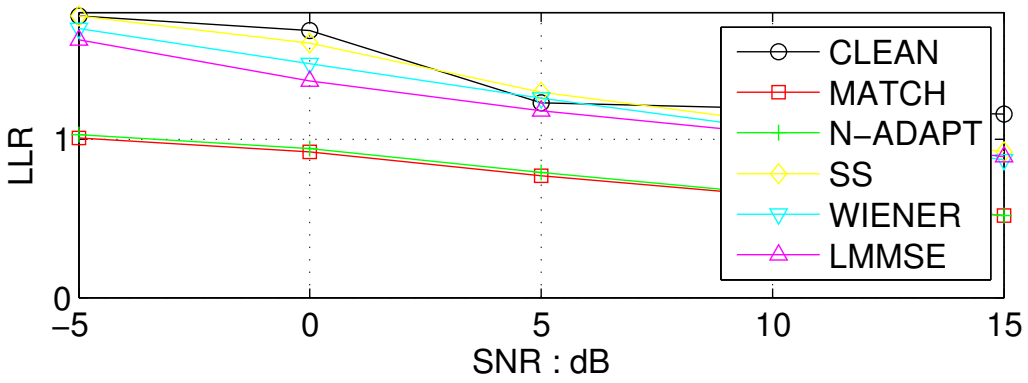
from the enhanced speech. LLRs using clean trained models (CLEAN) increased rapidly as SNRs fell but adapting the models to noise reduced LLRs substantially. In comparison to matched models (MATCH), which attained lowest LLRs, adapted models perform almost identically. Adaptation (N-ADAPT) performed much better than the three enhancement methods, which had a ranking of LLRs equal to that reported when measuring their respective speech quality (Loizou, 2007). Similar to Fig. 3, further tests varied the amount of noise adaptation data and found LLRs to converge after 1.5 seconds for white noise and 2.5 seconds for babble and destroyer noises.

*5.3.2. Speaker independent*

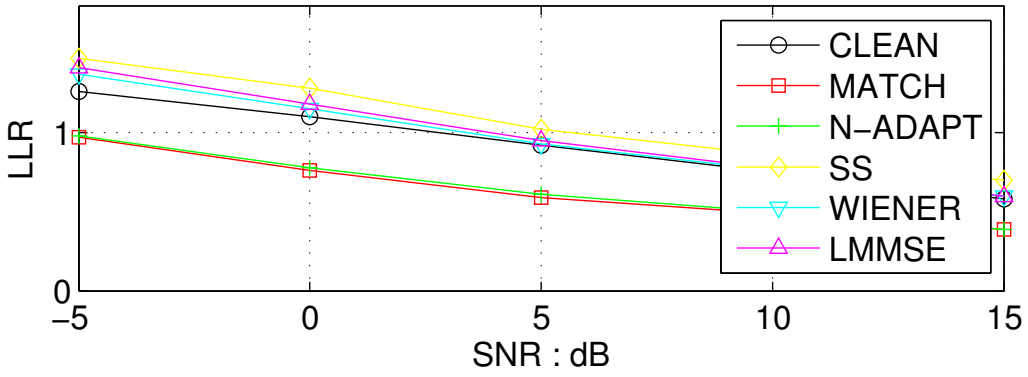
Fig. 10 shows LLRs for speaker-independent testing using clean models, noise adaptation, speaker and noise adaptation, spectral subtraction, Wiener filtering and log MMSE. Speaker independent results follow a similar trend to speaker-dependent testing where adapting the clean models to noise (N-ADAPT) gives a large reduction in LLR, and a further, but smaller, reduction when applying noise and speaker adaptation (NS-ADAPT). This is substantially lower than spectral subtraction, Wiener or log MMSE. For example, at an SNR of -5dB in white noise, the LLR for noise adapted models is 0.89 which is further reduced to 0.83 when adapting to both speaker and noise which compares to 1.40 for log MMSE.

## 6. Discussion

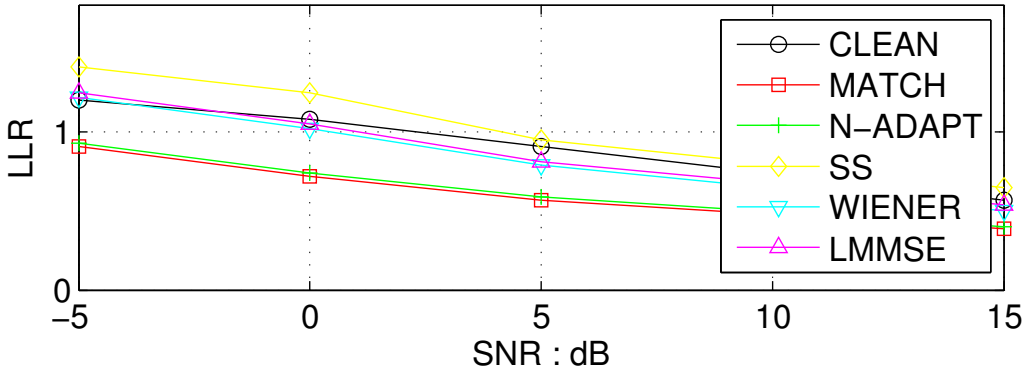
In noise-free conditions the clean trained models provide accurate estimates of acoustic features and outperform the comparative methods tested. As noise increases the statistics of the clean models become mismatched to




(a)

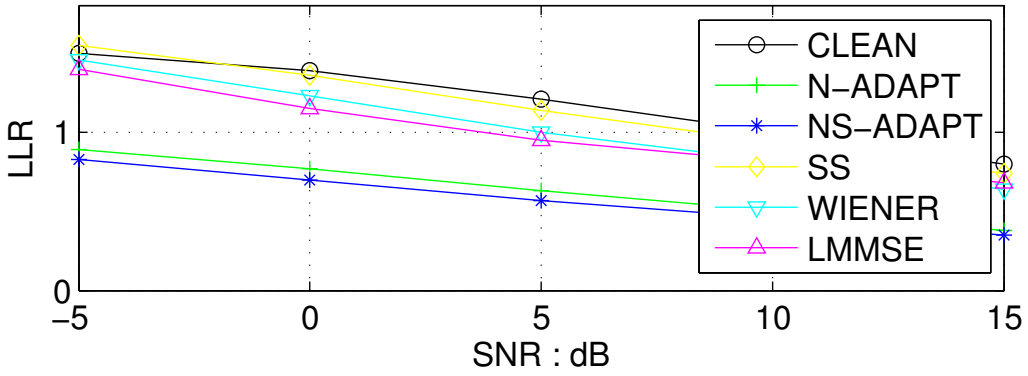



(b)

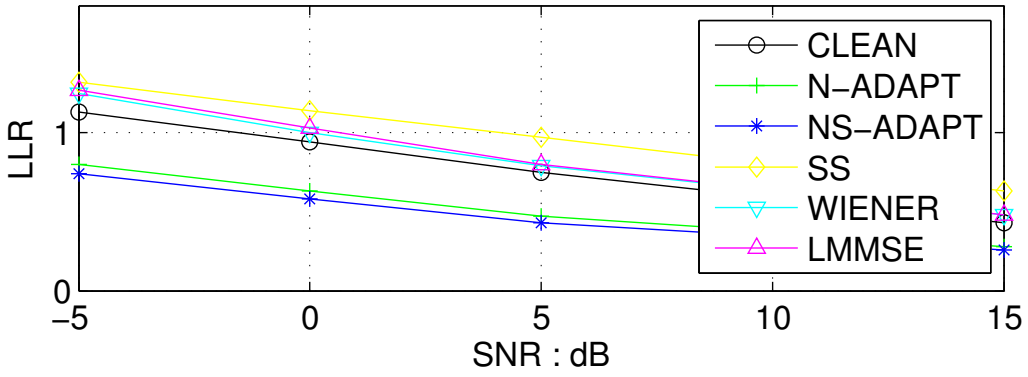



(c)

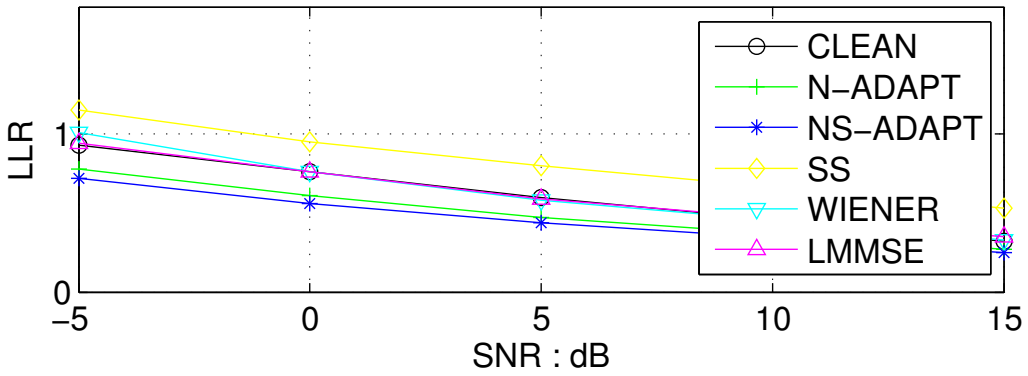
Figure 9: Speaker dependent LLRs in a) white noise, b) babble noise and c) destroyer noise, at SNRs from -5dB to 15dB for clean trained models, matched models, noise adapted models, spectral subtraction, Wiener filtering and log MMSE.




(a)




(b)




(c)

Figure 10: Speaker independent LLRs in a) white noise, b) babble noise and c) destroyer noise at SNRs from -5dB to 15dB for clean trained models, noise adapted models, noise and speaker adapted models, spectral subtraction, Wiener filtering and log MMSE.

the noisy speech and performance deteriorates rapidly. Training in the same noise conditions, whilst unrealistic practically, gives substantial improvements and is considered the target for performance.

Across all noises, and for speaker dependent and speaker independent systems, adapting the clean models to the noise conditions reduces the mismatch and gives large reductions in error that are close, and in some cases equal, to those of matched condition training. Applying speaker adaptation to the speaker independent models further reduces errors for fundamental frequency estimation but has less effect on voicing classification and spectral envelope estimation. This is attributed to the speaker independent distribution of fundamental frequency being broad while that of a single speaker is much more localised – this is illustrated in Fig. 1. Adapting the speaker independent acoustic models to the fundamental frequency range of the speaker under test is localises the estimation and improves accuracy. Conversely, for spectral envelope features, there is less speaker-specific variation and therefore adapting the speaker independent distribution to a new speaker has less effect which is reflected in the lower improvement in estimation accuracy.

## 7. Conclusion

This work has shown that the statistical framework proposed for estimating acoustic speech features is effective in clean conditions but deteriorates rapidly as SNRs fall and the models become mismatched to the test conditions. Analysis has shown that using models that have been prior trained to match the testing conditions gives best performance but is impractical from a practical perspective as noise and speaker characteristics change. Instead,

the proposed method of adapting the models to the current noise and speaker conditions has been shown effective and able to attain error rates close to, and in some cases equal to, that of the models trained under matched conditions. In comparison to a range of existing methods for estimating acoustic features the proposed method achieved lowest errors across both the stationary and non stationary noise conditions and the range of SNRs tested from -5dB to +15dB.

## References


Berouti, M., Schwartz, R., Makhoul, J., Apr. 1979. Enhancement of speech corrupted by acoustic noise. In: ICASSP. Vol. 4. USA, pp. 208–211.

Boersma, P., Weenik, D., 2001. Praat, a system for doing phonetics by computer. Tech. Rep. 132, Institute of Phonetic Sciences, University of Amsterdam,.

Cappe, O., Moulines, E., 1996. Regularization techniques for discrete cepstrum estimation. IEEE Signal Processing Letters 3 (4), 100–102.

Christensen, M., Oct. 2013a. Accurate estimation of low fundamental frequencies from real-valued measurements. IEEE Trans. Audio, Speech and Language Processing 21 (10), 2042–2056.

Christensen, M., May 2013b. Metrics for vector quantisation-based parametric speech enhancement and separation. Journal of the Acoustical Society of America 133 (5), 3063–3071.

Christensen, M., Jakobsson, A., 2009. Multi-pitch estimation. Synthesis Lectures on Speech and Audio Processing 5.

Chung, Y., Hansen, J., 2013. Compensation of SNR and noise type mismatch using an environmental sniffing based speech recognition solution. EURASIP Journal on Audio, Speech, and Music Processing 2013 (1), 1–14.

Darch, J., 2008. Robust acoustic speech feature prediction from mel frequency cepstral coefficients. Ph.D. thesis, University of East Anglia, UK, Norwich, UK.

Davis, S., Mermelstein, P., Aug. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoustics, Speech and Signal Processing 28 (4), 357–366.

de Cheveigné, A., Kawahara, H., Apr. 2002. YIN, a fundamental frequency estimator for speech and music. Journal of the Acoustical Society of America 111 (4), 1917–1930.

Dhananjaya, N., Yegnanarayana, B., 2013. Voiced/nonvoiced detection based on robustness of voiced epochs. IEEE Signal Processing Letters 17 (3), 273–276.

Ephraim, Y., Malah, D., Apr. 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoustics, Speech and Signal Processing 33 (2), 443–445.

ETSI, Nov. 2003. Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Extended advanced front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm. ES 202 212 version 1.1.1, ETSI STQ-Aurora DSR Working Group.

Faubel, F., McDonough, J., Klakow, D., 2008. A phase-averaged model for the relationship between noisy speech, clean speech and noise in the log-mel domain. In: Interspeech. pp. 553–556.

Gales, M., Young, S., Sep. 1996. Robust continuous speech recognition using parallel model combination. IEEE Trans. Speech and Audio Processing 4 (5), 352–359.

Gauvain, J.-L., Lee, C.-H., Apr. 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. Speech and Audio Processing 2 (2), 291–298.

Geiger, F., Wallhoff, F., Rigoll, G., 2010. GMM-UBM based open-set online speaker diarization. In: Interspeech. pp. 2330–2333.

Godsill, S., Davy, M., 2002. Bayesian harmonic models for musical pitch estimation and analysis. In: ICASSP. pp. 1769–1772.

Gonzalez, S., Brookes, M., Feb. 2014. PEFAC - a pitch estimation algorithm robust to high levels of noise. IEEE/ACM Trans. Audio, Speech and Language Processing 22 (2), 518–530.

Harding, P., 2013. Model based speech enhancement. Ph.D. thesis, University of East Anglia, UK.

Harding, P., Milner, B., Sep. 2012. On the use of machine learning methods for voice activity detection. In: Interspeech. Portland, USA, pp. 709–712.

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. Journal of the Acoustical Society of America 87 (4), 1738–1752.

Hirahara, T., Nov. 1988. On the role of the fundamental frequency in vowel perception. Second Joint Meeting of ASA and ASJ, Journal of the Acoustical Society of America 84 (S1), S156.

Hu, Y., Huo, Q., 2006. An HMM compensation approach using unscented transformation for noisy speech recognition. Lecture Notes in Computer Science, Chinese Spoken Language Processing 4274, 346–357.

Kaewtip, K., Tan, L. N., Alwan, A., 2013. A pitch-based spectral enhancement technique for robust speech processing. In: Interspeech. pp. 3284–3288.

Kawahara, H., Estill, J., Fujimura, O., Sep. 2001. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. In: Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA).

Kawahara, H., Nisimura, R., Irino, T., Morise, M., Takahashi, T., Banno, H., 2009. Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown. In: ICASSP. pp. 3905–3908.

Koriyama, T., Nose, T., Kobayashi, T., 2014. Statistical parametric speech synthesis based on gaussian process regression. IEEE Selected Topics in Signal Processing 8 (2), 173–183.

Lei, X., Siu, M., Hwang, M.-Y., Ostendorf, M., Lee, T., 2006. Improved tone modeling for Mandarin broadcast news speech recognition. In: Interspeech. pp. 1237–1240.

Li, H., Stoica, P., Li, J., 2000. Computationally efficient parameter estimation for harmonic sinusoidal signals. Elsevier Signal Processing 80 (9), 1937–1944.

Loizou, P., 2007. Speech Enhancement: Theory and Practice. CRC Press, Inc.

Ma, N., Green, P., Barker, J., Coy, A., Dec. 2007. Exploiting correlogram structure for robust speech recognition with multiple speech sources. Speech Communication 49 (12), 874–891.

Makhoul, J., 1975. Linear prediction: a tutorial review. Proc. IEEE 63 (4), 561–580.

Martin, R., Jul. 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech and Audio Processing 9 (5), 504–512.

McAulay, R., Champion, T., Apr. 1990. Improved interoperable 2.4 kb/s LPC using sinusoidal transform coder techniques. In: ICASSP. Vol. 2. Albuquerque, NM, USA, pp. 641–643, DOI: 10.1109/ICASSP.1990.115818.

Milner, B., 2002. A comparison of front-end configurations for robust speech recognition. In: ICASSP. pp. 797–800.

Milner, B., Darch, J., Feb. 2011. Robust acoustic speech feature prediction from noisy mel-frequency cepstral coefficients. IEEE Trans. Audio, Speech and Language Processing 19 (2), 338–347.

Morales-Cordovilla, J. A., Ma, N., Sanchez, V., Carmona, J. L., Peinado, A. M., Barker, J., May 2011a. A pitch based noise estimation technique for robust speech recognition with missing data. In: ICASSP. Czech Republic, pp. 4808–4811.

Morales-Cordovilla, J. A., Peinado, A. M., Sanchez, V., Gonzalez, J. A., Apr. 2011b. Feature extraction based on pitch-synchronous averaging for robust

speech recognition. IEEE Trans. Audio, Speech and Language Processing 19 (3), 640–651.

Moreno, P., Raj, B., Stern, R., 1996. A vector Taylor series approach for environment-independent speech recognition. In: ICASSP. pp. 733–736.

Nielsen, J. K., Jensen, T. L., Jensen, J. R., Christensen, M. G., Jensen, S. H., 2016. Fast and statistically efficient fundamental frequency estimation. In: ICASSP. pp. 86–90.

Oppenheim, A., Schafer, R., 1975. Digital Signal Processing. Prentice Hall.

Rangachari, S., Loizou, P., Feb. 2006. A noise estimation algorithm for highly nonstationary environments. Speech Communication 48 (2), 22—231.

Robinson, T., Fransen, J., Pye, D., Foote, J., Renals, S., 1995. WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition. In: ICASSP. Vol. 1. pp. 81–84.

Scalart, P., Vieira-Filho, J., 1996. Speech enhancement based on a priori signal to noise estimation. In: ICASSP. pp. 629–632.

Schäck, T., Sledz, C., Muma, M., Zoubir, A. M., 2015. A new method for heart rate monitoring during physical exercise using photoplethysmographic signals. In: EUSIPCO. pp. 2666–2670.

Soong, F., 1984. Line spectrum pair (LSP) and speech data compression. In: ICASSP. pp. 37–40.

Stoica, P., Söderström, T., 1989. On reparameterization of loss functions used in estimation and the invariance principle. Elsevier Signal Processing 17, 383–387.

Swindlehurst, A., Stoica, P., 1998. Maximum likelihood methods in radar array signal processing. Proceedings of the IEEE 86 (2), 421–441.

Syrdal, A., Steele, S., Nov. 1985. Vowel F1 as a function of speaker fundamental frequency. 110th Meeting of ASA, Journal of the Acoustical Society of America 78 (S1), S56.

Tabrikian, J., Dubnov, S., Dickalov, Y., Jan. 2004. Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model. IEEE Trans. Speech and Audio Processing 12 (1), 76–87.

Taghia, J., Taghia, J., Mohammadiha, N., Sang, J., Bouse, V., Martin, R., 2011. An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments. In: ICASSP. pp. 4640–4643.

Talkin, D., 1995. A robust algorithm for pitch tracking (RAPT). In: Kleijn, W., Paliwal, K. (Eds.), Speech Coding and Synthesis. Elsevier, Ch. 14, ISBN: 0-444-82169-4.

Varga, A., Steeneken, H., Jul. 1993. Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Communication 12 (3), 247–251.

Vaseghi, S., Milner, B., Jan. 1997. Noise compensation methods for hidden

Markov model speech recognition in adverse environments. IEEE Trans. Speech and Audio Processing 5 (1), 11–21.

Woodland, P., Aug. 2001. Speaker adaptation for continuous density HMMs: A review. In: ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition.