# Analysis and prediction of acoustic speech features from mel-frequency cepstral coefficients in distributed speech recognition architectures

Jonathan Darch and Ben Milner[a)]
*School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, United Kingdom*

Saeed Vaseghi
*Department of Electronic and Computer Engineering, Brunel University, Uxbridge UB8 3PH, United Kingdom*

(Received 19 December 2007; revised 5 September 2008; accepted 10 September 2008)

The aim of this work is to develop methods that enable acoustic speech features to be predicted from mel-frequency cepstral coefficient (MFCC) vectors as may be encountered in distributed speech recognition architectures. The work begins with a detailed analysis of the multiple correlation between acoustic speech features and MFCC vectors. This confirms the existence of correlation, which is found to be higher when measured within specific phonemes rather than globally across all speech sounds. The correlation analysis leads to the development of a statistical method of predicting acoustic speech features from MFCC vectors that utilizes a network of hidden Markov models (HMMs) to localize prediction to specific phonemes. Within each HMM, the joint density of acoustic features and MFCC vectors is modeled and used to make a maximum *a posteriori* prediction. Experimental results are presented across a range of conditions, such as with speaker-dependent, gender-dependent, and gender-independent constraints, and these show that acoustic speech features can be predicted from MFCC vectors with good accuracy. A comparison is also made against an alternative scheme that substitutes the higher-order MFCCs with acoustic features for transmission. This delivers accurate acoustic features but at the expense of a significant reduction in speech recognition accuracy. © *2008 Acoustical Society of America.*
[DOI: 10.1121/1.2997436]

PACS number(s): 43.72.Ar [DOS] Pages: 3989–4000

## I. INTRODUCTION

Acoustic speech features, namely, formants, fundamental frequency, and voicing, are traditionally estimated from time-domain waveforms of speech or some other representation that is derived from the time domain. For example, effective methods of fundamental frequency estimation apply autocorrelation analysis, frequency-domain analysis, or cepstral analysis to extract candidates that can be tracked and smoothed using dynamic programing.[1] Similarly, formant frequencies can be estimated from pole positions obtained from linear predictive analysis or from spectral peaks, followed by tracking, using, for example, Kalman filtering.[2]

In recent years, distributed speech recognition (DSR) architectures have been developed as a robust method for achieving accurate speech recognition over mobile and IP networks.[3] However, within a DSR architecture, no time-domain waveform is sent to the remote back-end. Instead, only a stream of mel-frequency cepstral coefficient (MFCC) vectors is received, which prohibits the use of conventional methods of acoustic feature estimation. In many situations, it is desirable to have acoustic features available at the remote back-end. An example is to enable a speech signal to be reconstructed at the back-end from the received MFCC vectors, as may be required in cases of legal dispute of potential speech recognition errors. The MFCC vectors themselves can be inverted to provide spectral envelope information, but to facilitate speech reconstruction, source information such as voicing and fundamental frequency are also required. One solution has been to explicitly estimate and transmit voicing and fundamental frequency from the terminal device to the remote back-end.[3] While this provides the necessary information, it creates considerable overheads both in terminal-side processing and in additional bit-rate requirements. For example, the ETSI Aurora standard uses 800 bits/s to represent voicing and fundamental frequency. An alternative approach is to predict voicing and fundamental frequency from the received MFCC vectors themselves.[4,5] This has been achieved by modeling the joint density of MFCC vectors and the fundamental frequency to enable a maximum *a posteriori* (MAP) prediction of fundamental frequency from a MFCC vector. This has been applied first to a constrained connected digit vocabulary[4] and then to unconstrained free speech.[5] Using a similar probabilistic framework, formant frequencies have also been successfully predicted from MFCC vectors.[6]

The work presented in this paper extends the previous work in four ways. First, instead of predicting fundamental frequency/voicing and formant frequencies separately, an acoustic feature vector is defined, which comprises fundamental frequency and formant frequencies together with voicing and speech/nonspeech information.[7] This allows a simultaneous prediction of the set of acoustic features from a

[a)]Electronic mail: b.milner@uea.ac.uk



Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 134.83.1.241 On: Tue, 19 Aug 2014 09:49:58

MFCC vector and is achieved by defining three speech classes (voiced, unvoiced, and nonspeech). Second, a comprehensive analysis of the correlation between acoustic features and MFCC vectors is made. This examines the effect on the acoustic feature to MFCC vector correlation when applying constraints such as measuring correlation within individual speakers (speaker dependent), within gender types (gender dependent), and across all speakers (gender independent). Restrictions into the speech class are also made, which consider correlation either globally across all speech sounds or within individual phoneme classes. Third, a detailed analysis into the accuracy of acoustic feature prediction is made. This examines prediction accuracy under speaker-dependent, gender-dependent, and gender-independent constraints and allows conclusions to be drawn into the effect that a high acoustic feature to MFCC vector correlation has on prediction accuracy. Finally, a comparison of the proposed method is made against a scheme that replaces higher-order MFCCs in the transmitted feature vector by acoustic features (estimated on the terminal device). The acoustic features are compressed using the same method as the MFCCs, and the effect of this on acoustic feature estimation is measured, together with the effect on recognition accuracy of losing higher-order MFCCs.

The remainder of this paper is organized as follows. Section II presents a detailed study into the correlation between acoustic features and MFCC vectors. Section III defines the acoustic feature vector and describes two methods of predicting acoustic features from MFCC vectors using either phoneme-specific or global MAP methods. Experimental results are presented in Sec. IV, which examines the accuracy of acoustic feature prediction from MFCC vectors under constraints of speaker independence, gender dependence, and gender independence. Finally, Sec. V examines the effect of replacing higher-order MFCCs by acoustic features on both acoustic feature estimation and speech recognition performance.

## II. ANALYSIS OF CORRELATIONS BETWEEN ACOUSTIC SPEECH FEATURES AND MFCC VECTORS

This section examines the correlation between acoustic speech features and MFCC vectors. This is motivated by the need to establish the existence and level of correlation before developing a system to predict acoustic speech features from MFCC vectors. The first part of this section explains how the acoustic feature to MFCC vector correlation is measured, while the latter parts analyze the correlation. Various constraints are made in the analysis, such as measuring correlation globally or phoneme-specifically and applying speaker and gender constraints.

### A. Measuring the acoustic speech feature to MFCC vector correlation

A correlation analysis is performed by first defining an acoustic speech feature vector, $\boldsymbol{f}_i$, which comprises the fundamental frequency, $f0$, and the frequencies of the first four formants, F1, F2, F3, and F4, at time frame $i$,

$$\boldsymbol{f}_i = [f0, \mathrm{F1}, \mathrm{F2}, \mathrm{F3}, \mathrm{F4}]. \tag{1}$$

For voiced speech, multiple correlations are measured between each element of the acoustic feature vector and the MFCC vector. For unvoiced speech, multiple correlations are only calculated between formant frequencies and MFCC vectors.

The multiple correlation between each acoustic speech feature and the MFCC vector is measured using multiple linear regression.[8] A linear model is computed to describe the relationship between MFCC vectors (independent variables) and acoustic speech feature vectors (dependent variables). Each acoustic feature at frame $i$, $f_i(j)$, is represented in terms of the $i$th MFCC vector, $\boldsymbol{x}_i = [x_i(1), x_i(2), \ldots, x_i(M)]$, using a set of $M+1$ regression coefficients, $[\beta_{0,j}, \ldots, \beta_{m,j}, \ldots, \beta_{M,j}]$, which are specific to the $j$th acoustic feature,[7]

$$f_i(j) = \beta_{0,j} + \beta_{1,j} x_i(1) + \beta_{2,j} x_i(2) + \cdots + \beta_{M,j} x_i(M) + \varepsilon,$$

$$0 \leqslant i \leqslant I-1, \quad 1 \leqslant j \leqslant 4 \quad \text{for unvoiced speech},$$

$$0 \leqslant i \leqslant I-1, \quad 0 \leqslant j \leqslant 4 \quad \text{for voiced speech}, \tag{2}$$

where $\varepsilon$ is an error term, $I$ is the number of MFCC vectors, and $M$ is the dimensionality of the MFCC vector. In matrix notation, Eq. (2) can be written as

$$\boldsymbol{F} = \boldsymbol{X_1}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{3}$$

$\boldsymbol{F}$ is a matrix of acoustic vectors, $\boldsymbol{F} = [\boldsymbol{f}_0, \boldsymbol{f}_1, \ldots, \boldsymbol{f}_{I-1}]^T$. $\boldsymbol{X_1} = [\mathbf{1}_I, \boldsymbol{X}]$, where matrix $\boldsymbol{X} = [\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{I-1}]^T$ and $\mathbf{1}_I$ is a vector of ones of length $I$. $\boldsymbol{\beta}$ is a matrix of regression coefficients and $\boldsymbol{\varepsilon}$ is a matrix of acoustic feature errors.

From a set of training data, least squares estimation can provide an estimate of the regression coefficients, $\hat{\boldsymbol{\beta}}$,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X_1}^T \boldsymbol{X_1})^{-1} \boldsymbol{X_1}^T \boldsymbol{F}. \tag{4}$$

This enables a prediction of an acoustic feature vector, $\hat{\boldsymbol{f}}_i$, from a MFCC vector, $\boldsymbol{x}_i$,

$$\hat{\boldsymbol{f}}_i = \boldsymbol{x}_i \hat{\boldsymbol{\beta}}. \tag{5}$$

The multiple correlation, $R$, between the $j$th acoustic feature, $f(j)$, and the MFCC vector is determined from the $R^2$ term, which is defined as

$$R(j)^2 = 1 - \frac{\Sigma_i \{f_i(j) - \hat{f}_i(j)\}^2}{\Sigma_i \{f_i(j) - \bar{f}(j)\}^2} = \frac{\Sigma_i \{\hat{f}_i(j) - \bar{f}_i(j)\}^2}{\Sigma_i \{f_i(j) - \bar{f}(j)\}^2}, \tag{6}$$

where $\bar{f}(j)$ is the mean of the $j$th acoustic feature.

### B. Correlation analysis

The correlations between acoustic features and MFCC vectors are measured on the test set of the VTRFormants database[9]—see Sec. IV for specific details. MFCC vectors are computed as specified in the ETSI DSR front-end,[3] resulting in a stream of 14-dimensional MFCC vectors at a rate of 100 vectors/s.

Two methods of calculating multiple correlations are considered: global, $R^G(j)$, and phoneme specific, $R^P(j)$. The global multiple correlation between each acoustic speech



Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 134.83.1.241 On: Tue, 19 Aug 2014 09:49:58

TABLE I. Multiple correlations between acoustic features and MFCC vectors for male and female speaker-dependent speech, calculated globally and by phoneme for unvoiced and voiced speech.

| Gender | Method | Voicing | F1 | F2 | F3 | F4 | *f*0 |
|---|---|---|---|---|---|---|---|
| Male | Global | Unvoiced | 0.5095 | 0.6688 | 0.5730 | 0.4655 | ⋯ |
| | | Voiced | 0.7520 | 0.8685 | 0.8128 | 0.5781 | 0.4874 |
| | Phoneme specific | Unvoiced | 0.7349 | 0.7994 | 0.7615 | 0.7375 | ⋯ |
| | | Voiced | 0.9017 | 0.9335 | 0.9238 | 0.8895 | 0.8728 |
| Female | Global | Unvoiced | 0.5172 | 0.5752 | 0.4869 | 0.3772 | ⋯ |
| | | Voiced | 0.7741 | 0.8369 | 0.7844 | 0.5366 | 0.7469 |
| | Phoneme specific | Unvoiced | 0.7495 | 0.7720 | 0.7545 | 0.7287 | ⋯ |
| | | Voiced | 0.9010 | 0.9140 | 0.8946 | 0.8615 | 0.9184 |

feature and MFCC vector is measured by pooling features together from all speech sounds and applying Eq. (6). Phoneme-specific correlations are measured by first segmenting the data into phoneme classes using reference annotations. Within each phoneme class, Eq. (6) is used to measure the multiple correlation between each acoustic feature and MFCC vector. A weighted averaged is then computed to give the mean phoneme-specific multiple correlation,

$$R^P(j) = \frac{1}{N_G} \sum_{w=1}^{W} N_w R^w(j), \tag{7}$$

where $N_w$ is the number of vectors corresponding to phoneme $w$ and $R^w(j)$ is the multiple correlation between the $j$th acoustic feature and the MFCC vector for the $w$th phoneme. $W$ represents the total number of phonemes, and $N_G$ is the total number of vectors.

Further in the two methods of measuring correlation, three scenarios are considered, which allows the effect of speaker and gender on correlations to be examined. The three scenarios are described in order of decreasing dependence on gender and speaker,

(1) Speaker dependent: Multiple correlations are calculated for each speaker separately. This scenario is also inherently gender dependent.
(2) Gender dependent: Multiple correlations are calculated first from all male speech and then from all female speech.
(3) Gender independent: All speech is used to compute the regression coefficients in Eq. (4). Multiple correlations are then measured separately for male and female speech to enable gender comparison.

Note that both the gender-dependent and gender-independent scenarios are also speaker independent as more than one speaker is used. Besides being listed by gender and method of calculating correlation (global or phoneme specific), correlation results are also calculated separately for voiced and unvoiced speech.

### 1. Speaker-dependent correlations

The global and phoneme-specific multiple correlations of formant frequencies (F1–F4) and fundamental frequency to MFCC vectors are shown in Table I. The correlation measures are broken down into male and female speech and unvoiced and voiced speech.

The results show that multiple correlations between acoustic features and MFCC vectors are consistently higher when calculated within individual phonemes rather than globally over all speech. This is to be expected since restricting multiple regression to model correlations from a small cluster of related sounds with similar formant structures and fundamental frequencies is more likely to produce higher correlation than generalizing across all speech sounds.

The high correlations between formant frequencies and MFCC vectors, especially for lower formants, in comparison to the lower correlation between fundamental frequency and MFCC vectors are attributed to the shape and spacing of the mel filter bank used in the MFCC extraction process. The filter bank processing retains sufficient spectral envelope information to indicate formant positions but lacks much of the finer spectral structure that conveys fundamental frequency information.[6] The nonuniform distribution of the mel filter banks means that while the lowest frequency channels provide a relatively high spectral resolution (the two lowest filter bank channels are 62 Hz apart), higher frequency channels have lower resolutions so F4 is less accurately represented (the two highest channels are 312 Hz apart). In addition, the frequency of F4 is often above the 4 kHz bandwidth of the speech, although the presence of such formants will still affect the spectrum below 4 kHz.

The level of the fundamental frequency to MFCC vector correlation was larger than expected as MFCC vectors have traditionally been considered to remove source information. However, the close spacing of lower frequency filter bank channels does allow some of the low frequency harmonic source information to be retained. In addition, the correlation between the fundamental frequency and F1, observed by Syrdal and Steele,[10] also contributes to the fundamental frequency to MFCC vector correlation. A significant increase in the fundamental frequency to MFCC vector correlation is observed when comparing the global to phoneme-specific measurements. This indicates that fundamental frequency is influenced by phoneme, an observation that has also been reported by Hirahara.[11]

The correlation between formant frequencies and MFCC vectors is always greater during voiced, rather than unvoiced, speech. This result is consistent with traditional formant estimation methods, which are more accurate for



Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 134.83.1.241 On: Tue, 19 Aug 2014 09:49:58

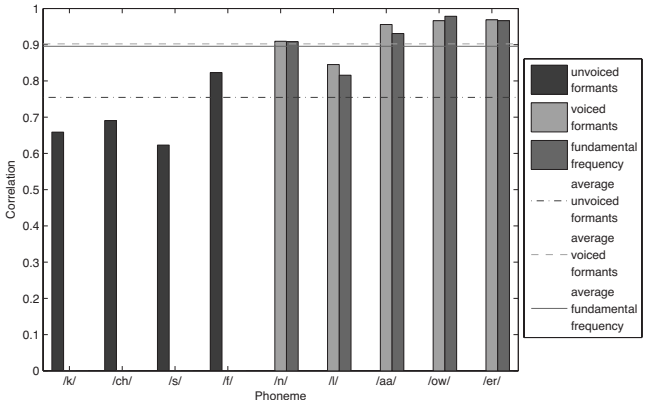


FIG. 1. Phoneme-specific correlations between acoustic features and MFCC vectors for nine example phonemes.

voiced speech. The spectral structure is better defined during voiced speech due to the high energy present at the harmonics of the fundamental frequency. For the noiselike structure of unvoiced speech, formants are less well defined; formant bandwidths are broader and amplitudes are usually lower.

Correlations between formant frequencies and MFCC vectors are generally higher for male speakers compared with female speakers. This is consistent with traditional signal processing methods of formant estimation, which also perform less well on female speech due to the wider spacing of fundamental frequency harmonics, which makes the precise localization of formant frequencies more difficult. For fundamental frequency, higher correlations are observed for female speech. This may be due to the higher frequencies associated with female speech, the harmonics of which span a wider range of mel filter bank channels than for male speech, making the identification of female fundamental frequency more accurate.

The acoustic feature to MFCC vector correlation was found to vary considerably across different phonemes. For example, Fig. 1 shows phoneme-specific correlations for nine voiced and unvoiced phonemes. For voiced phonemes, both fundamental frequency and mean formant frequency correlations (averaged across all four formants) are shown, while for unvoiced phonemes only the mean formant frequency correlation is shown. Also shown are the weighted mean phoneme-specific correlations, calculated over all phonemes, using Eq. (7), for fundamental frequency and voiced and unvoiced formant frequencies.

The bar chart shows that unvoiced phonemes have significantly lower formant frequency to MFCC vector correlation than voiced phonemes. This is confirmed by the lines showing mean formant frequency correlations, which for unvoiced phonemes is 0.76 in comparison to 0.91 for voiced phonemes. The higher correlation for voiced phonemes is attributed to a better spectral representation of formants that voiced excitation can provide through its harmonic structure and greater energy. Within the voiced phonemes, vowels exhibit higher levels of correlation than the semivowel /l/ and nasal /n/. For voiced phonemes, the level of the fundamental frequency to MFCC correlation follows closely that of the formant frequency correlation.

### 2. Gender-dependent correlations

The gender-dependent correlation of acoustic features to MFCC vectors is shown in Table II. These measurements were made by pooling data from all male speakers and all female speakers before applying multiple linear regression, thereby making the correlation analysis speaker independent as well. Comparing Tables I and II shows, without exception, that acoustic feature to MFCC vector correlations fall when moving from a speaker-dependent to a gender-dependent analysis.

The largest decreases in correlation are for phoneme-specific formant frequencies and fundamental frequency. For the gender-dependent and speaker-independent correlations shown in Table II, there is less difference in correlation when calculated globally or across individual phonemes. This is due to the increased variability as correlations are considered across all male speakers and across all female speakers, rather than for each speaker separately. There are instances where correlations are lower when calculated by phoneme rather than globally. The correlations between F1 and MFCC vectors for male speech provide such an example.

In Table II the largest increases in correlation when comparing global and phoneme-specific correlations occur for the fundamental frequency and F4. Compared with the cor-

TABLE II. Multiple correlations between acoustic features and MFCC vectors for male and female gender-dependent but speaker-independent speech, calculated globally and by phoneme for unvoiced and voiced speech.

| Gender | Method | Voicing | F1 | F2 | F3 | F4 | *f*0 |
|---|---|---|---|---|---|---|---|
| Male | Global | Unvoiced | 0.4578 | 0.6148 | 0.4954 | 0.3621 | ··· |
| | | Voiced | 0.7185 | 0.8137 | 0.7389 | 0.4515 | 0.3984 |
| | Phoneme specific | Unvoiced | 0.4543 | 0.6231 | 0.5472 | 0.4328 | ··· |
| | | Voiced | 0.7012 | 0.7959 | 0.7631 | 0.6001 | 0.5754 |
| Female | Global | Unvoiced | 0.4812 | 0.5326 | 0.4266 | 0.2332 | ··· |
| | | Voiced | 0.7554 | 0.8057 | 0.7384 | 0.3853 | 0.7267 |
| | Phoneme specific | Unvoiced | 0.5127 | 0.5792 | 0.4966 | 0.4134 | ··· |
| | | Voiced | 0.7645 | 0.7959 | 0.7455 | 0.5889 | 0.8122 |



Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 134.83.1.241 On: Tue, 19 Aug 2014 09:49:58

TABLE III. Multiple correlations between acoustic features and MFCCs for male and female gender-dependent and speaker-independent speech, calculated globally and by phoneme for unvoiced and voiced speech.

| Gender | Method | Voicing | F1 | F2 | F3 | F4 | *f*0 |
|---|---|---|---|---|---|---|---|
| Male | Global | Unvoiced | 0.4554 | 0.6037 | 0.4804 | 0.3240 | ··· |
| | | Voiced | 0.7055 | 0.7862 | 0.6999 | 0.3500 | 0.1815 |
| | Phoneme specific | Unvoiced | 0.4314 | 0.5887 | 0.4997 | 0.3839 | ··· |
| | | Voiced | 0.6661 | 0.7442 | 0.6966 | 0.4848 | 0.3532 |
| Female | Global | Unvoiced | 0.4724 | 0.4798 | 0.3894 | 0.0363 | ··· |
| | | Voiced | 0.7293 | 0.7663 | 0.6537 | 0.2031 | 0.6154 |
| | Phoneme specific | Unvoiced | 0.4195 | 0.4636 | 0.3417 | 0.1792 | ··· |
| | | Voiced | 0.7025 | 0.7235 | 0.6160 | 0.3707 | 0.6265 |

relations in Table I, gender has less effect on formant frequency correlations for the gender-dependent correlations in Table II.

### 3. Gender-independent correlations

Table III shows multiple correlations between acoustic speech features and MFCCs calculated using data that are gender independent and speaker independent. Data from all speakers were pooled to create a single set of gender-independent/speaker-independent regression coefficients. Phoneme-specific and global correlations for voiced and unvoiced speech were then calculated. The correlations are given for male and female speakers separately despite the regression coefficients used to estimate the acoustic speech features calculated from male and female speech pooled together. This enables comparisons with speaker-dependent and gender-dependent correlations shown in Tables I and II.

Comparing the gender-independent correlations in Table III to the gender-dependent correlations in Table II shows that correlations between acoustic speech features and MFCCs are always lower when calculated using gender-independent data because of the increased variation. The results also show higher formant correlation with voiced speech rather than unvoiced speech, and this is attributed to the better spectral representation of voiced speech.

The gender-independent correlations in Table III are, in general, slightly lower when calculated within phonemes rather than globally, expect for fundamental frequency and F4. For the speaker-dependent and gender-dependent correlations shown in Tables I and II, respectively, phoneme-specific correlations are usually greater than those calculated globally across all speech, as expected.

### 4. Summary

In this section multiple correlations between acoustic speech features and MFCC vectors have been measured in three scenarios to allow speaker-dependent, gender-dependent, and gender-independent analyses. The scenarios use the same data, but correlations are calculated over subsets of the data to vary constraints on speaker and gender. Figure 2 summarizes the observations by showing the multiple correlations between the fundamental frequency and MFCC vectors in Fig. 2(a) and the mean multiple correlations between formant frequencies and MFCC vectors in Fig. 2(b). The correlations are shown for the three scenarios introduced in Sec. II B but are averaged over male and female speech, and formant frequency correlations are averaged over all four formants.

For all conditions in Fig. 2, a higher acoustic feature to MFCC vector correlation is observed when measured within specific phonemes rather than globally across all speech. For formant frequencies, a higher multiple correlation with MFCC vectors occurs for voiced speech rather than unvoiced speech. The acoustic feature to MFCC vector multiple correlation peaks with speaker-dependent constraints and reduces with gender-dependent analysis, reaching its lowest values with gender-independent analysis due to the increased variability.

## III. PREDICTING ACOUSTIC SPEECH FEATURES FROM MFCC VECTORS

The correlation analysis in Sec. II suggests that sufficient information is retained during feature extraction to en-

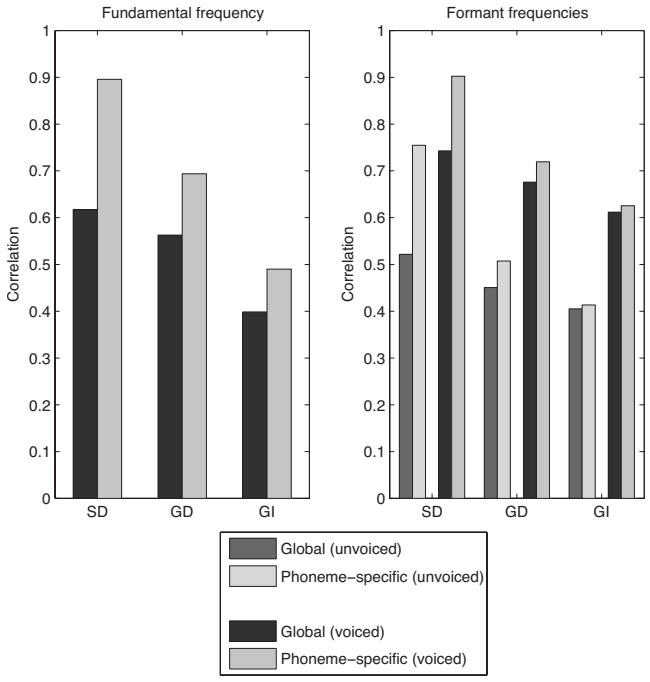


FIG. 2. Global and phoneme-specific correlations between MFCC vectors and (a) fundamental frequency and (b) formant frequencies for male and female speakers using speaker-dependent, gender-dependent, and gender-independent data.



Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 134.83.1.241 On: Tue, 19 Aug 2014 09:49:58

able fundamental and formant frequencies to be predicted from MFCC vectors. This section describes a MAP technique to utilize these correlations to predict acoustic features from MFCC vectors by employing statistical models of the joint density of acoustic features and MFCC vectors. Note that the term "prediction" is used to distinguish this statistical technique from signal processing methods of estimating acoustic speech features.

The analysis in the previous section showed that a higher acoustic feature to MFCC vector correlation is obtained when measured within specific phonemes rather than measured globally across all speech sounds. It is therefore expected that prediction will be more accurate using models of the joint density of acoustic features and MFCC vectors specific to individual phonemes rather than models that generalize all speech sounds together. To localize the modeling of the joint density of acoustic features and MFCC vectors to individual phonemes, a network of hidden Markov models (HMMs) is employed. Within each state of each HMM, the joint density of acoustic features and MFCC vectors is modeled by three Gaussian mixture models (GMMs), which represent voiced, unvoiced, and nonspeech audio. This "HMM-GMM" prediction of acoustic features from MFCC vectors comprises two parts: training of the HMM-GMMs and then prediction of acoustic features from a stream of MFCC vectors using the models.

### A. Phoneme-specific modeling of acoustic speech features and MFCC vectors

Phoneme-specific modeling of the joint density of acoustic features and MFCC vectors involves three stages of training. First, a set of phoneme HMMs is trained. Second, vector pools are created for each state of each phoneme model for voiced, unvoiced, and nonspeech vectors. Finally, from the vector pools, voiced, unvoiced, and nonspeech GMMs are trained to model the phoneme and state-specific joint density of acoustic features and MFCC vectors.

#### 1. HMM training

Training begins by creating of a set of $W+1$ MFCC-based HMMs, which model the $W$ phonemes in the database and also nonspeech. Left-right HMMs are used and comprise $S=3$ states with $H=8$ modes per state with diagonal covariance matrices. Phonemes are chosen as the speech class as they allow unconstrained speech to be processed when arranged in an unconstrained phoneme grammar.

#### 2. Phoneme-specific vector pools

Sets of vector pools for voiced speech, $\Omega^{v}_{s,w}$, unvoiced speech, $\Omega^{u}_{s,w}$, and nonspeech, $\Omega^{\mathrm{ns}}_{s,w}$, within each state $s$ of each phoneme model $w$ can now be created. This is achieved by first force aligning each training data utterance to the correct sequence of phoneme HMMs (using reference annotation labels) using VITERBI decoding.[12] This provides, for each training utterance, $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N]$ comprising $N$ MFCC vectors, phoneme model allocations, $\boldsymbol{m}=[m_1,m_2,\ldots,m_N]$, and state allocations, $\boldsymbol{q}=[q_1,q_2,\ldots,q_N]$, where $q_i$ and $m_i$ are the state and model associated with the $i$th MFCC vector, $\boldsymbol{x}_i$.

The MFCC vectors, $\boldsymbol{x}_i$, can be joined with their corresponding acoustic feature vectors, $\boldsymbol{f}_i$, to create joint feature vectors, $\boldsymbol{y}_i$, defined as

$$\boldsymbol{y}_i=[\boldsymbol{x}_i,\boldsymbol{f}_i]^T. \tag{8}$$

Voicing class information is contained within the joint feature vector through $\boldsymbol{f}_i$. In nonspeech the elements of $\boldsymbol{f}_i$ are all zero. For unvoiced speech F1–F4 contain nonzero values, but the fundamental frequency is zero. For voiced speech all elements of $\boldsymbol{f}_i$ are nonzero. Reference voicing classifications are provided by the ETSI extended front-end (XFE).[3]

Voiced augmented feature vectors, $\boldsymbol{y}_i$, associated with each state $s$ and phoneme model $w$ are pooled to form state and model dependent subsets of voiced feature vectors, $\Omega^{v}_{s,w}$, from the complete set of training data feature vectors, $Z$,

$$\Omega^{v}_{s,w}=\{\boldsymbol{y}_i\in Z:\ \mathrm{voicing}(\boldsymbol{y}_i)=\mathrm{voiced},\ q_i=s,\ m_i=w\},$$
$$1\leqslant s\leqslant S,\quad 1\leqslant w\leqslant W, \tag{9}$$

where voicing$(\boldsymbol{y}_i)$ is voiced, unvoiced, or nonspeech. Similar to Eq. (9), unvoiced, $\Omega^{u}_{s,w}$, and nonspeech, $\Omega^{\mathrm{ns}}_{s,w}$, vector pools are also created.

#### 3. Phoneme-specific GMMs

Expectation-maximization clustering[13] can be applied to each vector pool to create voiced, $\Phi^{v}_{s,w}$, unvoiced, $\Phi^{u}_{s,w}$, and nonspeech, $\Phi^{\mathrm{ns}}_{s,w}$, GMMs associated with each state $s$ of each phoneme model $w$. The set of state and model dependent GMMs that model the joint density of acoustic vectors and MFCC vectors for voiced speech is given by

$$\begin{aligned}p(\boldsymbol{y}|v,s,w)&=\Phi^{v}_{s,w}(\boldsymbol{y})\\&=\sum_{k=1}^{K}\alpha^{v}_{k,s,w}\phi^{v}_{k,s,w}(\boldsymbol{y})\\&=\sum_{k=1}^{K}\alpha^{v}_{k,s,w}\mathcal{N}(\boldsymbol{y};\boldsymbol{\mu}^{v,\boldsymbol{y}}_{k,s,w},\boldsymbol{\Sigma}^{v,\boldsymbol{yy}}_{k,s,w}),\end{aligned} \tag{10}$$

where $\phi^{v}_{k,s,w}(\boldsymbol{y})$ represents the $k$th Gaussian distribution in the voiced GMM for state $s$ and model $w$, which has associated with it the prior probability $\alpha^{v}_{k,s,w}$.

The set of voiced GMMs, $\Phi^{v}_{s,w}$, is represented by mean vectors, $\boldsymbol{\mu}^{v,\boldsymbol{y}}_{k,s,w}$, and covariance matrices, $\boldsymbol{\Sigma}^{v,\boldsymbol{yy}}_{k,s,w}$, for the $k$th cluster of state $s$ and model $w$ such that

$$\boldsymbol{\mu}^{v,\boldsymbol{y}}_{k,s,w}=\begin{bmatrix}\boldsymbol{\mu}^{v,\boldsymbol{x}}_{k,s,w}\\ \boldsymbol{\mu}^{v,\boldsymbol{f}}_{k,s,w}\end{bmatrix}\quad\text{and}\quad\boldsymbol{\Sigma}^{v,\boldsymbol{yy}}_{k,s,w}=\begin{bmatrix}\boldsymbol{\Sigma}^{v,\boldsymbol{xx}}_{k,s,w} & \boldsymbol{\Sigma}^{v,\boldsymbol{xf}}_{k,s,w}\\ \boldsymbol{\Sigma}^{v,\boldsymbol{fx}}_{k,s,w} & \boldsymbol{\Sigma}^{v,\boldsymbol{ff}}_{k,s,w}\end{bmatrix}. \tag{11}$$

Similar to Eq. (10), sets of unvoiced and nonspeech GMMs, $\Phi^{u}_{s,w}$ and $\Phi^{\mathrm{ns}}_{s,w}$, are also created. These are represented by mean vectors, $\boldsymbol{\mu}^{u,\boldsymbol{y}}_{k,s,w}$ and $\boldsymbol{\mu}^{\mathrm{ns},\boldsymbol{y}}_{k,s,w}$, covariance matrices, $\boldsymbol{\Sigma}^{u,\boldsymbol{yy}}_{k,s,w}$ and $\boldsymbol{\Sigma}^{\mathrm{ns},\boldsymbol{yy}}_{k,s,w}$, and prior probabilities, $\alpha^{u}_{k,s,w}$ and $\alpha^{\mathrm{ns}}_{k,s,w}$.

The voiced, unvoiced, and nonspeech vector pools also allow prior probabilities of the speech being voiced, $P(v|s,w)$, unvoiced, $P(u|s,w)$, and nonspeech, $P(\mathrm{ns}|s,w)$, to be computed for each state of each phoneme HMM. These are calculated from the number of voiced, unvoiced, and



Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 134.83.1.241 On: Tue, 19 Aug 2014 09:49:58

nonspeech vectors allocated to each state during training. For example, the prior probability of voiced speech in state $s$ of phoneme $w$, $P(v|s,w)$, is computed as

$$P(v|s,w) = \frac{N_{\Omega^v_{s,w}}}{N_{\Omega^v_{s,w}} + N_{\Omega^u_{s,w}} + N_{\Omega^{\mathrm{ns}}_{s,w}}}, \quad 1 \leqslant s \leqslant S, \quad 1 \leqslant w \leqslant W, \tag{12}$$

where $N_{\Omega^v_{s,w}}$ is the number of voiced vectors, $N_{\Omega^u_{s,w}}$ is the number of unvoiced vectors, and $N_{\Omega^{\mathrm{ns}}_{s,w}}$ is the number of nonspeech vectors in state $s$ of phoneme model $w$. Similar to Eq. (12), prior probabilities for unvoiced, $P(u|s,w)$, and nonspeech, $p(\mathrm{ns}|s,w)$, are computed.

### B. Phoneme-specific prediction of acoustic speech features from MFCCs

The prediction of acoustic features from a stream of MFCC vectors requires first an estimate of the phoneme and state sequences, which can be provided by VITERBI decoding using the network of HMMs discussed in Sec. III A 1. Next, the voicing class of each MFCC vector is predicted using the voiced, unvoiced, and nonspeech GMMs associated with the state that the MFCC vector is aligned to. For MFCC vectors predicted as voiced, formant and fundamental frequencies are predicted, while for unvoiced MFCC vectors only formant frequencies are predicted.

#### 1. Voicing prediction

VITERBI decoding of the input MFCC vector stream provides the state and phoneme sequences, $q_i$ and $m_i$, for each MFCC vector, $\boldsymbol{x}_i$. The prior voicing probabilities give an initial prediction of the voicing class of the MFCC vector based only on its phoneme and state allocation. However, VITERBI decoding errors can lead to an erroneous voicing class prediction, which in turn leads to incorrect decisions to predict or not to predict acoustic features. This is avoided by introducing posterior voicing probabilities, which utilize information contained within the voiced, unvoiced, and nonspeech GMMs.

The probability of the MFCC vector, $\boldsymbol{x}_i$, allocated to state $q_i$ and phoneme HMM $m_i$, belonging to the voiced GMM is given as

$$P(v|\boldsymbol{x}_i,q_i,m_i) = \frac{P(v|q_i,m_i)p(\boldsymbol{x}_i|v,q_i,m_i)}{p(\boldsymbol{x}_i|q_i,m_i)}, \tag{13}$$

where $P(v|q_i,m_i)$ is the prior probability of the MFCC vector being voiced based on its state and phoneme allocation, $p(\boldsymbol{x}_i|q_i,m_i)$ is the probability of vector $\boldsymbol{x}_i$, and $p(\boldsymbol{x}_i|v,q_i,m_i)$ is the probability of the MFCC vector being voiced according to the marginalized voiced GMM, $\Phi^{v,\boldsymbol{x}}_{q_i,m_i}$, where

$$\begin{aligned} P(\boldsymbol{x}_i|v,q_i,m_i) &= \Phi^{v,\boldsymbol{x}}_{q_i,m_i}(\boldsymbol{x}_i) \\ &= \sum_{k=1}^{K} \alpha^v_{k,q_i,m_i} p(\boldsymbol{x}_i|\phi^{v,\boldsymbol{x}}_{k,q_i,m_i}) \\ &= \sum_{k=1}^{K} \alpha^v_{k,q_i,m_i} \mathcal{N}(\boldsymbol{x}_i;\boldsymbol{\mu}^{v,\boldsymbol{x}}_{k,q_i,m_i},\boldsymbol{\Sigma}^{v,\boldsymbol{xx}}_{k,q_i,m_i}). \end{aligned} \tag{14}$$

The probabilities of the MFCC vector $\boldsymbol{x}_i$ belonging to the unvoiced GMM, $\Phi^{u,\boldsymbol{x}}_{q_i,m_i}$, and nonspeech GMM, $\Phi^{\mathrm{ns},\boldsymbol{x}}_{q_i,m_i}$, are similarly defined,

$$P(u|\boldsymbol{x}_i,q_i,m_i) = \frac{P(u|q_i,m_i)p(\boldsymbol{x}_i|u,q_i,m_i)}{p(\boldsymbol{x}_i|q_i,m_i)}, \tag{15}$$

$$P(\mathrm{ns}|\boldsymbol{x}_i,q_i,m_i) = \frac{P(\mathrm{ns}|q_i,m_i)p(\boldsymbol{x}_i|\mathrm{ns},q_i,m_i)}{p(\boldsymbol{x}_i|q_i,m_i)}. \tag{16}$$

The voicing of the MFCC vector is chosen by selecting the voicing class with the highest posterior voicing probability. Note that $p(\boldsymbol{x}_i|q_i,m_i)$ cancels from Eqs. (13), (15), and (16) and need not be calculated.

#### 2. Acoustic speech feature prediction

For MFCC vectors predicted as voiced, a MAP prediction of the $i$th acoustic speech feature vector, $\hat{\boldsymbol{f}}^k_i$, from the $k$th cluster of the voiced GMM, $\phi^{v,\boldsymbol{y}}_{k,q_i,m_i}$, can be made,

$$\hat{\boldsymbol{f}}^k_i = \arg\max_{\boldsymbol{f}_i}\{p(\boldsymbol{f}_i|\boldsymbol{x}_i,\phi^{v,\boldsymbol{y}}_{k,q_i,m_i})\}, \tag{17}$$

where $q_i$ and $m_i$ are the state and model that the MFCC vector $\boldsymbol{x}_i$ is allocated to. This evaluates to

$$\hat{\boldsymbol{f}}^k_i = \boldsymbol{\mu}^{v,\boldsymbol{f}}_{k,q_i,m_i} + \boldsymbol{\Sigma}^{v,\boldsymbol{fx}}_{k,q_i,m_i}(\boldsymbol{\Sigma}^{v,\boldsymbol{xx}}_{k,q_i,m_i})^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}^{v,\boldsymbol{x}}_{k,q_i,m_i}). \tag{18}$$

Predictions from all of the $K$ clusters can be combined by weighting using the voiced posterior probability, $h^v_{k,q_i,m_i}(\boldsymbol{x}_i)$, of the $i$th MFCC vector $\boldsymbol{x}_i$, belonging to the $k$th cluster,

$$\hat{\boldsymbol{f}}_i = \sum_{k=1}^{K} h^v_{k,q_i,m_i}(\boldsymbol{x}_i)\{\boldsymbol{\mu}^{v,\boldsymbol{f}}_{k,q_i,m_i} + \boldsymbol{\Sigma}^{v,\boldsymbol{fx}}_{k,q_i,m_i}(\boldsymbol{\Sigma}^{v,\boldsymbol{xx}}_{k,q_i,m_i})^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}^{v,\boldsymbol{x}}_{k,q_i,m_i})\}. \tag{19}$$

The voiced posterior probability, $h^v_{k,q_i,m_i}(\boldsymbol{x}_i)$, is given by

$$h^v_{k,q_i,m_i}(\boldsymbol{x}_i) = \frac{\alpha^v_{k,q_i,m_i} p(\boldsymbol{x}_i|\phi^{v,\boldsymbol{x}}_{k,q_i,m_i})}{\sum_{k=1}^{K} \alpha^v_{k,q_i,m_i} p(\boldsymbol{x}_i|\phi^{v,\boldsymbol{x}}_{k,q_i,m_i})}, \tag{20}$$

where $p(\boldsymbol{x}_i|\phi^{v,\boldsymbol{x}}_{k,q_i,m_i})$ is the marginal distribution of the MFCC vector for the $k$th cluster of the voiced GMM in state $q_i$ and model $m_i$.

For MFCC vectors predicted as unvoiced, acoustic features comprising only formant frequencies are predicted. Equations (19) and (20) are used, but voiced GMMs and voiced probabilities are replaced by unvoiced GMMs and probabilities.

### C. Global prediction of acoustic speech features

The phoneme-specific prediction can be reduced to a global prediction by replacing the network of HMMs with a single one-state HMM—i.e., setting $W=1$ and $S=1$. This removes the need to decode the MFCC vector stream into a state sequence. For each MFCC vector, a voicing classification is made from the voiced, unvoiced, and nonspeech GMMs in the single state, followed by a prediction of the appropriate acoustic features. This is a more simple method of prediction and provides a useful comparison to the more



Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 134.83.1.241 On: Tue, 19 Aug 2014 09:49:58

sophisticated phoneme-specific method. As such, the experimental evaluation of the acoustic feature prediction compares the performance of the phoneme-specific acoustic feature prediction with the global prediction.

## IV. RESULTS

The aim of the experiments in this section is to investigate the accuracy of the acoustic feature prediction methods under various constraints of speaker and gender dependence.

For gender-dependent and gender-independent testing, the VTRFormants database has been used.[9] This is a subset of the TIMIT database and comprises 324 utterances for training, spoken by 173 speakers, and 192 utterances for testing, spoken by a different set of 24 speakers. The database is supplied with the first four formant frequencies, the first three of which are hand corrected. Originally the database was sampled at 16 kHz, although for this work it has subsequently been downsampled to 8 kHz. Reference voicing classifications have been created using the ETSI XFE tool,[3] and reference fundamental frequency has been extracted using the YIN algorithm.[14]

For speaker-dependent testing preliminary experiments revealed that insufficient speaker-dependent data were available within the VTRFormants database to reliably train the joint densities needed for an acoustic feature prediction. Instead, two further databases (UEAChris and UEACath) were used to provide male and female speaker-dependent speech. These databases, in addition to the audio, also contain laryngograph recordings, which (after a minor hand correction) provide reference voicing and fundamental frequency. Reference formant frequency data were obtained using a combination of linear predictive coding (LPC) analysis and Kalman filtering.[12] Each utterance comprises phonetically rich sentences that have been downsampled to 8 kHz for this work. The male speaker, UEAChris, provides 601 training utterances and 246 testing utterances, while the female speaker, UEACath, provides 579 training utterances and 246 testing utterances.

### A. Voicing classification

This section examines voicing classification accuracy for systems trained on speaker-dependent, gender-dependent, and gender-independent speech. Figures 3 and 4 show graphical voicing class confusion matrices for both global and phoneme-specific predictions. The global GMMs comprised 32 clusters, while the state and phoneme specific GMMs comprised 16 clusters. A detailed analysis of the effect of the number of clusters in the GMMs is given in Sec. IV B and IV C.

Within the two confusion matrices, bar graphs show the percentage of MFCC vectors correctly and incorrectly classified for each class of speech (voiced, unvoiced, and nonspeech). Results are shown for speaker-dependent, gender-dependent, and gender-independent systems and separately for male and female speech.

Examining the leading diagonals of the two figures reveals the phoneme-specific prediction to be more accurate than the global prediction of speech class. Considering the

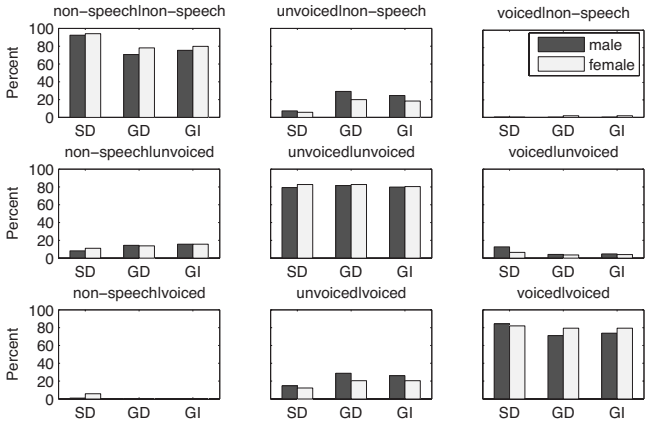


FIG. 3. Graphical confusion matrix for global (GMM) voicing classification for speaker and gender constrained speech.

off diagonal entries, which show speech class confusions, in general, the phoneme-specific prediction is more accurate with the exception of misclassifying slightly more nonspeech and unvoiced frames as voiced. The lowest errors occur for classifying voiced speech as nonspeech, which is important as such errors would erroneously cause no prediction of acoustic features to be made in some regions of voiced speech. Very low error rates are also observed in the phoneme-specific system for classifying unvoiced speech as nonspeech. Examining the effect of gender reveals female speech to be more accurately classified than male speech in most of the conditions tested.

### B. Fundamental frequency prediction

For evaluation purposes, the reference voicing is used to determine from which MFCC vectors fundamental frequency is predicted. This removes the effect of voicing classification errors from fundamental frequency evaluation by ensuring that in all tests the same vectors are selected. However, in practical situations the predicted voicing would be used. The percentage fundamental frequency error $E^{f0}$ is used to measure fundamental frequency prediction errors and is computed as

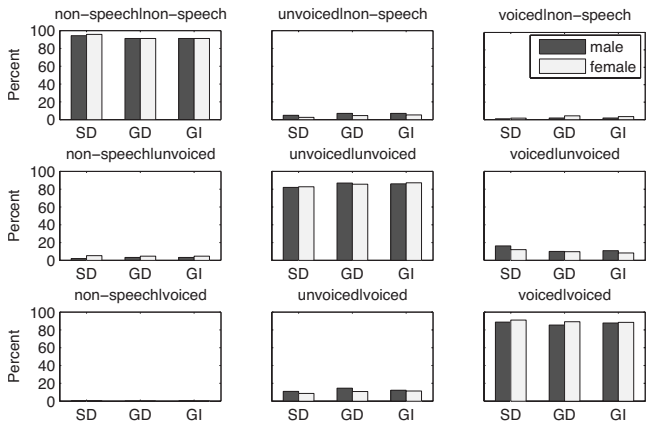


FIG. 4. Graphical confusion matrix for phoneme-specific (HMM-GMM) voicing classification for speaker and gender constrained speech.



Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 134.83.1.241 On: Tue, 19 Aug 2014 09:49:58

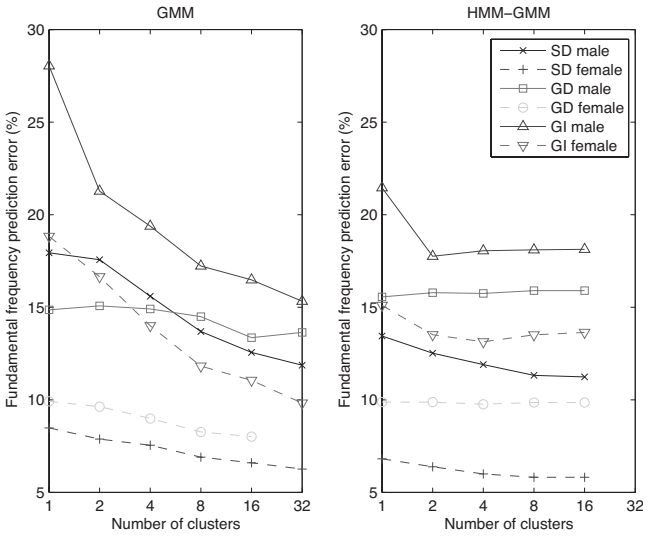


FIG. 5. Fundamental frequency prediction error with increasing number of clusters for (a) global (GMM) and (b) phoneme-specific (HMM-GMM) predictions using speaker-dependent, gender-dependent, and gender-independent speech.

$$E^{f0} = \frac{1}{N_v} \sum_{i=1}^{N_v} \frac{|\hat{f}_i(0) - f_i(0)|}{f_i(0)} \times 100\,\%\,, \tag{21}$$

where $N_v$ is the number of voiced MFCC vectors. Figure 5 shows percentage fundamental frequency prediction errors for global prediction and phoneme-specific prediction as the number of clusters in the GMMs is increased from 1 to 32. For both methods, speaker-dependent, gender-dependent, and gender-independent systems are evaluated separately for male and female speakers.

Considering first the effect of increasing the number of clusters in the GMMs, the results show that for a global prediction of fundamental frequency from a single GMM, errors reduce substantially as the number of clusters is increased. In contrast, for a phoneme-specific prediction from the network of HMM-GMMs, the effect of increasing the number of clusters gives much less reduction in error. This is attributed to the fact that the network of HMM-GMMs itself provides a relatively detailed model of the feature space, while the single GMM, which in itself has to model the entire feature space, requires substantially more clusters to achieve this. In addition, for each GMM in the HMM-GMM system, substantially less training data are available, which makes training of the state-specific GMMs more difficult. The results also reveal lower fundamental frequency errors with female speech than with male speech. This is consistent with the correlation analysis made in Sec. II B, which found a higher fundamental frequency to MFCC vector correlation for female speech than for male speech.

The results show that for speaker-dependent speech, the localization of prediction to specific phonemes through the HMM-GMM system gives lower errors than a global prediction from the single GMM system. However, for both gender-dependent and gender-independent speech, prediction errors are higher with the phoneme-specific system in comparison to the global system. This can be attributed to a lack of training data in both the gender-dependent and gender-independent systems. Examining the prediction errors for the HMM-GMM system in Fig. 5(b) reveals both the male and female speaker-dependent systems to decrease in error as the number of clusters increases. However, for both the gender-dependent and gender-independent systems, errors increase as more clusters are used, which suggests a lack of training data. The gender-dependent and gender-independent systems both use the VTR-Formants database, which comprises approximately 60 000 male speech frames and 42 000 female speech frames. In contrast, the speaker-dependent tests use the UEAChris and UEACath databases, which contain, respectively, about 233 000 and 261 000 speech frames. Further tests are presented in Sec. IV D, which use a larger gender-dependent database to explore the effect of insufficient training data.

### C. Formant frequency prediction

For the purposes of evaluation, formant frequencies are predicted from MFCC vectors labeled as speech according to the reference voicing. Separate GMMs are trained and tested for voiced and unvoiced speech as these were found to be more accurate than a single GMM due to the different structures of voiced and unvoiced speech. The formant frequency prediction error is averaged across all four formants and is measured using the mean formant frequency prediction error, $E^F$, which is defined as

$$E^F = \frac{1}{4N_v} \sum_{i=1}^{N_v} \sum_{j=1}^{4} \frac{|\hat{f}_i(j) - f_i(j)|}{f_i(j)} \times 100\,\%\,, \tag{22}$$

where $N_s$ is the number of speech MFCC vectors. Figure 6 shows global and phoneme-specific mean formant frequency prediction errors separately for voiced and unvoiced speech and for male and female speaker-dependent, gender-dependent, and gender-independent scenarios.

The results show that the formant frequency prediction from voiced speech is considerably more accurate than that from unvoiced speech. This observation is consistent with the findings of the correlation analysis in Sec. II B, which showed a higher formant frequency to MFCC vector correlation for voiced speech than for unvoiced speech; this is attributed to the better definition of formants in voiced speech than unvoiced speech. The trend of formant frequency prediction errors, when considering the number of clusters, follows a similar pattern to those found in the fundamental frequency prediction. Global prediction errors decrease as more clusters are used, but for the phoneme-specific HMM-GMM system, errors generally increase due to the lack of training data, with the exception of the larger speaker-dependent systems. This is again attributed to the larger amount of training data available for HMM-GMM training with the speaker-dependent database over that available for the gender-dependent and gender-independent systems.



Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 134.83.1.241 On: Tue, 19 Aug 2014 09:49:58

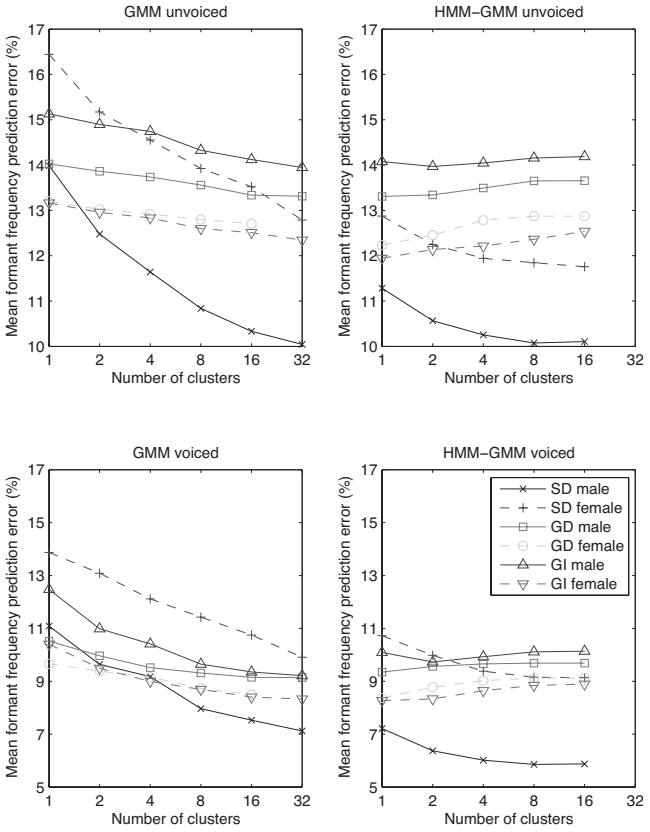


FIG. 6. Mean formant frequency prediction error with increasing number of clusters for (a) global (GMM) and (b) phoneme-specific (HMM-GMM) predictions using speaker-dependent, gender-dependent, and gender-independent speech. Top row: unvoiced; bottom row: voiced.

### D. Effect of using a larger gender-dependent speech database

The aim of this section is to investigate further the relatively poor performance of the phoneme-specific gender-dependent and gender-independent systems over the global prediction system. As stated in Secs. IV B and IV C, these tests use the relatively small VTRFormants database. In this section gender-dependent male speaker experiments are carried out using a subset of the larger WSJCAM0 database.[15] This contains approximately 777 000 training data frames in comparison to 60 000 male speech training data frames for the VTRFormants database. For the WSJCAM0 male subset, a reference fundamental frequency is provided by the YIN algorithm,[14] and formant frequencies are provided by a combined LPC/Kalman filtering method.[2]

Figure 7 shows male gender-dependent fundamental frequency errors for the global and phoneme-specific prediction systems as the number of clusters is varied from 1 to 32. As the number of clusters is increased, fundamental frequency errors decrease for the global prediction system but increase for the phoneme-specific system, although the level of change is much smaller than that observed for the VTRFormants database. However, in comparison to Fig. 5. The performances of the global and phoneme-specific systems are much closer. Using the VTRFormants database, the global prediction outperformed the phoneme-specific prediction by over 2%, while for the larger WSJCAM0 database the difference is about 0.2%.

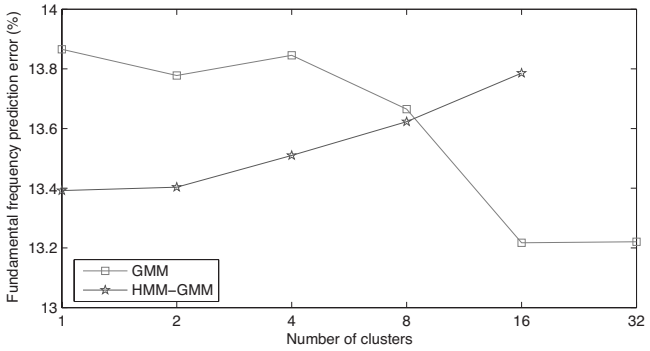


FIG. 7. Global (GMM) and phoneme-specific (HMM-GMM) gender-dependent fundamental frequency prediction errors.

Figure 8 shows male gender-dependent formant frequency errors for unvoiced and voiced speech for the global and phoneme-specific prediction systems. For both voiced and unvoiced speech, the phoneme-specific prediction outperforms the global prediction by about 1%. The results also reveal that phoneme-specific prediction errors decrease as more clusters are used, which was not observed with the smaller VTRFormants database.

These results suggest that to exploit the benefits of the more localized modeling that phoneme-specific prediction offers over global prediction, it is necessary to have sufficient training data to reliably create the state and phoneme specific GMMs.

## V. ENCAPSULATION OF ACOUSTIC SPEECH FEATURES

This final section presents an alternative method for obtaining acoustic speech features at the back-end and is included as a comparison to the proposed prediction method. In this method, the fundamental frequency and formant frequencies are computed at the front-end using conventional estimation methods applied to the time-domain signal.[2,3] The four highest-order MFCCs in the feature vector are then replaced by the acoustic features. This leads to two issues: first,

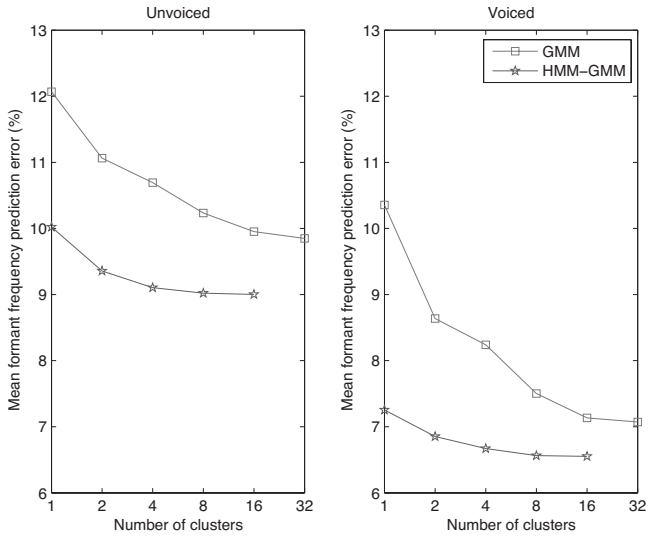


FIG. 8. Global (GMM) and phoneme-specific (HMM-GMM) gender-dependent (a) unvoiced and (b) voiced formant frequency prediction errors.



Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 134.83.1.241 On: Tue, 19 Aug 2014 09:49:58

TABLE IV. Unconstrained monophone accuracy using 12 D and 8-D MFCC vectors on speaker-dependent and speaker-independent speech.

| | MFCCs 1–12 | MFCCs 1–8 |
|---|---|---|
| Speaker dependent | 73.7% | 72.8% |
| Speaker independent | 53.1% | 48.4% |

to what extent speech recognition accuracy is reduced by the loss of higher-order MFCCs, and second, the error that estimation and quantization introduces into the acoustic features.

### A. Effect on recognition accuracy

The effect of reducing the feature vector from MFCCs 0–12 to MFCCs 0–8 is examined using the unconstrained phoneme recognition task described in Sec. III A for determining the phoneme sequence. Both speaker-dependent and speaker-independent tasks are examined, with results shown in Table IV. For the easier speaker-dependent task, recognition accuracy reduces by 1% as a result of removing MFCCs 9–12. However, on the more difficult speaker-independent task, accuracy falls by 5%.

### B. Quantization of acoustic features

Quantization of acoustic features uses the same method as the ETSI Aurora standard for MFCCs. This specifies that pairs of MFCCs are vector quantized, with MFCCs 9 and 10 being allocated 6 bits (to give 64 centroids) while MFCCs 11 and 12 are allocated 5 bits (to give 32 centroids). For vector quantization (VQ) of the acoustic features, the fundamental frequency and F1 are paired and allocated 6 bits while F2 and F3 are paired and allocated 5 bits. To examine quantization errors, VQ codebooks were trained on a set of 80 000 vectors and tested on a set of 40 000 vectors. Percentage quantization errors, using Eqs. (21) and (22), were 5.3%, 2.3%, 4.7%, and 2.8% for f0, F1, F2, and F3, respectively. In addition to quantization errors, acoustic features are subject to errors made by the front-end-based estimation methods. An investigation into the accuracy of acoustic feature estimation methods, using hand-corrected data, showed that the lowest fundamental frequency errors of about 2% were obtained using the YIN algorithm, and the lowest mean formant frequency errors of about 8% were obtained from a LPC-Kalman method.[16]

Combining these estimation errors with the VQ errors, for fundamental frequency, gives higher errors than predicted on the speaker-dependent female task, but lower errors on the gender-dependent and gender-independent tasks. For the formant frequency estimation, the combination of errors leads to higher errors than predicted across most of the speaker and gender constraints. The effect of quantizing the MFCC vectors using the VQ scheme proposed by ETSI was found to have an insignificant effect on acoustic feature prediction accuracy—for example, the prediction of fundamental frequency reduced by 0.1%. This result is consistent with speech recognition results that found little difference between using quantized or unquantized MFCC vectors.[17]

## VI. CONCLUSIONS

This work has shown that correlation exists between acoustic speech features and MFCC vectors and can be increased by placing constraints on the speech. The analysis also revealed that a higher correlation is obtained within individual phonemes rather than globally across all speech sounds. This led to the development of a MAP prediction of acoustic features from MFCC vectors using a combined HMM-GMM framework. Evaluations found that, given sufficient training data, phoneme-specific prediction of acoustic features is more accurate than global prediction. A problem with the HMM-GMM system is the need for substantially more training data to train GMMs within each state of each HMM. In cases where sufficient training data were unavailable, global prediction was found to be better.

A comparison to prediction was also made where higher-order MFCCs were replaced by acoustic features. Fundamental frequency accuracy was comparable to speaker-dependent prediction and was higher than gender-dependent and gender-independent predictions. For formant frequencies, the combination of estimation errors and VQ errors led to the technique generally performing worse than prediction. A significant downside of replacing higher-order MFCCs was found to be the reduction in speech recognition accuracy. Maximizing recognition accuracy is paramount in designing front-ends, and this decline in performance was judged unacceptable. If more accurate estimates are required, a better alternative is to adopt the XFE proposed by ETSI, which uses additional 800 bits/s to transmit fundamental frequency and voicing.[3]


[1]D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, edited by W. Kleijn and K. Paliwal (Elsevier, New York, 1995), Chap. 14.

[2]Q. Yan, S. Vaseghi, E. Zavarehei, B. Milner, J. Darch, P. White, and I. Andrianakis, "Formant tracking linear prediction model using HMMs and Kalman filters for noisy speech processing," Comput. Speech Lang. **21**, 543–561 (2007).

[3]ETSI, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; extended front-end feature extraction algorithm; compression algorithms; back-end speech reconstruction algorithm," ES 202 211, Version 1.1.1, ETSI STQ-Aurora DSR Working Group, 2003.

[4]X. Shao and B. Milner, "Predicting fundamental frequency from mel-frequency cepstral coefficients to enable speech reconstruction," J. Acoust. Soc. Am. **118**, 1134–1143 (2005).

[5]B. Milner and X. Shao, "Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction," IEEE Trans. Audio, Speech, Lang. Process. **15**, 24–33 (2007).

[6]J. Darch, B. Milner, and S. Vaseghi, "MAP prediction of formant frequencies and voicing class from MFCC vectors in noise," Speech Commun. **48**, 1556–1572 (2006).

[7]J. Darch, B. Milner, I. Almajai, and S. Vaseghi, "An investigation into the correlation and prediction of acoustic speech features from MFCC vectors," in ICASSP. Honolulu, HI (2007), Vol. **4**, pp. 465–468.

[8]S. Chatterjee and A. Hadi, *Regression Analysis by Example*, 4th ed. (Wiley, New York, 2006).

[9]L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in ICASSP, Toulouse, France (2006), Vol. **1**, pp. 369–372.

[10]A. Syrdal and S. Steele, "Vowel F1 as a function of speaker fundamental frequency," J. Acoust. Soc. Am. **78**, S56 (1985).

[11]T. Hirahara, "On the role of the fundamental frequency in vowel perception," J. Acoust. Soc. Am. **84**, S156 (1988).

[12]L. Rabiner, "A tutorial on hidden Markov models and selected applica-




Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 134.83.1.241 On: Tue, 19 Aug 2014 09:49:58

tions in speech recognition," Proc. IEEE **77**, 257–286 (1989).
[13]A. Webb, *Statistical Pattern Recognition*, 2nd ed. (Wiley, New York, 2002).
[14]A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am. **111**, 1917–1930 (2002).
[15]J. Fransen, D. Pye, T. Robinson, P. Woodland, and S. Young, "WSJCAM0 corpus and recording description," Technical Report No. CUED/F-INFENG/TR.192, Cambridge University Engineering Department, Cambridge, UK, 1994.
[16]J. Darch and B. Milner, "A comparison of estimated and MAP-predicted formants and fundamental frequencies with a speech reconstruction application," in Proceedings of Interspeech, Antwerp, Belgium (2007), pp. 542–545.
[17]H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in Automatic Speech Recognition: Challenges for the New Millenium, ISCA ITRW, Paris, France (2000), pp. 181–188.



Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 134.83.1.241 On: Tue, 19 Aug 2014 09:49:58