Available online at www.sciencedirect.com
ScienceDirect
Speech Communication xxx (2014) xxx–xxx
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Phonetic feature extraction for context-sensitive glottal source processing

John Kane [a,*], Matthew Aylett [b,c], Irena Yanushevskaya [a], Christer Gobl [a]

[a] Phonetics and Speech Laboratory, School of Linguistic, Speech and Communication Sciences, Trinity College Dublin, Ireland
[b] School of Informatics, University of Edinburgh, UK
[c] CereProc Ltd., UK

Received 19 September 2013; received in revised form 28 November 2013; accepted 24 December 2013

Abstract

The effectiveness of glottal source analysis is known to be dependent on the phonetic properties of its concomitant supraglottal features. Phonetic classes like nasals and fricatives are particularly problematic. Their acoustic characteristics, including zeros in the vocal tract spectrum and aperiodic noise, can have a negative effect on glottal inverse filtering, a necessary pre-requisite to glottal source analysis. In this paper, we first describe and evaluate a set of binary feature extractors, for phonetic classes with relevance for glottal source analysis. As voice quality classification is typically achieved using feature data derived by glottal source analysis, we then investigate the effect of removing data from certain detected phonetic regions on the classification accuracy. For the phonetic feature extraction, classification algorithms based on Artificial Neural Networks (ANNs), Gaussian Mixture Models (GMMs) and Support Vector Machines (SVMs) are compared. Experiments demonstrate that the discriminative classifiers (i.e. ANNs and SVMs) in general give better results compared with the generative learning algorithm (i.e. GMMs). This accuracy generally decreases according to the sparseness of the feature (e.g., accuracy is lower for nasals compared to syllabic regions). We find best classification of voice quality when just using glottal source parameter data derived within detected syllabic regions.

Keywords: Voice quality; Phonation type; Glottal source; Expressive speech; Speech synthesis

1. Introduction

Glottal source analysis refers to the process of trying to parameterise the important and salient aspects of the excitation source for voiced speech, created (mainly) by the vibration of the vocal folds at the larynx. Compared to many other feature extraction methods used in contemporary speech processing, glottal source analysis is relatively complex and involves making several simplifications of the speech production process (for a more comprehensive review of glottal source analysis please refer to: Alku (2011) or Walker and Murphy (2007)). For instance, glottal source analysis typically requires a process known as glottal inverse filtering as a pre-requisite. Glottal inverse filtering is the process of deconvolving a model of the vocal tract transfer function from the speech signal. The process involves making two key (and potentially over-reaching) assumptions.

The first is that speech production can be represented as a Linear Time-Invariant (LTI) system, which facilitates the linear separation of glottal source and vocal tract components (Fant, 1960). This representation is somewhat justified when using short analysis frames, as the articulators in the vocal tract are relatively slowly moving. However, as outlined in several previous publications (see e.g., Lin (1987), Fant and Lin (1987) and Fant et al. (1985b)) source-filter interactions effects exist. These interactions

* Corresponding author. Tel.: +353 1 896 1348.
E-mail addresses: kanejo@tcd.ie (J. Kane), matthewa@cereproc.com (M. Aylett), yanushei@tcd.ie (I. Yanushevskaya), cegobl@tcd.ie (C. Gobl).

0167-6393/$ - see front matter © 2013 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.specom.2013.12.003

are most significant in speech regions, for instance, where there is rapid transition of the vocal tract setting within a given analysis frame. The interactions may also be significant when there is a high f_0 and low first formant frequency, as commonly occurs in high vowels. Glottal inverse filtering of such analysis frames may result in an ineffective estimation of the glottal source component.

A second assumption is typically that the vocal tract can be modelled using an all-pole representation. This treatment is usually effective for oral sounds (due to the single-tube characteristic of the vocal tract), but for nasals (i.e. both nasal consonants and nasalised vowels) the different acoustic system is thought to create additional resonances and anti-resonances, and hence pole-zero pairs (Gobl and Mahshie, 2013). The presence of zeros in the vocal tract spectrum may also be true for laterals. As a result, glottal inverse filtering of such regions may be negatively affected by the lack of suitability of the vocal tract all-pole model. Furthermore, it has often been reported that signal processing methods for estimation of the all-pole vocal tract model can be sub-optimal for analysing higher-pitched voices (Alku et al., 2013; Alku and Vilkman, 1994).

One should note that despite these shortcomings for glottal source analysis and criticisms from the literature (notably from Teager and Teager (1990)) the use of glottal source feature data has brought significant benefits to a range of speech technology applications, including: speaker recognition (Chan et al., 2007; Zheng et al., 2007; Murty and Yegnanarayana, 2006), emotion classification (Cullen et al., 2013; Iliev et al., 2010; Lugger and Yang, 2008), characterisation of speaking styles in expressive speech data (Kane et al., 2013a; Székely et al., 2012; Campbell and Mokhtari, 2003), etc. Furthermore, one of the most natural sounding statistical parametric speech synthesisers currently available (Raitio et al., 2011) involves separate modelling of glottal source and vocal tract components, and also allows greater flexibility of voice characteristics compared to conventional methods (Raitio et al., 2013).

However, aside from parametric speech synthesis, which requires modelling of the glottal source for all voiced speech regions, for many other applications (such as those listed above) it may be preferable to use a lesser volume of glottal source feature data but which has been calculated from regions where is most likely to have been derived successfully. Such an approach of deriving glottal source feature data from selective speech regions has previously been suggested (Mokhtari and Campbell, 2003; Mokhtari and Campbell, 2002). Their method involves automatically detecting *centres of reliability*, which they define as vocoids involving high sonorant energy in steady regions where formant estimation is believed to be most reliable. Although they demonstrate the phonetic dependence of a certain glottal source parameter and that this parameter derived in these *centres of reliability* can be effective at discriminating certain affective labels, they do not formally assess the effect of using their selection method compared with not using it.

Recently, we proposed an alternative method for selecting optimal regions for glottal source analysis based on the presence or absence of certain phonetic features (Kane et al., 2013b). In that study we automatically determined the presence of a small number of phonetic features using Mel-Frequency Cepstral Coefficients (MFCCs) as input to Artificial Neural Networks (ANNs). That study revealed that by excluding glottal source feature data in detected nasal and fricative regions significant improvements could be achieved in voice quality classification. Despite these gains, there is still room-for-improvement, in particular in terms of accuracy of the phonetic feature extraction.

Different approaches have been used to automatically derive information on phonetic features from continuous speech. King and Taylor (2000) describe a method based on MFCCs used as inputs to recurrent neural networks and report accuracy in excess of 85% for many features (including vocalic, consonantal, nasal and strident features). However, as the results reported are the % of correct frames (and not, for instance, F-statistics), it is unclear exactly how well the classification performed for sparse features like nasals.

Previous to this, Ali et al. (1999) outlined a system which categorised speech into 4 components (sonorants, stops, fricatives and silences), before further subdividing these into 19 phonetic classes. Experiments on the TIMIT database demonstrated high accuracy, however as before % accuracy is not a very illuminating metric when analysing sparse features. More recently (Tarek and Carson-Berndsen, 2003; Kanokphara et al., 2006), a Hidden Markov Model (HMM) approach to phonetic feature extraction was developed and once more evaluated on the TIMIT database.

Several previous publications have described approaches involving the use of phonetic feature extraction as part of automatic speech recognition systems (Siniscalchi and Lee, 2009; Launay et al., 2002). More recently, authors have looked to exploit the discriminative power of deep neural networks in order to improve phonetic feature extraction accuracy (Siniscalchi et al., 2013; Yu et al., 2012). However, aside from our recent work (Kane et al., 2013b) to the best of our knowledge such approaches have not been investigated in terms of improving glottal source analysis.

1.1. Research questions and aims

The present paper looks to advance the work on phonetic feature extraction by: (1) carrying out a formal evaluation of detection of a range of phonetic features using three different classifiers and (2) by investigating the usefulness of such automatically derived information for glottal source analysis. The research questions can be stated explicitly as:

Please cite this article in press as: Kane, J. et al., Phonetic feature extraction for context-sensitive glottal source processing, Speech Comm. (2014), http://dx.doi.org/10.1016/j.specom.2013.12.003

RQ 1: How do different classifiers perform at detecting a set of binary phonetic features?
- *Hypothesis 1.1:* Accuracy will deteriorate with increasing sparseness (as expected following findings in Tarek and Carson-Berndsen (2003))
- *Hypothesis 1.2:* SVMs will deal relatively well with the sparseness problem.

RQ 2: Can the effectiveness of glottal source analysis be improved by restricting glottal parameter data to that occurring in certain phonetic contexts?
- *Hypothesis 2.1:* Avoiding nasal and voiced fricative regions will improve voice quality classification (following evidence from Kane et al. (2013b))

2. Phonetic feature extraction

2.1. Speech data

Two speech databases are used in the evaluation of the phonetic feature extraction, one for training and cross-validation and the other for optimising the classifier parameters. The speech data used here are summarised in Table 1.

For training and validation we use a large set of data recorded as part of the development of the CereVoice speech synthesis system. The database includes sub-corpora of speech produced using lax, neutral and tense phonation types in order to produce subtle changes in emotion (Aylett and Pidcock, 2007). The acoustic characteristics of identical phonemes produced in different phonation types can be markedly different (e.g., with differences in spectral tilt, presence of noise in the spectrum etc). As we intend to use the developed phonetic feature extraction on various types of speech data (in future work), including expressive and conversational speech, incorporating this variety in the training data is likely to increase the robustness of the feature extraction when applied to novel data. These sub-corpora have been recorded over a five year period across several languages (English, French, German, Italian, Japanese), however we include only the English data here. The data covers different accents of English (RP, General American, Scottish accent, Irish accent, Northern England, Midlands).

Table 1
Summary of speech data used in training and validating of the phonetic feature extraction.

Set	Speaker ID	Database	Gender	Utterances
Training & validation	ABM	CereVoice	Female	4724
	CJI	CereVoice	Male	7136
	FES	CereVoice	Female	5400
	FMM	CereVoice	Female	5580
	GTV	CereVoice	Female	4869
	JDH	CereVoice	Male	4982
	MAN	CereVoice	Male	4982
	NEN	CereVoice	Male	5785
	OAS	CereVoice	Female	4981
	PAH	CereVoice	Male	5017
	RRH	CereVoice	Female	4806
	SGT	CereVoice	Male	4363
	SMO	CereVoice	Female	6414
	SPA	CereVoice	Male	5829
	VDE	CereVoice	Female	6281
Development	AWB	ARCTIC	Male	1138
	BDL	ARCTIC	Male	1142
	CLB	ARCTIC	Female	1132
	SLT	ARCTIC	Female	1132

For optimisation of classifier parameters, we use data from 4 speakers (2 female, 2 male) from the ARCTIC database (Kominek and Black, 2004). We label this as the 'development' set. Note that the use of a completely separate database for parameter optimisation is done purposely to avoid biasing results on the training and validation database.

2.2. Classification

The approach used here is to develop classifiers of a set of binary phonetic features. The phonetic features used here are: {voiced, syllabic,[1] fricative, plosive, liquid, nasal}. Although this set is not as exhaustive as that proposed in Chomsky and Halle (1968), it does cover a reasonably large set of phonetic features which are relevant to the speech processing tasks considered in the present study. More specifically, for glottal source analysis it is clearly important to detect voiced sounds. The turbulent air present in fricatives and the potential zeros in the vocal tract spectrum for nasals and liquids, may negatively affect the glottal inverse filtering process. Similarly the rapid transitioning in plosives is likely to cause difficulty for glottal analysis.

The classification task here is to map from a set of acoustic features to binary labels, identifying the presence or absence of a given phonetic feature.[2]

More formally, the classification problem involves mapping from the feature space I, in $\mathbb{R}^n$, to the target space T (in this case {0, 1}, i.e. the individual phonetic feature binary target).

2.2.1. Acoustic features and target labelling

The standard Mel-frequency cepstral coefficients (MFCCs) are used as the acoustic features in the present

[1] We interpret the term *syllabic* following Chomsky and Halle (1968) whereby the feature is used to differentiate vowels from other classes of sounds. Note that consonants (such as liquids and nasals) that under certain circumstances may be [+syllabic] are not labelled as syllabic in this study.

[2] A note should be made here regarding the terminology. The classification problem addressed here, in fact, involves mapping from acoustic features to phonological labels. Indeed King and Taylor (2000) (and others) use the term *phonological feature extraction* which may appear more suitable. However, the use of binary phonological targets does not detract from the fact that phonetic variation within such phonological labels will inevitably affect the acoustic features and hence the classification output. Some authors have sought to circumvent this problem by using the term *articulatory feature extraction* (Tarek and Carson-Berndsen, 2003), but this may conjure up connotations of physiological measurements. As a result we opt for the term *phonetic feature extraction* despite its acknowledged shortcomings.

Please cite this article in press as: Kane, J. et al., Phonetic feature extraction for context-sensitive glottal source processing, Speech Comm. (2014), http://dx.doi.org/10.1016/j.specom.2013.12.003

study. The 13 MFCCs are measured on 25 ms Hanning windowed frames with a 10 ms shift. The 0th cepstral coefficient, corresponding to signal energy, is normalised to the maximum value for a given utterance. First (Δ) and second (Δ^2) derivatives are also included, resulting in a 39-dimensional feature vector, **x**.

The binary target label for each phonetic feature is set based on phonological labels derived following the forced alignment (described below) of the speech data. For example, for the phonetic feature 'fricatives', labels including /f/ and/z/ are assigned the target 1, with non-fricative labels assigned the target 0.

Forced alignment was carried out using the CereProc voice building system (Aylett and Pidcock, 2007). The alignment is a flat start monophone system which allows pronunciation variation. The underlying system used to carry out the alignment is HTK (Young et al., 2007) using a 10 ms frame rate, a five state model, and based on MFCCs of order 12 (plus log energy), and also first (Δ) and second (Δ^2) derivatives. The process is very similar to forced alignment described for Festival in Richmond et al. (2007), however, CereProc also employs proprietary techniques for refining pause insertion and dealing with multiple pronunciations. Tested against the CMU KED TIMIT database[3] with just over 21 min of speech, the CereVoice aligner did substantially better than the included festival based alignment (12% difference in insertion/deletion of segment boundary compared to 21% in the CMU KED TIMIT automatic labels, and a mean error of 10.1 ms for matching segment boundaries compared to 11.2 ms error in the Festival alignment). The speaker databases used in this study contained over 10 times the material in this evaluation corpus for each speaker and alignment results are likely to be improved over this baseline evaluation.

2.2.2. Artificial Neural Networks – ANNs

The first classification approach included in the present study is based on Artificial Neural Networks (ANNs). ANNs are in general used for learning the mapping function f from I to $T : f(\mathbf{x}) : \mathbf{x} \in I \rightarrow \mathbf{y} \in T$, where **x** denotes the input vector and **y** the output of the approximator f.

The ANN implementation we use here is based on the multi-layer perceptron (MLP) as the network type of choice which is said to fulfil the universal approximator theorem (Hornik, 1991). We use a two layer MLP, with a single hidden layer. The number of neurons used in the hidden layer is set below (Section 2.4). tanh is the transfer function used by the hidden layer, while the output layer uses a linear transfer function. Weight training is done using the back-propagation algorithm (Bishop, 2006).

2.2.3. Gaussian Mixture Models – GMMs

The second classification approach utilises Gaussian Mixture Models (GMMs). GMMs are a generative learning algorithm which involve modelling a given class of data using a mixture of multi-variate Gaussians. In our current implementation we train a GMM for the data where the given phonetic feature is *present* λ_1 and a separate GMM for where it is *absent* λ_0.

A given GMM, λ, has the probability density function:

$$p(\mathbf{x}|\boldsymbol{\lambda}) = \sum_{k=1}^{K} P_k \cdot \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\mathbf{k}, \Sigma_k) \quad (1)$$

where K is the number of multi-variate Gaussians, P_k is the prior probability of the kth Gaussian and each Gaussian can be written as:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma_k) = \frac{1}{(2\pi)^{-\frac{m}{2}}|\Sigma_k|^{-2}} e\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{k})^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)\right) \quad (2)$$

where **x** is the m dimensional feature vector (here m is 39, see Section 2.2.1), $\boldsymbol{\mu}_\mathbf{k}$ is its mean vector and Σ_k is its m-by-m covariance matrix. Here we use a diagonal covariance matrix, and K is optimised on the development set as described below (Section 2.4). The model parameters are trained using the Expectation–Maximisation (EM) algorithm (Bishop, 2006) with an initialisation step using K-means clustering. A given phonetic feature is considered to be present if:

$$p(\mathbf{x}|\lambda_1) > p(\mathbf{x}|\lambda_0) \quad (3)$$

that is, if a given feature vector, **x**, is more likely to have come from *present* phonetic feature GMM, λ_1, than the *absent* one, λ_0.

2.2.4. Support Vector Machines – SVMs

The final classifier included in the present study is an implementation of Support Vector Machines (SVMs). SVMs in general look to find a separating hyperplane which maximises the functional margin between the two classes. In our SVM implementation we utilise a Radial Basis Function (RBF) kernel (Bishop, 2006) which is used to project the feature data into a higher-dimensional space in order to derive a more effective separating hyperplane.

2.3. Experimental procedure

In order to validate the various classifiers used here for the purpose of extracting phonetic features, we carry out speaker independent leave-one-speaker-out validation experiments. Here classifiers are trained on all but one speaker's data, and are then tested on the held out data. The held out speaker is then rotated until all speakers have been covered. The procedure is repeated for each of the six phonetic features: {voiced, syllabic, fricative, plosive, liquid and nasal}. We use three metrics to evaluate the performance at the frame level. As percentage of errors (i.e. percentage of false positives and false negatives) is not a very suitable metric for assessing classification for sparse features (like nasals) we use the F1 score:

[3] http://festvox.org/dbs/dbs_kdt.html.

$$F1 = \frac{2 \cdot Tp}{2 \cdot Tp + Fp + Fn} \in [0, 1] \quad (4)$$

where Tp is the number of true positives, Fp is the number of false positives ad Fn is the number of false negatives. We also used False Positive Rate (FPR):

$$FPR = \frac{Fp}{Fp + Tn} \cdot 100 \quad (5)$$

and True Positive Rate:

$$TPR = \frac{Tp}{Tp + Fn} \cdot 100 \quad (6)$$

Note that during training the decision threshold, θ, in the ANN classifier is optimised by maximising F1 score on the training set.

2.4. Classifier optimisation

In order to use our classifiers in our experiments we must first optimise some of their parameters. These parameters are optimised on the development set summarised in Table 1.

First we look to optimise the number of neurons used in the hidden layer of the ANN classifier. This is done by carrying out a 10-fold cross validation procedure, where the F1 score is recorded for each fold. In Fig. 1 we illustrate the effect of increasing the number of neurons used in the hidden layer of the ANN by averaging across validation folds and phonetic features (ALL – black line), and we also show the effect separately for a selected sparse feature (Nasal – red line) and for a well represented feature (Syllabic – blue line). Overall there is no dramatic effect of increasing the number of neurons on the feature extraction averaged across all phonetic features. Similarly, for the syllabic feature, increasing the number of neurons does not have a significant positive effect and there is even some deterioration in F1 for higher numbers of neurons. For the nasal feature, however, there is a clear improvement with a higher number of neurons up to 64, after which the effect plateaus. Based on this we opt to use 64 neurons in our ANN implementation.

For the GMM classifier we carry out the same procedure, but this time varying the number of Gaussians (i.e. K) used in the GMM. The impact of this variation is illustrated in Fig. 2. One can observe that the F1 score for the syllabic feature plateaus from K set to 8. A similar effect is observed the sparser nasal feature, and indeed for all features combined, with no clear improvement observed for K greater than 16. As a result, 16 Gaussians are used our GMM implementation.

2.5. Results

Classification results for the phonetic features are shown below for F1 (Fig. 3), FPR (Fig. 4) and and TPR (Fig. 5). F1 score provides a good summary of detection performance and will, hence, receive the most attention, although FPR and TPR results will be referred to in order to help explain the F1 score. Note that the results here are used to determine which classifier should be used for each

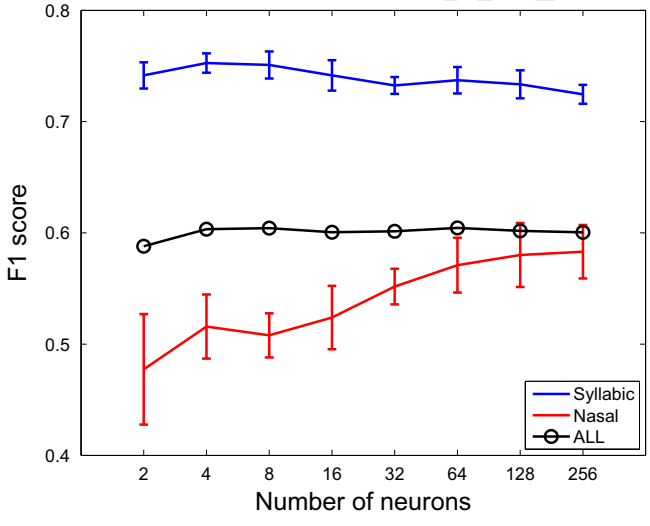

Fig. 1. Effect of varying number of neurons used in the ANN classifier on F1 score. Data is expressed as mean ± standard deviation.

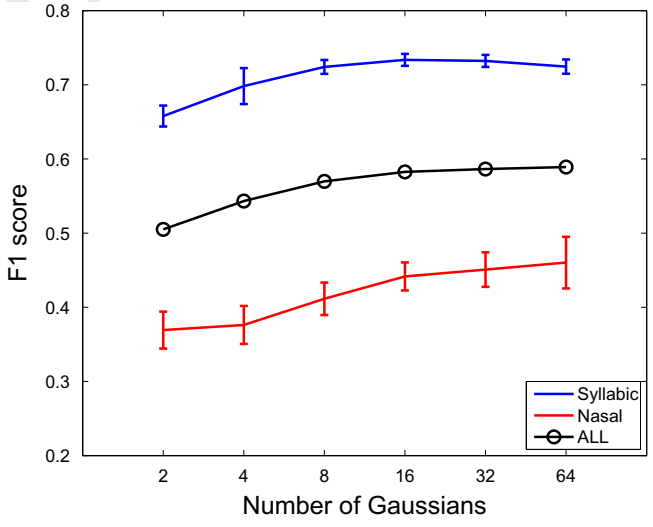

Fig. 2. Effect of varying number of Gaussians used in the GMM classifier on F1 score. Data is expressed as mean ± standard deviation.

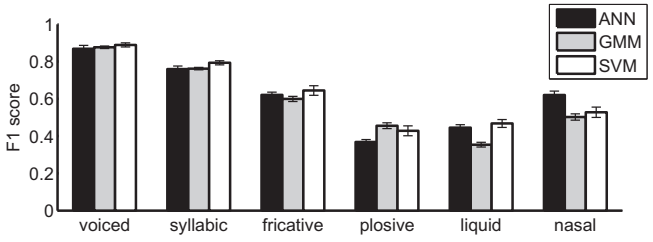

Fig. 3. **F1 score** plotted as a function of phonetic class (ranked in ascending order of sparseness) for the three classifiers. Data is expressed as mean ± standard error of the mean.

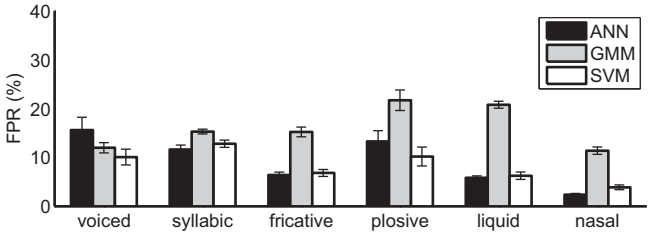

Fig. 4. **False positive rate (FPR)** plotted as a function of phonetic class (ranked in ascending order of sparseness) for the three classifiers. Data is expressed as mean ± standard error of the mean.

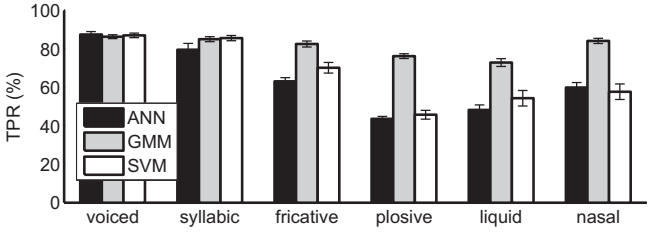

Fig. 5. **True positive rate (TPR)** plotted as a function of phonetic class (ranked in ascending order of sparseness) for the three classifiers. Data is expressed as mean ± standard error of the mean.

phonetic feature in the subsequent sections of this paper. If there are no significant differences, we default to the ANN classifier which is both computationally efficient at run-time and which also can be used to output a contour which can be interpreted as the posterior probability of the given feature.

A two-way ANOVA with F1 score treated as the dependent variable reveals a significant effect of both independent variables: phonetic features [$F_{(5,252)} = 343.77$, $p < 0.001$] and classifier type [$F_{(2,252)} = 5.74, p < 0.01$], as well as the interaction of the two independent variables [$F_{(10,252)} = 5.75, p < 0.001$. Pair-wise comparisons using Tukey's Honestly Significant Difference (HSD) test shows the SVM classifier produces significantly higher F1 scores compared to the GMM classifier ($p < 0.01$). Although the ANN classifier had a higher mean F1 score compared to the GMM method, the difference was not found to be significant ($p = 0.07$).

For the three phonetic features: voiced, syllabic and fricative, no significant differences are observed between the three classifiers, however SVM has a slightly higher mean F1 owing to a relatively lower false positive rate. For plosives, although the GMM classifier shows a higher false positive rate, its higher true positive rate results in a significantly higher F1 compared to the ANN classifier ($p < 0.05$), though no significant difference compared to the SVM. For liquid, the higher false positive rate for the GMM method causes a significantly lower F1 compared to both SVM and ANN classifiers ($p < 0.05$), though no difference is observed between SVM and ANN. Finally, for nasals the ANN classifier is found to have a significantly higher F1 compared to both the GMM ($p < 0.001$) and SVM ($p < 0.05$) methods.

Following the results observed in this section, we opt to use the ANN classifier for all phonetic features except for plosives, where the GMM classifier is instead used. Considering these particular phonetic feature extractors we briefly assess here the false positives observed in the speaker-independent experiments. Fig. 6 summarises the distribution of false positives for each of the phonetic feature extractors. Note that this figure shows just the 5 most common false positives. For voiced, /s/ and/t/ are the main false positives. It is not uncommon that these phonologically voiceless sounds would be subject to contextual voicing due to the presence of adjacent voiced segments, e.g., inter-vocalically. For the remainder of the phonetic features, false positives are more evenly distributed across the different sounds and in the majority of cases the substitution may be somewhat explained by the co-articulatory influence of surrounding segments. For fricative, for instance, devoicing of /l/, /r/ and /ə/, and aspirated or lenited realisation of /t/ may partly explain the identification of these segments as fricatives.

An example output of the entire phonetic feature extraction process is given in Fig. 7. Besides the binary output of the GMM feature extractor (used for plosives), one can observe a continuous output for the ANN extractor. Having continuous values for features like voiced and syllabic provide additional information than simply the binary decision, and may be useful for measuring aspects of the speech signal like degree of voicing.[4] Focusing on the output for nasals (third panel down) one can observe a clear peak for the only nasal consonant present (/n/) at around 0.8 s. The liquid /r/ is detected (fourth panel down) at around 0.45 and 1.25 s.

The output of the plosive GMM-based feature extraction (fifth panel down) reveals some interesting information to do with the proposed approach. One can observe that the /d/ (at around 0.6 s) and the /t/ (at around 1.15 s) are correctly identified. However, the first detected plosive region at around 0.5 s ('dh' which corresponds to/ð/) is counted as a false positive. In terms of the phonological label it is in fact a false positive, but careful phonetic analysis (using both auditory and spectrographic analysis) shows that the degree of constriction is likely higher than an idealised/ð/. This is of course a frequently occurring process in continuous speech where the voiced fricative /ð/ is realised as a plosive. The observation also seems to further justify the terminology used of *phonetic feature extraction* rather than *phonological feature extraction*. This may also somewhat explain the detected plosive at 0.95 s, however the spectrogram shows acoustic characteristics which look less like a plosive suggesting that this indeed may be a *true* false positive.

Further, it is interesting to observe in the fricative contour (sixth panel down), whereas the voiceless fricative

[4] Note that although it is hypothesised that the ANN output may be an indicator of the *degree* of a certain phonetic feature, such a correspondence is not formally assessed in the present work.

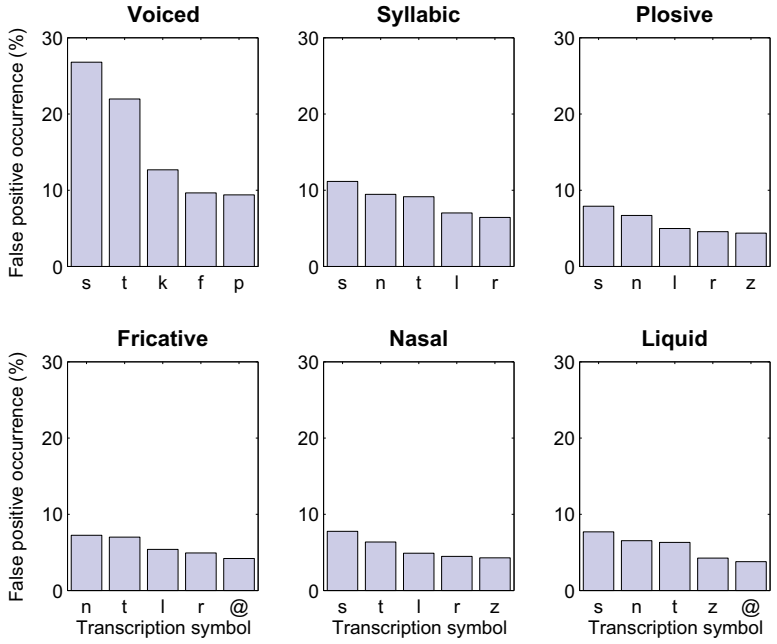

Fig. 6. Summary of the false positives across all phonetic feature extractors. Note that the @ corresponds to/ə/.

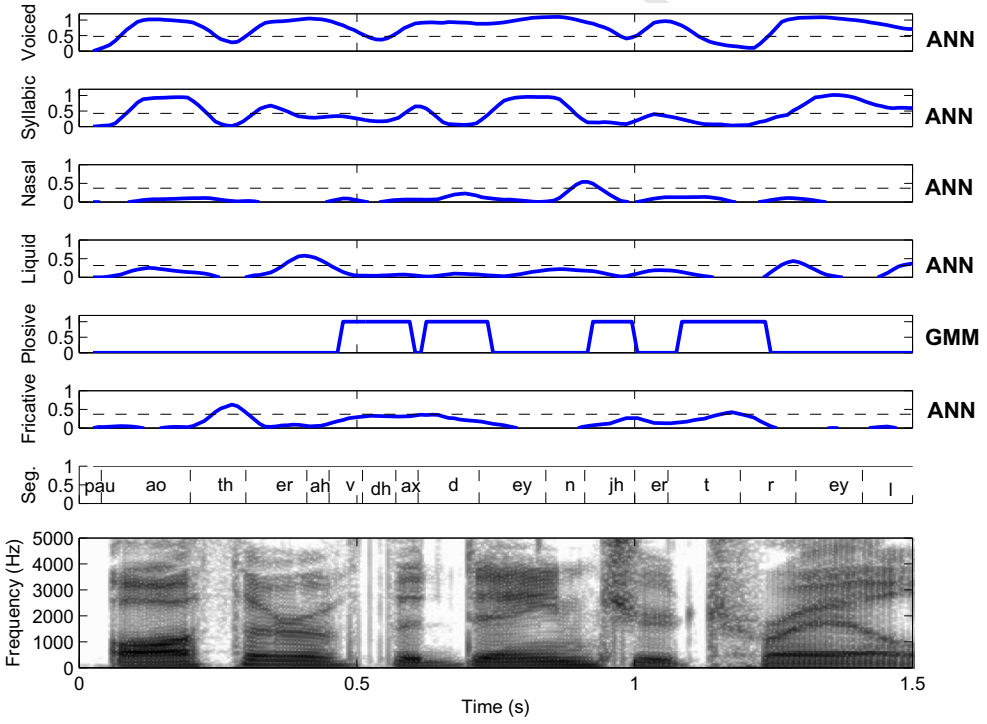

Fig. 7. Output of the phonetic feature extraction, along with the segmentation and broadband spectrogram for the utterance "Author of The Danger Trail ...". The decision threshold is given as a horizontal dashed line for the relevant phonetic features.

('th' which corresponds to/θ/) is clearly detected by the feature extractor, the first 'mistakenly' detected plosive /ð/ does not show a clear detection in the fricative contour. This supports our claim here of higher degree of constriction in the phonetic realisation of the sound, and also that the feature extraction approach does indeed correspond closely to the phonetics. For the /t/ at around 1.2 s, the fricative contour slightly exceeds the decision threshold which can be interpreted as a false positive. However, in the spectrogram one can observe the strong presence of noise and it is likely that the feature extractor is detecting the aspiration often accompanying voiceless stops as well as allophonic lenition which entails these sounds being produced as fricatives. This once more highlights that potential for strong

Please cite this article in press as: Kane, J. et al., Phonetic feature extraction for context-sensitive glottal source processing, Speech Comm. (2014), http://dx.doi.org/10.1016/j.specom.2013.12.003

variation in the phonetic realisation of certain phonological labels.

3. Context-sensitive glottal source processing

This section aims to utilise the information provided by the automatic phonetic feature extraction to improve the effectiveness of glottal source processing. The quantitative assessment of glottal source analysis is known to be problematic. Some authors use methods including: analysis of synthetic speech signals where parameter values are known (Drugman et al., 2011; Kane and Gobl, 2013b), analysis of natural speech with simultaneous Electroglottographic recordings (from which reference parameters can be derived, Kane and Gobl (2013a) and Sturmel et al. (2006)) or analysis-synthesis procedures. All of these methods have their own serious shortcomings. In this study we look to quantitatively evaluate the effectiveness of the glottal source analysis implicitly through voice quality classification experiments. The assumption here being that for speech data involving voice quality variation brought about by changes in laryngeal activity, effective glottal source analysis will inevitably lead to successful discrimination of voice quality. Contrastingly, ineffective glottal source parameterisation should result in a lack of discrimination of voice quality.

3.1. Speech data

In order to evaluate the glottal source analysis, we use a subset of the speech data originally used in Kane and Gobl (2013c). In this database 17 TIMIT sentences were spoken by 3 females and 3 males in a range of phonation types. In the present study we use only those sentences spoken in breathy, modal and tense phonation types. Additionally, we include speech data from 3 male speakers, saying 10 sentences again in breathy, modal and tense phonation types. This speech data was previously used in Kane and Gobl (2013d), and details of the recording conditions and setup are available in that publication.

3.2. Glottal source parameters

We use as feature data, both glottal source parameters derived as direct measures from estimated glottal pulses as well as parameters derived following the fitting of a mathematical model to the pulses.[5] For both sets of parameters there are some prerequisites. First, glottal closure instants (GCIs) are automatically detected from the speech data using the SE-VQ algorithm (Kane and Gobl, 2013c), which can be effective for analysis of non-modal phonation types. We then use the iterative and adaptive inverse filtering (IAIF) algorithm (Alku, 1992) in order to derive an estimate of the glottal source signal. The IAIF algorithm works by a sequence of all-pole modelling and inverse filtering of vocal tract and glottal source components, with increasing prediction order. Our IAIF implementation is carried out pitch-synchronously, on GCI-centred analysis frames with a duration of twice the local glottal period.

3.2.1. Direct measures

Four parameters measured directly from the estimated glottal source signal are included in the present study. Their inclusion is partly due to their effectiveness at discriminating voice quality on a lax-tense dimension, as demonstrated in Airas and Alku (2007). The first parameter is the normalised amplitude quotient (NAQ, Alku et al., 2002), which is derived using:

$$\mathrm{NAQ} = \frac{f_{ac}}{d_{peak} \cdot T_0} \tag{7}$$

where f_{ac} is the maximum amplitude of a given glottal flow pulse, d_{peak} is the maximum negative amplitude of the glottal derivative pulse (see Fig. 8) and and T_0 is the local glottal period. The quasi-open quotient (QOQ, Hacki, 1989) is derived by normalising the *quasi-open phase* (see top panel of Fig. 8)) to T_0. The quasi-open phase is defined as the duration between time points previous to and following the maximum amplitude of the glottal flow pulse that descend below 50% of this peak amplitude.

Two frequency domain parameters are also included. The first is the difference in amplitude between the first two harmonics of the narrowband glottal flow derivative spectrum (H1–H2, Hanson, 1997). The spectrum is derived using GCI-centred frames of duration three times the local glottal period (to ensure clear harmonics) from the estimated glottal flow derivative signal. Harmonic amplitudes are measured by searching for peaks in the vicinity of integer multiples of the local f_0 in the spectrum. The final parameter included is the so-called parabolic spectral

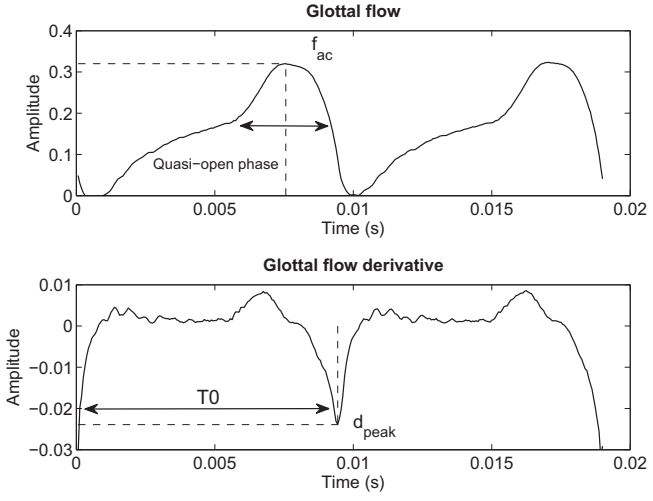

Fig. 8. Glottal flow (top panel) and glottal flow derivative (bottom panel) pulses estimated by IAIF. Highlighted are the measurements required for calculating NAQ (i.e. f_{ac} and d_{peak}) and QOQ (i.e. the quasi-open phase).

[5] Note that many of the algorithms used here are freely available on the COVAREP repository:https://github.com/covarep/covarep.

parameter (PSP, Alku et al., 1997). The parameter is derived by fitting a parabola to the low-frequency part of the spectrum of a single glottal flow pulse.

3.2.2. Model-based measures

We also include glottal source parameters derived from the Liljencrants–Fant (LF) glottal source model (Fant et al., 1985a) fitted to estimated glottal flow derivative pulses. We use the recently proposed dyProg-LF algorithm (Kane and Gobl, 2013a). The method utilises a dynamic programming algorithm, the weights for which are optimised using manually-obtained glottal source analysis. The target cost consists of a weighted time-domain and frequency domain error measurement to ensure comprehensive modelling of glottal pulses. A transition cost is incorporated to ensure sensibly smooth parameter trajectories. The transition cost is modulated by a spectral-stationarity measure to allow rapidly varying parameter values in certain speech regions (e.g., voice-offset, creaky or harsh voice). Three parameters derived from the LF model fit are used as part of the present feature data: Rg (normalised frequency of the glottal formant), Rk (a measure of glottal skew, and inverse of the commonly used speed quotient) and Ra (a measure of the glottal return phase).

3.3. Experimental procedure

The 7 glottal parameters (i.e. {NAQ, QOQ, H1-H2, PSP, Rg, Rk, Ra}) are extracted from the speech data at locations corresponding to GCIs. Only GCIs in voiced regions (as determined using the phonetic feature extraction method) are used. Along with the glottal parameters, we also extract and record the output of the optimal phonetic feature extractors at these locations. This makes up our feature data to be used.

For the classification, we utilise an SVM implementation with a one-against-one multi-class architecture. As with the phonetic feature extraction, we use a RBF kernel. The targets used in the classification experiments are the three voice quality labels: {breathy, modal, tense}. 10-fold cross-validation experiments are carried out where the data is randomly separated into 10 equal sized sets. Training is carried out on 9 of the sets with testing on the one held out set. The procedure is repeated by varying the held out set until all 10 sets have been covered. Classification error and confusion matrices are recorded.

In order to examine the effect of including only selected glottal feature data we repeat the cross-validation experiments for 6 different feature sets:

All: Including glottal feature data from all voiced regions (used as a baseline).
No-liquid: Baseline feature set, excluding data from detected liquid regions.
No-nasal: Baseline feature set, excluding data from detected nasal regions.
No-fricative: Baseline feature set, excluding data from detected fricative regions.
No-plosive: Baseline feature set, excluding data from detected plosive regions.
Only-syllabic: Only feature data from detected syllabic regions

3.4. Results

The results from the voice quality classification experiments are illustrated in Fig. 9, where classification error (%) from the 10-fold cross-validation is plotted as a function of feature set used. It is clear from Fig. 9 that choosing to include or exclude glottal source feature data from certain phonetic regions has a significant effect on the classification error. This observation is supported by results from a one-way ANOVA where feature set (i.e. the independent variable) is found to have a highly significant effect $[F_{(5,54)} = 64.0, p < 0.0001]$ on the classification error (i.e. dependent variable).

A further statistical analysis using Tukey's Honestly Significant Difference (HSD) test allows pairwise comparisons of the various feature sets. Excluding liquid regions actually increases the median classification error (to 38.5%) but with no significant difference compared to the baseline (i.e. feature data from all regions, which gives a median classification error of 35.6%). This suggests that glottal feature data derived in liquid regions is in fact beneficial, rather then harmful, to voice quality classification.

Excluding nasal regions brings a slight reduction in the median classification error (34.2%), but again with no sig-

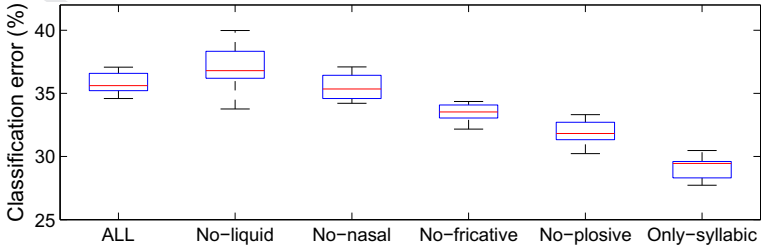

Fig. 9. Voice quality classification error (%) plotted as a function of feature sets used in the SVM classifier. Display order is All (baseline) first, then the rest in descending order of median classification error. The red centre-line indicates the median, the boxes show the inter-quartile range (IQR) and the whiskers are set as extensions of 1.5 times the IQR from the 25th and 75th percentiles.

Table 2
Confusion matrices for the voice quality classification experiment, shown for the classifier trained with all data (left column) and with only data in syllabic regions (right column).

	ALL			SYLLABIC		
	Breathy	Modal	Tense	Breathy	Modal	Tense
Breathy	59	30	11	**66**	26	8
Modal	23	61	16	22	**67**	11
Tense	7	21	72	6	16	**78**

nificant difference compared to the baseline. This finding does not strictly corroborate the initial findings reported in Kane et al. (2013b), where we found a significant improvement in classification when excluding nasal regions. However, despite the improvement being significant the amplitude of the difference was relatively small. Another important difference is that our detection of nasal regions is significantly more accurate in the present study compared to that in Kane et al. (2013b). It may be that the false positives resulting from the previous nasal detection method were in fact also useful to exclude from the features used in the classifier.

Excluding feature data from fricative regions brings a highly significant ($p < 0.0001$) reduction in classification error (median error of 33.6%), corroborating our previous findings (Kane et al., 2013b). Excluding plosive regions results in an even larger reduction in classification error (median error of 32.1 %) relative to the baseline ($p < 0.0001$) and also compared with results from removing fricative regions ($p = 0.05$). The largest reduction in classification error is achieved by only utilising glottal feature data obtained in detected syllabic regions (28.2% median error), with a 7.4% reduction in median classification error compared to using feature data from all speech regions. The reduction is further reported from the pairwise comparisons which reveal a highly significant difference ($p < 0.0001$) compared to every other feature set.

Finally, confusion matrices for the 'all' feature set and the 'syllabic' feature set are shown in Table 2. The matrices demonstrate an even classification improvement across the three voice quality labels. This suggests that the approach of isolating syllabic regions is helpful generally for improving the classification of voice quality and not solely for one voice quality class.

4. Discussion & conclusion

This study looked to implement and evaluate a variety of approaches for automatically determining information on the presence of an array of binary phonetic features. We then looked to apply this information to allow glottal source processing which is sensitive to the underlying phonetic context. In particular, we implicitly evaluated the effectiveness of glottal source analysis through a set of voice quality classification experiments.

In response to the first research question (RQ 1, at the end of Section 1) we implemented and evaluated classifiers based on ANNs, GMMs and SVMs on a vast speech dataset covering a range of speakers. The data consisted of speech produced in a variety of phonation types which is likely to enhance the robustness of the feature extraction when applied to expressive speech. In terms of hypothesis 1.1, we indeed do generally observe a decrease in accuracy with increasing sparseness of the given feature. However, sparseness in not the only issue affecting accuracy, as demonstrated by the higher accuracy for nasals compared to the less sparse plosives and liquids. This is likely due to the more stable spectral characteristics of nasals compared to plosives, which often display a relatively long period with very low signal energy.

We in general observe higher accuracy for the discriminative classifiers (i.e. ANNs and SVMs) compared to the generative classifier, GMM. This is with the exception of plosives, where the GMM-based classifier gives the best accuracy. It is rather difficult to speculate on why the GMM classifier is most effective for plosives. One explanation, however, could be that plosives, unlike the other classes of speech sounds included here, are highly varied, dynamic events with combinations of a hold phase and release burst. Using multiple Gaussians in a GMM may be useful for modelling these separate acoustic characteristics which both come under the single class 'plosive' and, hence, this approach may be most effective for modelling this specific speech feature.

For SVMs, which we initially hypothesised (hypothesis 1.2 to be effective with handling sparse features, we in general observe a similar level of performance to the ANN classifier (with nasals being an exception). Note that SVMs are not found, for any phonetic feature, to significantly outperform the ANNs. Although the SVMs provide significantly better detection of liquids than the GMMs, for the other sparse features the SVMs provide a similar or worse level of detection compared to the GMMs. Therefore, we cannot confirm hypothesis 1.2 and conclude that SVMs are particularly suited to the classification of sparse phonetic features.

We address RQ 2 by investigating the extent to which this information can be useful for improving the effectiveness of glottal source analysis. Evidence from the voice quality classification experiments strongly suggests that indeed the effectiveness of glottal source analysis can be significantly improved. In relation to hypothesis 2.1, nasals only appear to be slightly problematic for glottal analysis, as suggested by the minor reduction in classification by excluding feature data derived in nasal regions. This finding is somewhat at odds to the findings in Gobl and Mahshie (2013), however in that paper the authors found that nasalisation had the main negative impact on glottal return phase parameter estimation. As the voice quality classification experiments used in this study exploited a variety of parameters to do with both the glottal open and return phases the overall accuracy was not negatively affected for nasals. One must also bear in mind that voice quality classification can only provide a rather crude assessment

of glottal source analysis. Nevertheless, this approach is necessary for quantitative evaluation on a large body of data.

Removal of feature data from fricative regions, however, brings a significant improvement in classification error. The best classification accuracy is achieved by only using glottal feature data derived in detected syllabic regions. This finding supports the previous strategy applied by Mokhtari and Campbell (2003) for targeting specific speech regions for voice quality analysis. Recall, however, that those authors did not explicitly examine the improvement in voice quality classification using their selection approach compared to using all voiced speech regions. Our findings quantitatively demonstrate that syllabic regions are indeed the most reliable phonetic region for effective glottal source analysis and that the proposed phonetic feature extraction is a suitable and robust means for determining this information automatically. Also, as is discussed in the introduction, syllabic regions may be the parts of speech where we most portray our vocal timbre, so we must consider that this too may have affected the classification results favourably.

We intend to apply the proposed phonetic feature extraction approach, in particular the determination of syllabic regions, to our analysis of expressive and conversational speech. Furthermore, we wish to investigate whether the information provided by the phonetic feature extraction can be used to enable an adaptive vocal tract model to improve glottal inverse filtering, and indeed the parameterisation of speech in general. Finally, the approach of phonetic feature extraction may be exploited in clinical settings to help allow clinicians analyse read and spontaneous speech and alleviate some of the problems of analysing sustained vowels (e.g., the unnatural 'singing' production which has little in common with the habitual voice of a speaker).

Acknowledgements

The first, third and fourth authors are supported by the Science Foundation Ireland Grant 09/IN.1/I2631 (FASTNET).

References

Airas, M., Alku, P., 2007. Comparison of multiple voice source parameters in different phonation types. In: Proceedings of Interspeech 2007, Antwerp, Belgium, pp. 1410–1413.

Ali, A.M.A., der Spiegel, J.V., Mueller, P., Haentjens, G., Berman, J., 1999. An acoustic-phonetic feature-based system for automatic phoneme recognition in continuous speech. In: Proceedings of the IEEE International Symposium on Circuits and Systems 3, pp. 118–121.

Alku, P., 1992. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Commun. 11 (2-3), 109–118.

Alku, P., 2011. Glottal inverse filtering analysis of human voice production – a review of estimation and parameterization methods of the glottal excitation and their applications. Sadhana 36 (5), 623–650.

Alku, P., Vilkman, E., 1994. Estimation of the glottal pulseform based on discrete all-pole modeling. In: Proceedings of the Third International Conference on Spoken Language Processing, pp. 1619–1622.

Alku, P., Strik, H., Vilkman, E., 1997. Parabolic spectral parameter – a new method for quantification of the glottal flow. Speech Commun. 22 (1), 67–79.

Alku, P., Bäckström, T., Vilkman, E., 2002. Normalized amplitude quotient for parameterization of the glottal flow. J. Acoust. Soc. Am. 112 (2), 701–710.

Alku, P., Pohjalainen, J., Vainio, M., Laukkanen, A., Story, B., 2013. Formant frequency estimation of high-pitched vowels using weighted linear prediction. J. Acoust. Soc. Am. 134 (2), 1295–1313.

Aylett, M.P., Pidcock, C.J., 2007. The CereVoice characterful speech synthesiser SDK. In: Artificial Intelligence and Simulation of Behaviour (AISB). Newcastle, UK.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, New York.

Campbell, N., Mokhtari, P., 2003. Voice quality: the 4th prosodic dimension. In: Proceedings of the 15th International Congress of Phonetic Sciences, pp. 2417–2420.

Chan, W., Zheng, N., Lee, T., 2007. Discrimination power of vocal source and vocal tract related features for speaker segmentation. IEEE Trans. Audio Speech Lang. process. 15 (6), 1884–1892.

Chomsky, N., Halle, M., 1968. The Sound Pattern of English. MIT Press, Cambridge, MA.

Cullen, A., Kane, J., Drugman, T., Harte, N., 2013. Creaky voice and the classification of affect. In: Proceedings of WASSS, Grenoble, France.

Drugman, T., Bozkurt, B., Dutoit, T., 2011. A comparative study of glottal source estimation techniques. Comput. Speech Lang. 26, 20–34.

Fant, G., 1960. The Acoustic Theory of Speech Production, 2nd ed. Mouton, Hague, 1970.

Fant, G., Lin, Q., 1987. Glottal source – vocal tract acoustic interaction. KTH, Speech Transmission Laboratory, Quarterly Report 28 (1), pp. 13–27.

Fant, G., Liljencrants, J., Lin, Q., 1985. A four parameter model of glottal flow. KTH, Speech Transmission Laboratory, Quarterly Report 4, pp. 1–13.

Fant, G., Lin, Q., Gobl, C., 1985b. Notes on glottal flow interaction. KTH, Speech Transmission Laboratory, Quarterly Report 2-3, 21–45.

Gobl, C., Mahshie, J., 2013. Inverse filtering of nasalized vowels using synthesized speech. J.Voice 27 (2), 155–169.

Hacki, T., 1989. Klassifizierung von glottisdysfunktionen mit hilfe der elektroglottographie. Folia Phoniatrica, 43–48.

Hanson, H.M., 1997. Glottal characteristics of female speakers: acoustic correlates. J. Acoust. Soc. Am. 10 (1), 466–481.

Hornik, K., 1991. Approximation capabilities of multilayer feedforward networks. Neural Networks 4 (2), 251–257.

Iliev, I., Scordilis, M., Papa, J., Falco, A., 2010. Spoken emotion recognition through optimum-path forest classification using glottal features. Comput. Speech Lang. 24 (3), 445–460.

Kane, J., Gobl, C., 2013a. Automating manual user strategies for precise voice source analysis. Speech Commun. 55 (3), 397–414.

Kane, J., Gobl, C., 2013. Evaluation of automatic glottal source analysis. In: Proceedings of NOLISP, Mons, Belgium, pp. 1–8.

Kane, J., Gobl, C., 2013c. Evaluation of glottal closure instant detection in a range of voice qualities. Speech Commun. 55 (2), 295–314.

Kane, J., Gobl, C., 2013d. Wavelet maxima dispersion for breathy to tense voice discrimination. IEEE Trans. Audio Speech Lang. Process. 21 (6), 1170–1179.

Kane, J., Scherer, S., Aylett, M., Morency, L., Gobl, C., 2013. Speaker and language independent voice quality classification applied to unlabelled corpora of expressive speech. In: Proceedings of ICASSP, Vancouver, Canada.

Kane, J., Yanushevskaya, I., Dalton, J., Gobl, C., NíChasaide, A., 2013. Using phonetic feature extraction to determine optimal speech regions for maximising the effectiveness of glottal source analysis. In: Proceedings of Interspeech, Lyon, France.

Kanokphara, S., Macek, J., Carson-berndsen, J., 2006. Comparative study: HMM and SVM for automatic articulatory feature extraction.

In: Proceedings of the 19th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems.

King, S., Taylor, P., 2000. Detection of phonological features in continuous speech using neural networks. Comput. Speech Lang. 14, 333–353.

Kominek, J., Black, A., 2004. The CMU ARCTIC speech synthesis databases. ISCA speech synthesis workshop, Pittsburgh, PA, pp. 223–224. <http://festvox.org/cmuarctic/>

Launay, B., Siohan, O., Surendran, A., Lee, C., 2002. Towards knowledge-based features for hmm based large vocabulary automatic speech recognition. In: Proceedings of ICASSP, Orlando, Florida, USA, pp. 817–820.

Lin, Q., 1987. Nonlinear interaction in voice production. KTH, Speech Transmission Laboratory, Quarterly Report 28 (1), pp. 1–12.

Lugger, M., Yang, B., 2008. Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters. In: Proceedings of ICASSP, Las Vegas, Nevada, USA, pp. 4945–4948.

Mokhtari, P., Campbell, N., 2002. Automatic detection of acoustic centres of reliability for tagging paralinguistic information in expressive speech. In: Proceedings of Language Resources and Evaluation (LREC).

Mokhtari, P., Campbell, N., 2003. Automatic measurement of pressed/breathy phonation at acoustic centres of reliability in continuous speech. IEICE Transactions on Information and Systems (special issue on speech information processing) E-86-D(3), pp. 574–582.

Murty, K., Yegnanarayana, B., 2006. Combining evidence from residual phase and mfcc features for speaker recognition. IEEE Signal Processing Lett. 13 (1), 52–55.

Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., Alku, P., 2011. HMM-based speech synthesis utilizing glottal inverse filtering. IEEE Trans. Audio Speech Lang. process. 19 (1), 153–165 (1).

Raitio, T., Suni, A., Vainio, M., Alku, P., 2013. Synthesis and perception of breathy, normal, and lombard speech in the presence of noise. Computer Speech and Language (in press).

Richmond, K., Strom, V., Clark, R., Yamagishi, J., Fitt, S., 2007. Festival multisyn voices for the 2007 blizzard challenge. In: Proc. Blizzard Challenge Workshop (in Proc. SSW6), Bonn, Germany.

Siniscalchi, S., Lee, C., 2009. A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition. Speech Commun. 51 (11), 1139–1153.

Siniscalchi, S., Yu, D., Deng, L., Lee, C., 2013. Exploiting deep neural networks for detection-based speech recognition. Neurocomputing 106, 148–157.

Sturmel, N., d'Alessandro, C., Doval, B., 2006. A spectral method for estimation of the voice speed quotient and evaluation using electroglottography. In: 7th Conference on Advances in Quantitative Laryngology, Groningen, The Netherlands.

Székely, É., Kane, J., Scherer, S., Gobl, C., Carson-Berndsen, J., 2012. Detecting a targeted voice style in an audiobook using voice quality features. In: Proceedings of ICASSP, Kyoto, Japan, 4593–4596.

Tarek, A., Carson-Berndsen, J., 2003. HARTFEX: a multi-dimentional system of HMM based recognisers for articulatory features extraction. In: Proceedings of Non-Linear Speech Processing Workshop (NOLISP03).

Teager, H.M., Teager, S.M., 1990. Evidence for nonlinear sound production mechanisms in the vocal tract. In: Hardcastle, W.J., Marchal, A. (Eds.), Speech Production and Speech Modelling. Kluwer Academic, pp. 241–261.

Walker, J., Murphy, P., 2007. A review of glottal waveform analysis. In: Stylianou, Y., Faundez-Zanuy, M., Esposito, A. (Eds.), Progress in Nonlinear Speech Processing. Springer Verlag, pp. 1–21.

Young, Steve J., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 2007. The HTK Book Version 3.4. Cambridge University Press.

Yu, D., Siniscalchi, S., Deng, L., Lee, C., 2012. Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition. In: Proceedings of ICASSP, pp. 4169–4172.

Zheng, N., Lee, T., Ching, P., 2007. Integration of complementary acoustic features for speaker recognition. IEEE Signal Processing Lett. 14 (3), 181–184.

Please cite this article in press as: Kane, J. et al., Phonetic feature extraction for context-sensitive glottal source processing, Speech Comm. (2014), http://dx.doi.org/10.1016/j.specom.2013.12.003