EXPLORING DIFFERENCES BETWEEN PHONETIC CLASSES IN SLEEP APNOEA SYNDROME PATIENTS USING AUTOMATIC SPEECH PROCESSING TECHNIQUES

Jose Luis Blanco, Rubén Fernández,
Eduardo López and Luis Alfonso Hernández
Departamento de Señales, Sistemas y Radiocomunicaciones,
Universidad Politécnica de Madrid, Spain
e-mail: jlblanco@gaps.ssr.upm.es, ruben@gaps.ssr.upm.es,
eduardo@gaps.ssr.upm.es, luis@gaps.ssr.upm.es

Abstract

Early detection of obstructive sleep apnoea syndrome (OSA syndrome) using automatic speech processing techniques has become of great interest because the current diagnostic methods are expensive and time-consuming. Pioneering research in this field has recently yielded some promising results based on differences noted when comparing voices recorded from patients suffering from apnoea and those from healthy people. However, the relationship between this condition and the noted vocal abnormalities is still unclear, because the speech signals have not been systematically described. Most of the information used to describe the vocal effects of apnoea comes from a perceptual study where phoneticians were asked to compare voices from apnoea patients with a control group. These results revealed abnormalities in articulation, phonation and resonance.

This work is part of an on-going collaborative project between the medical and signal processing communities to promote new research efforts on automatic OSA diagnosis. In this paper, we explore the differences noted in phonetic classes (inter-phoneme) across groups (control/apnoea) and analyze their utility for OSA detection. Using statistical models, inter-phoneme scores were evaluated to quantify the predictive capability associated with each phonetic class for identifying this pathology. A global predictive power of 72% was obtained by combining inter-phoneme scores from four phonetic classes. We also compared these scores to identify the most discriminative phonetic classes. This process will help us improve our overall understanding of the effects of OSA on speech. Finally, using the Kullback-Leibler distance, significant differences were found for vowel production in nasal vs. non-nasal contexts. This was probably the result of the abnormal coupling of the oral and nasal cavities observed in apnoea patients. This finding represents a relevant result for future research.

1 Introduction

Obstructive sleep apnoea (OSA) is a highly prevalent disease (Fox & Monoson 1989), and it is estimated that in middle-age adults many as 9% of women and 24% of men are affected, undiagnosed and untreated (Lee et al. 2008). This disorder is characterized by recurring episodes of sleep-related collapse of the upper airway at the level of the pharynx and it is usually associated with loud snoring and increased daytime sleepiness. OSA is a serious threat to an individual's health if not treated. This condition is a risk factor for hypertension and, possibly, cardiovascular diseases (Coccagna et al. 2006). It is usually a factor in traffic accidents caused by somnolent drivers (Lee et al. 2008; Coccagna et al. 2006; Lloberes et al. 2000), and it can lead to a poor quality of life and impaired work performance. Current diagnostic procedures require a full overnight sleep study to confirm the presence of the disorder. This procedure involves recording neuroelectrophisiological and cardiorespiratory variables (ECG), which then results in a 90% accuracy rate in detecting OSA. Nevertheless, this is an expensive and time-consuming diagnostic protocol, and, in some countries such as Spain, patients have to remain on a waiting list for several years before the test can be completed. This is because the demand for consultations and diagnostic studies for OSA has significantly increased (Lee et al. 2008). There is, therefore, a strong need for methods of early diagnosis of apnoea patients in order to alleviate these considerable delays and inconveniences.

In over 25 years of research, a number of factors have been related to the upper airway (UA) collapse during sleep-time. Essentially, pharyngeal collapse occurs when the normal reduction in pharyngeal dilator muscle tone at the onset of sleep is superimposed on a narrowed and/or highly compliant pharynx. This suggests that OSA may be a heterogeneous disorder rather than a single disease, involving the interaction of anatomic and neural state-related factors resulting in pharyngeal collapse. However, it is interesting to consider that OSA is an anatomic illness that might have been favoured by evolutionary adaptations in the human's upper respiratory tract (Davidson 2003). Anatomic changes include shortening of the maxillary, ethmoid, palatal and mandibular bones; acute oral cavity-skull base angulation, pharyngeal collapse with anterior migration of the foramen magnum, posterior migration of the tongue into the pharynx with descent of the larynx, and shortening of the soft palate with loss of the epiglottic–soft palate lock-up.

Phoneticians have also taken a look into OSA from their own perspective (for instance Fox & Monoson 1989) and concluded that although articulatory, physiologic and acoustic anomalies are somewhat unclear, results involving combinations of factors have some explanatory power. Nevertheless, such an anomaly should result not only in respiratory, but also in speech dysfunction. Consequently, the occurrence of speech disorder in the OSA population should be expected, and it would likely involve anomalies in articulation, phonation and resonance. The most representative of these abnormalities are described in Section 2.

In this paper, we investigate the acoustic characteristics of speech in patients suffering from OSA by using techniques taken from automatic speech and speaker recognition. Using generative statistical models to describe the acoustic space, we explore the differences between phonetic classes and their possible application to automatic detection of OSA. These phonetic classes have offered a good trade-off between data complexity and recognition rate in speaker verification scenarios (Hébert & Heck 2003), especially when sparse data are available. The differences in the variability observed between and within a group of healthy speakers and those suffering from OSA are significant enough to motivate further research and reflect what phoneticians had observed in their previous experiments.

The remainder of this paper is organized as follows: in Section 2 the physiological and acoustic characteristics described in the previous literature on speakers suffering from severe apnoea syndrome are reviewed. On the basis of the limited information available about the side-effects of this condition on speech, a specific speech corpus was designed to test differences between both a normal and patient population. The design of this corpus, i.e., a brief analysis of the sentences it contains, is presented in Section 3; while Section 4 briefly describes the characteristics of the recorded speech database and provides several physical characteristics of the speakers in both groups. In Section 5, our approach, based on modelling the acoustic space using statistical models is presented. Once the experimental framework has been set, Section 6 describes the actual phonetic classes identified and provides details on their representation using statistical models. In Section 7, experimental results exploring differences between OSA and healthy speakers are presented using inter-phoneme and intra-phoneme scores. Finally, some conclusions and a brief outline on the future work are provided in Section 8.

2 Physiological and acoustic characteristics of OSA speakers

Currently, the articulatory/physiological settings as well as the acoustic characteristics of speech in speakers suffering from apnoea syndrome (for simplicity we will refer them as apnoea speakers), are still unclear. Most of the more valuable information in this field can be found in Fox and Monoson’s work (1989), where a perceptual study with skilled judges was presented comparing voices from apnoea patients and a control group (hereafter referred to as “healthy” speakers). This study revealed differences between both groups of speakers, however acoustic cues for these differences were somewhat contradictory and unclear. What did seem to be clear was that speakers in the apnoea group exhibited abnormal resonances that might appear due to the altered structure or function of the upper airway. Theoretically this anomaly should result not only in a respiratory but also in a speech dysfunction, which is our primary hypothesis. The abnormalities previously identified are the following:

Articulatory anomalies: Fox and Monoson (1989) pointed out that neuromotor dysfunctions could be found in a sleep apnoea population as a “lack of regulated innervations to the breathing musculature or upper airway muscle hypotonus”. This type of dysfunction is normally related to speech disorders, especially dysarthria. There are several types of dysarthria, each incorporating different acoustic features.

However, all types of dysarthria affect the articulation of consonants and vowels causing the slurring of speech. Another common pair of features in apnoea patients is hyper- and hypo-nasality, as well as a number of problems related to respiration.

Phonation anomalies: Phonation anomalies may appear due to the fact that heavy snoring in sleep apnoea patients can cause inflammation in the upper respiratory system and affect the vocal cords.

Resonance anomalies: The analysis of resonance characteristics for the sleep apnoea group in Fox and Monoson's work (1989) did not yield a clear conclusion. It was only recently that resonance disorders affecting speech quality have been associated with vocal tract damping features, distinct from airflow imbalance between the oral and nasal cavities. The term applied to this particular speech disorder is "cul-de-sac" resonance, and refers to a specific type of hyponasality. However, researches could only conclude that resonance abnormalities in apnoea patients could be perceived both as hyponasality (no nasalization is produced when the sound should be nasal) or hypernasality (nasalization is observed during production of non-nasal –voiced oral– sounds). Furthermore, and perhaps more importantly, speakers with apnoea seemed to exhibit smaller intra-speaker differences between non-nasal and nasal vowels due to this dysfunction, when vowels ordinarily require either a nasal or a non-nasal quality. Additionally, due to pharyngeal anomalies, differences in formant values can be expected. This was confirmed by Robb's work (Robb et al. 1997), in which vocal tract acoustic resonance was evaluated in a group of OSA males. Statistically significant differences were found in formant frequencies and bandwidth values between apnoea and healthy groups. In particular, the results of the formant frequency analysis showed that F1 and F2 values among the OSA group were generally lower than for the non-OSA group.

Finally, these anomalies can occur either in isolation or in combination. However, none of them was found to be sufficient on its own to allow accurate assessment of the OSA condition. In fact, all three descriptors were necessary to differentiate and predict whether the subject belonged either to the healthy or the OSA groups.

3 Speech corpus

The speech corpus was specifically designed to test differences between healthy people and those suffering from OSA. It contains four sentences in Spanish that are repeated three times by each speaker (Fernández et al. 2008). Keeping Fox and Monoson's work in mind, the sentences were designed so that they include instances of the following specific phonetic contexts:

- In relation to **articulatory anomalies** we collected voiced sounds affected by preceding phonemes that have their primary locus of articulation near the back of the oral cavity, specifically, velar phonemes, such as the Spanish velar approximant /g/. This anatomical region has been known to display physical anomalies in speakers suffering from apnoea (Davidson 2003). Thus, it is reasonable to suspect that different coarticulatory effects may occur with these phonemes in speakers with and without apnoea. In particular, in our corpus, we collected instances of transitions from the Spanish voiced velar plosive /g/ to vowels, in order to analyse the specific impact of articulatory dysfunctions in the pharyngeal region.

- With regard to **phonation anomalies**, we included continuous use of voiced sounds to measure possible irregular phonation patterns related to muscular fatigue noted in apnoea patients.
- Finally, to look at **resonance anomalies**, we designed sentences that allowed intra-speaker variation measurements; that is, measuring differential voice features for each speaker, for instance to compare the degree of vowel nasalization within and without nasal contexts.

Moreover, all sentences were designed to exhibit a similar melodic structure, and speakers were asked to try reading them with a specific rhythm under the supervision of an expert. We followed this controlled rhythmic recording procedure hoping to minimise non-relevant inter-speaker linguistic variability. The sentences chosen were the following, with the different melodic groups underlined separately:

(1) Francia, Suiza y Hungría ya hicieron causa común.
'fraNθja 'sujθa i uŋ 'gri a ya i 'θje roŋ 'kaw sa ko 'mun

(2) Julián no vio la manga roja que ellos buscan, en ningún almacén.
xu 'ljan no 'βjo la 'maŋ ga 'ro xa ke 'e λoz 'βus kan en niŋ 'gun al ma 'θen

(3) Juan no puso la taza rota que tanto le gusta en el aljibe.
xwan no 'pu so la 'ta θa 'řo ta ke 'taN to le 'γus ta en el al 'xi βe

(4) Miguel y Manu llamarán entre ocho y nueve y media.
mi 'γel i 'ma nu λa ma 'ran 'eN tre 'o tʃo i 'nwe βe i 'me ðja

The first phrase was taken from the Albayzin database, a standard phonetically balanced database for Spanish (Moreno et al. 1993). It was selected because it contains an interesting sequence of successive /a/ and /i/ vowel sounds.

The second and third phrases, both negative, have a similar grammatical and intonation structure. They are potentially useful for contrastive studies of vowels in different linguistic contexts. Some examples of these contrastive pairs arise from comparing a nasal context, “m**a**ng**a** roj**a**” (*'maŋ ga 'řo xa*), with a neutral context, “t**a**z**a** rot**a**” (*'ta θa 'řo ta*). These contrastive analyses could be very helpful to confirm whether the voices of speakers with apnoea had an altered overall nasal quality and displayed smaller intra-speaker differences between non-nasal and nasal vowels due to velopharyngeal dysfunction.

The fourth phrase has a single and relatively long melodic group, containing largely voiced sounds. The rationale for this fourth sentence was that apnoea speakers usually show fatigue in the upper airway muscles. Therefore, this sentence might help us discover anomalies during the generation of voiced sounds. This sentence also contains several vowel sounds embedded in nasal contexts that could be useful to study phonation and articulation of nasalized vowels. Finally, with regard to the resonance anomalies found in the literature and previously described, one of the possible traits of apnoea speakers is **dysarthria.** This last sentence could also be used to analyse dysarthric voices that typically show differences in vowel space when compared to healthy (control) speakers (Turner et al. 1995).

4 OSA database collection

The database, which in the rest of the paper will be referred to as OSA database, was recorded in the Respiratory Department at Hospital Clínico Universitario de Málaga, Spain. It contains the readings of 80 male subjects; half of them suffering from severe sleep apnoea (high Apnoea – Hipoapnoea Index values, AHI > 30), and the other half were either healthy subjects or had mild OSA (AHI < 10). Subjects in both groups had similar physical characteristics, such as age and Body Mass Index (BMI, i.e. weight divided by the square of height) - see Table 1.

Table 1. Distribution of healthy and pathological speakers in the OSA database

	Number	Mean Age	Std. dev. Age	Mean BMI	Std. dev. BMI
Control	40	42.2	8.8	26.2	3.9
Apnoea	40	49.5	10.8	32.8	5.4

Our selection of speakers for each group attempted to avoid the influence of the external predisposing factors associated with the condition. Such an approach ensures that the results are most likely related to group factors and can be generalized to a homogeneous population.

Moreover, speech was recorded using a sampling rate of 16 kHz in an acoustically isolated booth. The recording equipment consisted of a standard laptop computer with a conventional sound card equipped with a SP500 Plantronics headset microphone with A/D conversion and digital data exchange accomplished through a USB-port.

5 Statistical modelling of the acoustic space

The discrimination of normal and pathological voices using automatic acoustic analysis and speech recognition technology is becoming an alternative method of diagnosis for researchers in laryngological and speech pathologies, because of its nonintrusive nature and its potential for providing quantitative data relatively quickly. State-of-the-art speech recognition technology can be briefly described as the use of machine learning techniques to train a statistical model from acoustic features representing a known acoustic space (see [Huang et al. 2001] for a complete introduction to speech technology). These acoustic features are extracted from a training speech database where the speech from specific speakers is recorded and properly annotated. So, in **speaker recognition,** these acoustic features come from a known speaker’s voice, while in **speech recognition,** the acoustic space is generally covered by a set of phoneme-like units representing a given language. After training, the acoustic features coming from an unknown speaker or spoken sentence are recognized based on the likelihood scores obtained supposing that the unknown acoustic features were generated by a statistical model representing a particular speaker (speaker recognition) or linguistic unit (speech recognition). So, for example, in speaker recognition, a certain speaker is recognized when values from the acoustic features being tested are more likely (i.e. higher likelihood score) for the speaker’s own statistical model, rather than any other model in the system.

Given this brief overview of speech recognition and the expected speech abnormalities in patients with apnoea syndrome, it can be seen that the use of this technology to explore differences between apnoea and control speaker could be utilized in two complementary ways: 1) statistical models trained on control (or healthy) speech, when used to test acoustic features coming from apnoea speakers should provide lower likelihood scores (i.e., control models will be “less likely” to generate apnoea speech due to OSA-related anomalies) than when testing control speakers (regarded that a consistent cross-validation scheme is used); and 2) apnoea/control classification can be considered as a speaker recognition problem using only two different statistical models, one trained for the apnoea group and the other for the control population. In this research we will explore the first way, as the second one has been considered in our previous work (Fernández et al. 2009).

5.1 Acoustic features

The **front-end** in any speech recognition system is the process involved in extracting a set of acoustic features from the speech signal, so that it provides an efficient representation of speech without losing discriminative information. These acoustic features should also correspond to the assumptions made by the actual modelling techniques (generally statistical independence between features). Selecting a proper parameterization is therefore a relevant task, and one that depends significantly on the specific problem we are dealing with. According to Fox and Monoson’s (1989) perceptual experiments, some abnormalities can be directly identified by listening to the recordings. Therefore, conventional MFCC (Mel-Frequency Cepstral Coefficients) parameterization was applied in this research as it provides both, relative independent coefficients, and high discrimination between sounds based on its similarity with human perception processing (Huang et al. 2001). We acknowledge that an optimized representation, similar to that of Godino et al. (2006) for laryngeal pathology detection, could produce better results in terms of classification efficiency, but for the present work, we are not focusing on maximizing the accuracy rate, but in exploring differences within the acoustic space according to the same **principia** described in the preceding perceptual experiments.

5.2 Speech segmentation

To train different statistical models for different acoustic or linguistic units, the acoustic feature vectors resulting from the front-end pre-processing must be segmented or grouped into different training sets. Since we are interested in studying specific phonetic classes, all of the utterances in our OSA database had to be segmented into phonetic units. This phonetic segmentation allowed us to group acoustic feature vectors with specific phonetic classes, and then to train a specific statistical model for each phonetic class.

All sentences in our apnoea database (both for control and apnoea speakers) were automatically segmented into phonemes through forced recognition. That is, each sentence was forced to be recognized using the sequence of phonemes corresponding to its known transcription (optional silences between words were allowed). This forced alignment provided the start and ending time boundaries for each sound in the sentence. Automatic forced alignment avoids the need for time-consuming and costly manual annotation, but, as will be discussed in Section 6, it must guarantee an appropriate level of segmentation precision. In our case automatic phonetic segmentation was carried out with the open-source HTK tool (Young 2002). We use 24 left-to-right, 3-state, context-independent **Hidden Markov Models** (HMMs) to represent the basic set of 24 Spanish phonemes. These context-independent HMM phoneme models were trained from an available manually segmented, phonetically-balanced speech subcorpus of Albayzin, a reference large speech database for Spanish (Moreno et al. 1993).

5.3 Statistical modelling

After phonetic segmentation, due to the fact that Mel-Frequency Cepstral Coefficients may follow any statistical distribution on different phonetic classes, the **Gaussian Mixture Model** (GMM) approach, broadly applied in speaker recognition systems (Reynolds et al. 2000), was chosen to approximate the actual statistical distribution of the selected acoustic space. In our experimental setup we started by training GMM models for different phonetic classes using a large speech database: the Albayzin database (Moreno et al. 1993). By doing so we provide a set of stable initial models from which, using adaptation techniques, more specific GMMs were derived (tuned to particular characteristics of the speakers' population, recording conditions, etc.). A MAP (*Maximum A Posteriori*) adaptation algorithm, also commonly used in speaker verification (Reynolds et al. 2000), was applied to derive those specific GMMs representing our OSA database peculiarities: limited in the amount of speech and more specific in their phonetic and population coverage. Additionally, MAP adaptation is known to increase the robustness of the models, especially when sparse speech material is available. Besides, as it is also a common practice in speaker verification systems, only the means of the gaussian components in the GMMs were adapted. For our experiments, MAP adaptation to GMM models was estimated with the BECARS open source tool (Blouet et al. 2004).

6 Modelling phonetic classes for OSA analysis

The basic unit to convey linguistic meaning is the phoneme. Each phoneme can be considered to be a code that consists of a unique set of articulatory gestures, which includes the type and location of sound excitation, as well as the position of the vocal tract articulators. Additionally, other factors, such as the resonances produced within the vocal tract and the response of the vocal folds decisively affect the way in which those phonemes are pronounced. However, in this work we are not interested in the meaning, but in exploring the acoustic information embedded within speech signals. Consequently, we are not subjected to the traditional approach followed in speech recognition and may choose any other unit.

While specific instances of the individual phonemes are quite limited within short segments of direct speech, phonetic classes are easier to recognise than phonemes and occur much more frequently. Therefore models are easier to train with sparse data, as long as their internal complexity can be accommodated. Consequently a limited number of models can be trained when only sparse data are available. On the other hand, according to the previous literature dealing with the effects of OSA in speech signals, only a few phonetic classes seem to be relevant for our experiments. Bearing this in mind, four different groups of broad phonetic classes were defined:

Vowel sounds, VOW: vowel sounds represent one of the most relevant acoustic groups in speech processing applications, and have been intensively analyzed in the detection of pathological voices. Sustained vowel sounds typically are considered to be the best source of information. However, recent studies have pointed out that, at least for certain pathologies, vowel segments extracted from continuous speech might be as informative as those from sustained speech.

Nasal sounds, NAS: nasal sounds are especially relevant when considering resonance effects in speech signals involving both the oral and nasal cavities. The coupling and de-coupling of the nasal cavity, by means of the opening/closing of the velopharyngeal port, causes the most familiar resonance effect in speech. Nasal phonemes appear in conjunction with at least one vowel, and cause a singular unique transition from the vowel to the nasal (and vice versa) known as **nasalization.** This seems to be a particularly relevant situation (Davidson 2003), which we will be looking thoroughly at this paper.

Plosive sounds, PLO: in contrast to the two previous classes, plosive sounds represent non-stationary, fast transitions in the speech signal. Therefore, instead of cepstral coefficients, more specific acoustic measures (mainly voice-onset-time) are generally used for their study. Consequently, in our statistical models, built on cepstral coefficients information, plosive sounds could present lower variability rates. This is in contrast to vowel and nasal sounds, which are expected to exhibit variability when healthy and apnoea speakers are compared. However, due to co-articulation, and the flawed boundaries provided by our automatic segmentation process, the GMM model for plosive sounds could include acoustic information from transitions from adjacent phonetic classes. This could cause some differences in this class, when used as phonetic classifiers and thus become relevant to our research on apnoea speech.

Fricative sounds, FRI: an extra phonetic class is introduced in order to group all sounds which were not assigned to the previous classes. Considering our designed apnoea corpus, most of these sounds are fricative, although others, such as liquid sounds, will also be included in this fourth class. By grouping all of these sounds, we complete our classification of sounds, introducing a quite artificial group which includes sounds with rather different characteristics, though, as we just said, fricatives form the most significant subset.

Using these four phonetic classes, our purpose will now be to explore any differences that could be found between OSA and healthy speakers using speech recognition technology. But before that, we have to give some details on how GMM models, as described in Section 5, were trained and how differences between phonetic classes were measured.

6.1 Training data and GMM characteristics

Considering the previous description of the four phonetic classes, it is important to note that as we are modelling them using GMMs, the linguistic differences between phonetic classes will not generate non-confusable or non-overlapping models. Besides the overlapping of the acoustic spaces in particular realizations of each phonetic class, the discriminative power of GMM-s depends on different factors, such as the size of the model (i.e., number of gaussians), amount of training data and the acoustic front-end parameterization. In our case, the automatic segmentation of phonetic units can also be a source or errors that, as we discussed before, could lead to some overlap between the acoustic spaces modelled by different phonetic-class GMMs. Being aware of all of these differences from ideal acoustic models, the use of broad phonetic classes allows us to ensure that, as long as our segmentation of the utterances is precise enough, the number of spurious frames will be negligible compared to the amount of reliable data, so little distortion is expected in the estimation of acoustic parameters. As we will see in Section 7, the trained GMM models deliver a classification rate that is accurate enough not only to discriminate between phonetic classes, but also to measure differences in the acoustic realizations between OSA and control speakers.

In summary, the full acoustic space in our speech database was divided, through automatic phonetic segmentation, into the four phonetic classes previously described (see top of Figure 1). Consequently, the amount of data available to train each of the four phonetic classes was different as well as the internal complexity of their statistical distributions. However, as the speech corpus was designed to have a homogeneous coverage of main phonetic contexts relevant to OSA pathology, we decided to model each phonetic class using GMM models with equal number of gaussian components. So, based on the amount of available training data, 64 gaussians were considered enough to properly represent the different acoustic complexities of the different phonetic classes.

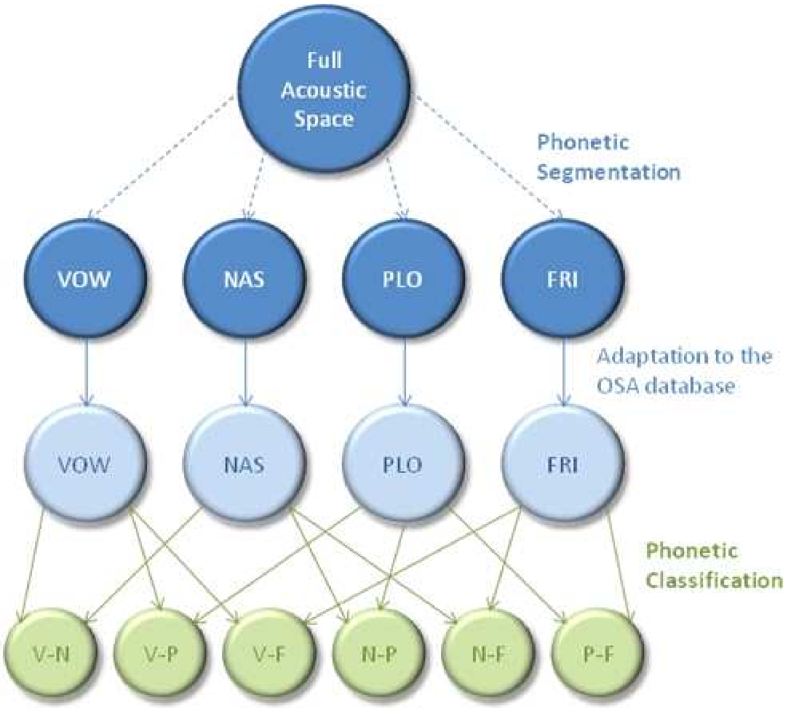

Figure 1. Brief description of the segmentation and adaptation processes (Section 6), as well as of the classification tests performed (Section 7)

6.2 Training reference GMMs for each phonetic class

As it was stated in Section 5, our aim is to explore acoustic differences between apnoea and control speakers using a set of GMMs trained from control or healthy voices. These voices served as reference GMMs to measure possible deviations of the **pathological voices** of apnoea speakers. As a result, these reference GMM models had to be trained from a control population, but, due to the limited amount of speech obtained from the control speakers in our OSA database, the first set of GMM models were trained using Albayzin database (Moreno et al. 1993). After that, as Figure 1 illustrates, these initial models were adapted to the control speaker population of our OSA database to generate the final four reference GMMs.

The whole training process can be described as follows: utterances from the Albayzin corpus, already manually segmented into phonemes, were labelled according to the four phonetic classes we defined. Once grouped, the feature vectors were used to estimate a GMM model for each phonetic class separately, resulting in four different models, namely: VOW, NAS, PLO and FRI. In the second step, as Figure 1 illustrates, a MAP algorithm (Reynolds et al. 2000; incorporating these initial GMM models) was used to adapt these models to the OSA database; specifically to its acoustic conditions (microphone and recording room) and population characteristics. To complete this adaptation process, speech utterances from the control speakers were automatically segmented and labelled using the process described in Section 5. Note that, as it was said before, only control speakers were used in the adaptation process to generate the final set of reference (healthy) GMM models for each phonetic class.

6.3 Measuring differences between phonetic classes

Once reference GMM models were trained, we tested their ability to classify different speech segments as belonging to different phonetic classes. Our hypothesis was that their discriminative power will be lower for apnoea speakers than for control speakers, as the pathological characteristics of apnoea voice make them more confusable. These differences in phonetic class discrimination then could be exploited to detect apnoea cases.

Thus, given a set of acoustic features corresponding to a particular phonetic class produced by a particular speaker, the discriminative measure used will be the difference between the logarithm of two likelihood scores (i.e. log-likelihood ratio, LLR, Reynolds et al. 2000): one score was the log-likelihood of generating the set of acoustic features using the true GMM model (i.e., the correct phonetic class), and the other score was the log-likelihood obtained using a different phonetic class. As can be seen in the bottom of the diagram in Figure 1, we explored the differences between every pair of GMM phonetic classes (V-N, V-P, V-F, N-P…).

We should bare in mind that the difference between likelihood scores are closely related to the Kullback-Leibler divergence (KLD) –also known as *discrimination information*– [17]. KLD is the most common approach to measure differences between the statistical distributions of two classes and to decide which of the two models most likely generated a certain sample. This theoretical measure can be estimated either by calculating the average likelihood ratio between two models over a set of feature vectors, or by considering an analytic approximation to it. This analytic approach will only be used in subsection 7.4, while likelihood averaging will be used in subsections 7.1 to 7.3.

7 Experimental results

Several experiments were developed to explore differences between the four phonetic classes, and all of them were based on the differences between log-likelihood ratios (LLRs) for control and apnoea speakers. To provide a fair test, both the adaptation of the reference GMM models and LLRs were estimated using the leave-one-out cross-validation test protocol. According to it, for all tests involving a particular speaker in the control group, the four reference GMMs were trained through MAP adaption using our OSA database, but excluding (leaving-out) this particular speaker's records. Z-score normalization was used to fairly compare results for the different phonetic classes and to consider their posterior fusion at the score level.

To quantify the acoustic mismatch between apnoea and control speech, two different approaches were considered, both of which were evaluated over a given sequence of acoustic features belonging to a particular phonetic class:

- First, our reference GMM models were used as classifiers of phonetic classes. We explored whether different performance rates in classification could be found for the control and apnoea populations. Note that in this case, we were not really classifying control/apnoea speakers but only exploring whether significant differences in phoneme classification exists across both groups. This result will provide some insights into the effects of apnoea on the speech of OSA patients.
- In a second set of experiments, control/apnoea classification was evaluated using average LLR from the reference GMM model corresponding to the true phonetic class, and the GMM of a different or competing phonetic class. Due to voice anomalies in apnoea patients, this average LLR was found to be different for control and apnoea speakers (i.e. higher for control speakers and lower –greater confusability– for apnoea speakers).

Finally we will conclude this Section by discussing how GMM models trained for vowel sounds in nasal and non-nasal contexts show an interesting distinctive pattern for apnoea speakers that should be explored in future research.

7.1 Differences in classifying phonetic classes

In this initial experiment, the discriminative power of the reference GMM models were evaluated using them for classification and comparing them across both the apnoea and control populations.

For each speaker in our database all the speech segments corresponding to the different phonetic classes were used to obtain the average LLR scores. Those were calculated as the mean difference of the log-likelihood values estimated for each speech sample by considering two reference GMM models (each of them corresponding to a phonetic class model). Thus, for each pair of phonetic classes, two different errors were possible: a) **missed recognition,** when a speech sample belonging to the first class was more likely to be generated by the second one, and b) a **false alarm,** when a speech sample belonging to the second phonetic class was more likely to have been generated by the first phonetic class model being evaluated. Depending on the decision threshold used across LLR scores, these two types of errors should be **opposite** (i.e. lower false alarm rates lead to higher missed recognitions, and vice versa) and can providing different operational points. Detection error trade-off (DET) curves have been widely used to represent the evolution for both types of errors (Reynolds et al. 2000), but also the discriminative power of a classifier can be described using a single Equal Error Rate value (EER). The EER corresponds to the operational point of equal missed recognition and false alarm errors. In Table 2, EER values representing the pair-wise phoneme class classification errors using the reference GMM models are presented. Different EER values are presented for both control and apnoea (bold values) populations.

From these results, we can see that classification rates are significantly different from one class to the other, though the results are reasonably good for all of them (the worst case being an EER of 11.7% classifying plosives vs. fricatives in the apnoea population). For vowels, results were particularly good when compared to those for nasals and plosives, as almost no errors appeared when testing over the whole data set for both groups of speakers. Other pairs do exhibit small, but meaningful, error values with quite different results. However, they all reflect a common trend: performance for the apnoea group is worse than the control group. This result suggest a systematic deviation from the reference acoustic phonetic classes in apnoea speakers which can be related to the physiological factors associated with OSA. For instance, the increase in EER value when comparing nasal sounds (NAS) to fricatives (FRI) and plosives (PLO) can be explained by the fact that patients suffering from the OSA syndrome exhibit abnormal velopharyngeal function, so this could alter the production of nasal sounds, introducing a slight oral plosive and fricative articulation due to partial palatal paralysis.

Table 2. EER values resulting from phonetic classification of all pairs for the four phonetic classes. Bold values correspond to the apnoea population, while the normal ones were estimated for the control group.

	VOW	NAS	PLO	FRI
VOW		0.0% **0.0%**	0.0% **3.3%**	1.7% **3.3%**
NAS	–		0.0% **4.2%**	1.7% **6.7%**
PLO	–	–		3.3% **11.7%**
FRI	–	–	–	

7.2 Phonetic classes for OSA detection

Based on the different classification results for control and apnoea populations previously described, we will now analyze whether the underlying differences in LLR scores could be used to classify a speaker as belonging to the control or apnoea population. LLR was evaluated in the same way as described in subsection 7.1: using two competing GMM models, but in this case, it was only averaged for speech segments corresponding to a single phonetic class. That is, in this experiment the phonetic class of the speech segment was known (as provided by the automatic phonetic segmentation process), but whether the speaker belonged to control/apnoea group was unknown.

Therefore, for a given speaker to be tested, 4 speech segments, one for each phonetic class, were used, and, for each segment, 3 different average LLRs were obtained. For example, for the speech segments corresponding to the vowel phonetic class, three different LLR scores were obtained using the V-N, V-P and V-F pairs of reference GMM models. Consequently, using each one of these three LLRs, three different control/apnoea classification results were considered. So far, when speech segments for all phonetic classes were used, and LLRs for all possible combinations of reference GMMs were used, a total of 12 control/apnoea classifiers were evaluated.

As in the previous experiment, evaluation for this set of control/apnoea classification systems was based on the **miss recognition** and **false alarm** errors, but in this case missed recognition meant that an apnoea speaker was incorrectly classified as a control speaker, and a false alarm signalled a control speaker being classified as having apnoea. Control/apnoea classification results, in terms of EER values for each one of the 12 classifiers, are presented in Table 3.

Table 3. EER values for control/apnoea classification using speech segments of the four phonetic classes and LLR scores for all pairs of reference GMM models

	VOW	NAS	PLO	FRI
VOW-NAS	46.7%	38.3%	–	–
VOW-PLO	42.5%	–	47.5%	–
VOW-FRI	47.5%	–	–	40.8%
NAS-PLO	–	37.5%	50.0%	–
NAS-FRI	–	39.2%	–	44.2%
PLO-FRI	–	–	46.7%	33.3%

From these results we can see that apnoea could be detected with an accuracy as high as 33% EER, which is rather surprising as this best classification result was obtained when considering fricative samples evaluated using LLR scores from fricative vs. plosive reference GMMs. In contrast, a very poor discrimination rate was attained when plosives were compared to fricatives. A possible explanation for this apparently odd result could be that in this experiment what we consider is not just the deviation from a perfect fit to the reference phonetic class models, but also the deviation towards a certain phonetic class. So in this case, fricative sounds in the apnoea group show a deviation towards plosive reference sounds. The same idea explains the results obtained when we compared nasals and vowels or nasals and plosives. Looking at other results in the Table, there are cases where both comparisons provided rather similar results for samples from both classes. This finding indicates that the distortion in one direction is about the same in the opposite one, just as it happens for vowels and plosives, vowels and fricatives or nasals and fricatives.

The results from nasal speech segments (NAS column in Table 3) require a more extensive explanation. According to the reviewed literature, abnormal resonances in speech are characteristic of OSA patients, particularly when considering the nasalization of connected vowels. Therefore, it was expected that nasal sounds would be useful cues in the design of an automatic system for OSA detection. In fact, Table 3 shows lower global EER values for the NAS column when compared to other phonetic classes. Consequently, the effects of vowel nasalization required from a specific analysis, which we describe in Section 7.4 by considering two different phonetic class subsets for vowels: those in nasal or non-nasal phonetic contexts.

7.3 Improving detection by the combination of pairs

From Table 3, it seems clear that classification results are poor for each of the individual classifiers. In this section, we will try to improve those results by combining all 12 classifiers into a single one. This is a complex task which generally requires a large amount of data to guarantee that the optimal combination is found. Since the current dataset is small, we could not implement an optimal approach, but used a suboptimal one, which was designed to iteratively improve binary classification.

The combination process used was based on the algorithm described by Al-Ani et al. (2003), though conditional mutual information calculations were substituted by EER estimations, which are in fact the posterior error probabilities discussed in that article. The idea was to improve classification rates by linearly weighting normalized scores and adding them up; but only if the overall results were noted to improve. In order to avoid any redundancies and spurious effects which could detrimentally affect the results, all combinations (successive pairs, triplets, quartets, etc.) were tested in order to identify the optimal one. However, as suggested by Al-Ani et al. (2003), good results (though suboptimal) can be obtained by iteratively combining the weighted classifier with the best and most uncorrelated spare classifier, reducing the computational complexity.

The results from all these combinations are presented in Figure 2 using DET curves. The final DET curve, corresponding to the combined system, returns a 28.33% EER. This final DET curve is presented along with a different set of DET curves in Figures 2a and 2b. In Figure 2b (right plot), the different successive DET curves illustrate how successive classification improvements are obtained during the iterative algorithm.

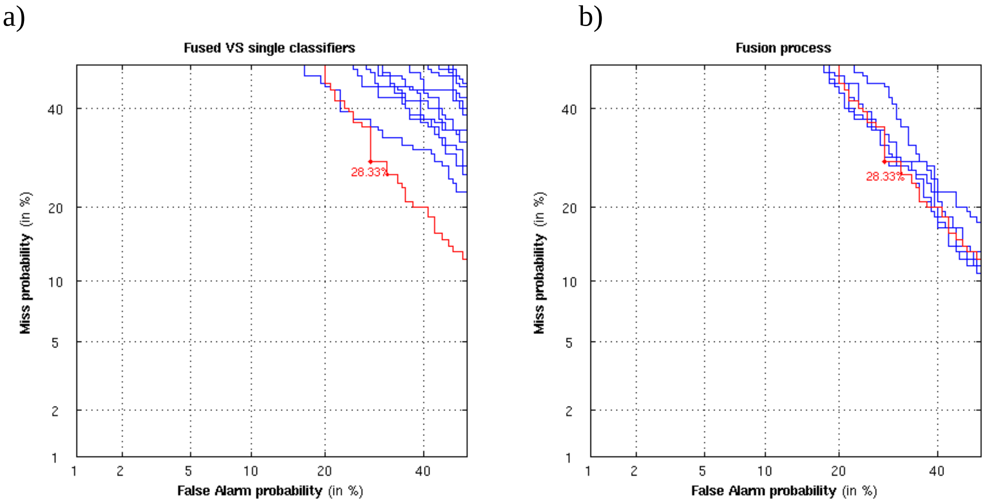

Figure 2. DET curve resulting from the combination of the 12 phonetic classifiers: the left one (a) compares the resulting DET curve with the ones estimated for each single classifier; the right one (b) compares the results from the iterative improvement algorithm.

As can be seen in this Figure, most of the improvement is achieved during the first iterations, while only marginal improvements occur later in the process. When comparing the final combination result to each of the twelve prior classifiers (the 12 upper right DET curves shown in Figure 2a), it becomes clear that there is a strong correlation among classifiers, although there is a considerable gap between the best prior classifier and the one resulting from the fusion process, improving the overall classification by 5%.

7.4 Exploring vowel nasalization

With the previous results estimated for control and OSA speakers in mind, nasalization effects (affecting both nasal and connected vowel sounds) seem to be a relevant phenomenon in apnoea detection. In order to improve our understanding of the side effects of the abnormal coupling and decoupling of the nasal and oral cavities, as well as to continue to rework Fox & Monoson's (1989) experiments by means of automatic speech processing, an additional exploratory experiment was carried out. The abnormal resonances described in Fox and Monoson's work could be perceived as a form of either hyponasality or hypernasality (no nasalization is produced when the sound should be nasal, or nasalization is produced during the pronunciation of non-nasal –voiced oral– sounds). In other words, OSA speakers will nasalize when they are not expected to, and/or vice versa. As a consequence, we will expect statistical models (GMMs) trained with such data to exhibit smaller differences when comparing models for vowel sounds in nasal and non-nasal contexts. This idea could be tested by measuring the distances between both models in each group of speakers.

Acoustic feature vectors for vowel sounds were grouped into two different subsets, based on whether their phonetic context was nasal or non-nasal, i.e. depending on whether they should be nasalized or not. The amount of available data for the original VOW phonetic class was enough to build the class model for the previous experiments, and is even big enough for our tests once we redistribute samples among these two nasal sub-classes. However, since we have reduced the size of the data set in this experiment, the KLD analytic approximation (Do 2003) was chosen. Therefore, four different models were trained by adapting the original VOW GMM: two GMMs adapted to vowels in non-nasal context (one for control and the other for apnoea speakers), and two GMMs for vowels in nasal contexts (also for control and apnoea voices).

As a test of the stability or consistency of our KLD approximation, these four GMM models were trained and the corresponding KLD distances were evaluated 40 times, each time using a different subset of 39 control and apnoea speakers extracted from our database. Figure 3 represents the resulting 40 KLD distance values obtained for GMM models for vowels in nasal and non-nasal contexts (speaker index in the Figure corresponds to the excluded speaker in the 39 speakers' subgroup). As it can be seen in Figure 3, significant differences in the nasal/non-nasal GMM distances were found for the control and apnoea speakers. This result suggests that acoustic differences between oral and nasal vowels are smaller in apnoea speakers and confirms the trend to an overall higher nasality level, as revealed in previous research.

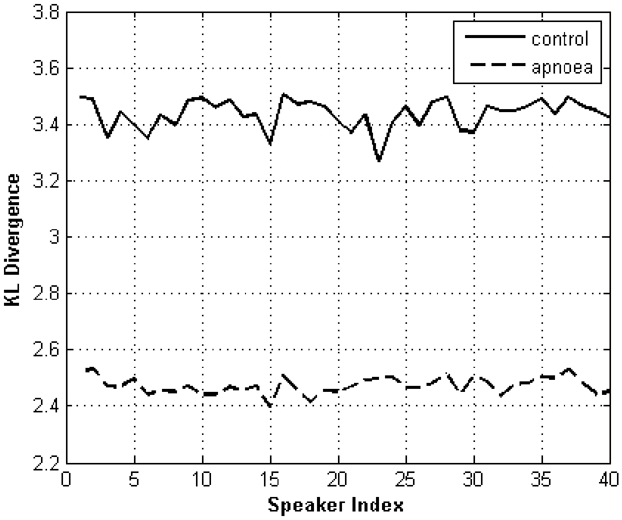

Figure 3. KLD approximation values between Gaussian Mixture Models for vowels in nasal and non-nasal contexts

8 Conclusions and future research

In this paper, some of the characteristic speech patterns that can be observed in speakers suffering from severe obstructive sleep apnoea (OSA) syndrome have been analyzed by comparing phonetic classes using a specifically designed speech corpus. This study offers an innovative perspective on how phonetic information can be used in pathological voices analysis using conventional automatic speech processing techniques.

Regarding Fox & Monoson's research as a reference, a "perceptual" representation of the speech signal using Mel-frequency cepstral coefficients was used. From this acoustic representation, experimental results were obtained using Gaussian Mixture Models (GMMs), which were initially trained on a large Spanish speech database and adapted to a control population. These GMMs were generated for the four broad phonetic classes and then used as reference patterns to explore possible acoustic mismatches in the voices of speakers with apnoea.

Differences in phonetic classification for control and apnoea populations were observed for the four phonetic classes. These results suggest that certain phonetic groups are more likely to be misclassified when the speaker suffers from apnoea. Using all different pair-wise reference GMM models, control/apnoea classification was also evaluated using log-likelihood ratio scores averaged over segments of speech corresponding to different phonetic classes. The minimum 33% EER obtained when using single classifiers, was improved to 28.3% when combining all of them through an iterative linear weighting algorithm.

Finally, various effects addressed in the previous literature were identified in our experiments, supporting the interpretation of the automatic speech recognition results. Reworking Fox and Monoson’s (1989) experiments has allowed us to come to the same conclusions they did. Though further analysis is needed, apnoea speakers certainly exhibit smaller intra-class differences during vowel nasalization. This side-effect is probably related to an abnormal coupling of the nasal cavity.

Our results are intended to shed some light on the peculiarities that phonetic classes exhibit when comparing healthy speakers to those suffering from OSA. Results obtained in control/apnoea classification were also promising, though still much work needs to be done. Besides, there is still a need for a larger speech database to continue study in this area. We shall focus on this need, while encouraging research to improve our understanding of the effects of OSA on speech signals.

9 Acknowledgments

The activities described in this paper were funded by the Spanish Ministry of Science and Technology as part of the TEC2009-14719-C02-02 project. The authors would like to thank the volunteers at Hospital Clínico Universitario of Málaga, Spain, and to Guillermo Portillo who made the speech and image data collection possible. José Luis Blanco also acknowledges Universidad Politécnica de Madrid scholarship’s support.

References

Al-Ani, A., Deriche, M. and Chebil, J. 2003. A new mutual information based measure for feature selection. *Intelligent Data Analysis,* 7(1): 43-57.

Blouet, R., Mokbel, C., Mokbel, H., Sanchez Soto, E., Chollet, G. and Greige, H. 2004. BECARS: a free software for speaker verification. In *Proceedings of The Speaker and Language Recognition Workshop.* ODYSSEY. 145-148.

Coccagna, G., Pollini, A. and Provini, F. 2006. Cardiovascular disorders and obstructive sleep apnoea syndrome. *Clinical and Experimental Hypertension,* 28: 217-224.

Davidson, T. M. 2003. The great leap forward: The anatomic evolution of obstructive sleep apnoea. *Sleep Medicine,* 4: 185-94.

Do, M. N. 2003. Fast approximation of Kullback-Leibler distance for dependence trees and Hidden Markov Models. *IEEE Signal Processing Letter,* 10: 115-118.

Fernández, R., Blanco, J. L., Hernández, L. A., López, E., Alcázar, J. and Torre, D. T. 2009. Assessment of severe apnoea through voice analysis, automatic speech, and speaker recognition techniques. In *EURASIP Journal on Advances in Signal Processing,* vol. 2009. 1-21.

Fernández, R., Hernández, L. A., López, E., Alcázar, J., Portillo, G., and Torre, D. 2008. Design of a multimodal database for research on automatic detection of severe apnoea cases. In *Proceedings of 6th Language Resources and Evaluation Conference*. Marrakech, Morocco. 1785-1790.

Fox, A. W. and Monoson, P. K. 1989. Speech dysfunction of obstructive sleep apnea. A discriminant analysis of its descriptors. *Chest Journal,* 96(3): 589-595.

Godino-Llorente, J. I., Gómez-Vilda, P., Sáenz, N., Blanco-Velasco, M., Cruz, F. and Ferrer, M. A. 2006. Support vector machines applied to the detection of voice disorders. In Hussain, A., Faundez-Zanuy, M. and Kubin, G. (eds.): *Lecture notes in computer science, Vol. 3817: Nonlinear analyses and algorithms for speech processing.* Heidelberg: Springer Verlag. 219-230.

Hébert, M. and Heck, L. P. 2003. Phonetic class-based speaker verification. In *Proceedings of the Eurospeech 2003 Conference*. Geneva, Switzerland. 1665-1668.

Huang, X., Acero, A. and Hon, H-W. 2001. *Spoken language processing: A guide to theory, algorithm and system development.* New Jersey: Prentice-Hall.

Kullback, S. and Leibler, R. A. 1951. On information sufficiency. *Annals of Mathematical Statistics,* 22(1): 79-86.

Lee, W., Nagubadi, S., Kryger, M. H. and Mokhlesi, B. 2008. Epidemiology of obstructive sleep apnea: a population-based perspective. *Expert Review of Respiratory Medicine,* 2(3): 349-364.

Lloberes, P., Levy, G., Descals, C. et al. 2000. Self-reported sleepiness while driving as a risk factor for traffic accidents in patients with obstructive sleep apnoea syndrome and in non-apnoeic snorers. *Respiratory Medicine,* 94: 971-976.

Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J. B. and Naude, C. 1993. ALBAYZIN speech database: Design of the phonetic corpus. In *Proceedings of Eurospecch 93*. Vol. 1. Berlin, Germany. 175-178.

Reynolds, D. A., Quatieri, T. F. and Dunn, R. B. 2000. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing,* 10: 19-41.

Robb, M., Yates, J. and Morgan, E. 1997. Vocal tract resonance characteristics of adults with obstructive sleep apnea. *Acta Otolaryngologica,* 117: 760-763.

Turner, G. S., Tjaden, K. and Weismer, G. 1995. The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis. *Journal of Speech and Hearing Research,* 38: 1001-1013.

Young, S. 2002. *The HTK Book (for HTK Version 3.2).* First published December 1995, Revised for HTK Version 3.2 December 2002.