Measurement, Analysis, and Detection of Nasalization in Speech

Xiaochuan Niu

B.S., Tsinghua University, 1994
M.S., Institute of Automation, Chinese Academy of Sciences, 1998

A dissertation presented to the faculty of the
OGI School of Science & Engineering
at Oregon Health & Science University
in partial fulfillment of the
requirements for the degree
Doctor of Philosophy
in
Computer Science and Electrical Engineering

March 2008

The dissertation "Measurement, Analysis, and Detection of Nasalization in Speech" by Xiaochuan Niu has been examined and approved by the following Examination Committee:

Jan P.H. van Santen
Professor
Thesis Research Adviser

Gary Weismer
Professor

John-Paul Hosom
Assistant Professor

Alexander Kain
Research Assistant Professor

Dedication

To my wife, Hui Yang, and my parents, Guangxiao Niu and Jing Zhao.

Acknowledgements

This work is a result of efforts by many people. I am grateful to all of them for making this work possible.

First of all, I would like to thank my adviser, Dr. Jan van Santen, for his constant support, encouragement and inspiration. Thanks are also due to my other committee members, Dr. Gary Weismer, Dr. John-Paul Hosom and Dr. Alexander Kain, for reviewing this document and for their suggestions to improve this work. I would like to say a special thanks to Alex, Esther, Johan and Paul, for so many inspiring discussions during the whole period of my Ph.D. program.

I would like to thank the following faculty members: Peter Heeman, Brian Roark, Esther Klabbers, Izhak Shafran, Lois Black, Xubo Song, Eric Wan, Todd Leen and Misha Pavel, for making CSLU and CSEE at OGI such a simulating place to study and to do research. I would also like to thank the departmental staff: Jacques de Villiers, Rachel Coulston, Margaret Mitchell, Pavel Chytil, Mark Bouza, Patricia Dickerson and Kim Basney, for always offering a helping hand.

Last but not least, I want thank my colleagues, including Fan Yang, Taniya Mishra, Qi Miao, Emily Tucker, Kristy Hollingshead, Akiko Kusumoto, Seeger Fisher, Nathan Bodenstab and Pete Jacobs, for their friendship. Some of them have helped me to record speech corpus, do perceptual experiments or proof-read papers, and we have shared a wonderful time during the years in OGI.

Contents

List of Tables

List of Figures

Abstract

Measurement, Analysis, and Detection of Nasalization in Speech

Xiaochuan Niu, B.S., M.S.

Ph.D., OGI School of Science & Engineering
at Oregon Health & Science University

March 2008

Supervising Professor: Jan P.H. van Santen

Nasalization refers to the process of speech production in which significant amounts of airflow and sound energy are transmitted through the nasal tract. In phonetics, nasalization is necessary for certain phonemes to be produced in normal speech; and it can also be a normal consequence of coarticulation. In disordered speech, however, inappropriate nasalization can be one of the causes that reduces the intelligibility of speech. Instrumental measurement and analysis techniques are needed for better understanding the relationship between the physiological status and the aerodynamic and acoustic effects of nasalization during speech. The main aim of the research work presented in this dissertation is to investigate the aerodynamic and acoustic effects of nasalization, and to develop objective approaches to measure, analyze, and detect the nasalized segments in speech. Based on an extensive survey of existing literature on the measurements of velopharyngeal function, the acoustic production models of speech, the analysis methods and results of normal nasalization, and the analysis methods of resonance disorders, it is understood that the

final acoustic representation of nasalization is a complex outcome that is affected by the degree of velopharyngeal opening, the variation of vocal tract configurations, the mixture of multiple acoustic channels and speaker differences. It is proposed to incorporate more available information besides single channel acoustic signals during the analysis of nasalization. In our research work, a parallel study of acoustic and aerodynamic signals reveals the complimentary information within the signals. In addition, dual-channel acoustic studies help to understand the acoustic relationship between the oral and nasal cavities, and show inherent advantages over the single-channel analysis. Based on the derivation and analysis of the dual-channel acoustic properties, automatic detectors of nasalization are developed and successfully tested. The techniques developed in these explorations provide novel instrumental and analysis approaches to possible applications such as phonetic studies of the normal nasalization process, clinical assessment of disordered nasal resonance, and special feature extraction for speech recognition.

Chapter 1

Introduction

1.1 Background

1.1.1 Nasalization in speech

Speech is a complex phenomenon involving voluntary movements of human vocal apparatus producing vibrations of air that can be perceived by human auditory systems. In order to understand this complex phenomenon, researchers sometimes regard the speech process as five separate but coordinating components (or subsystems), including respiration, phonation, resonance, articulation, and prosody (Duffy, 1995). The *respiration* process provides steady subglottal air pressure and airflow as the initial source of sound waves. The *phonation* process involves the adjustment or sustaining of a proper tension of vocal folds for vibration, producing a fundamental frequency perceived as pitch. The *resonance* process refers to the way in which the oral and nasal cavities are connected or separated to construct various acoustic resonators that result in different nasal qualities of the sound. The *articulation* process refers to the control of the shapes and movements of articulators including the tongue, lips and teeth, so as to modify the sound waves produced by the respiration, phonation and resonance processes. The *prosody* process is a higher level mechanism than other processes, which conveys information about syntactic factors such as lexical stress, phrase breaks and sentence structures, and para-linguistic factors such as attitudes and emotions of the speaker, by manipulating patterns of fundamental frequency, segmental duration, intensity, pauses, and speaking rate of speech.

Among these components, the *resonance process* is specifically achieved through the control of the velopharyngeal (VP) port inside the vocal tract. In physiology, the *VP port*

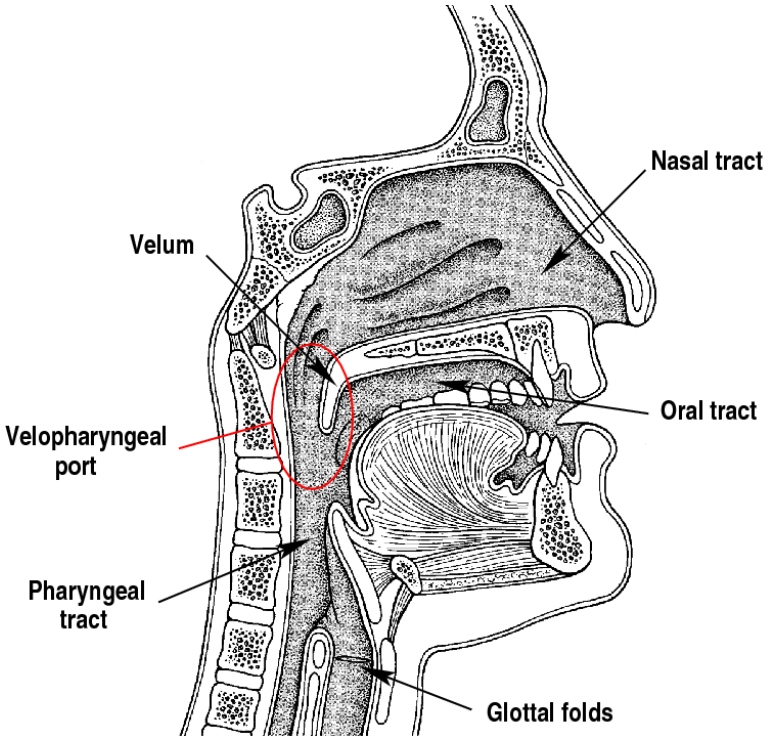


Figure 1.1: Schematic drawing of the vocal tract (adopted from Denes and Pinson, 1993). The section in the circle that connects the pharyngeal tract, the nasal tract, and the oral tract is called the velopharyngeal (VP) port.

refers to the section in the vocal tract that connects the nasal tract with the oral-pharyngeal tract (see Figure 1.1). The VP port usually stays open when one is not speaking, but can be either closed or opened voluntarily during speech. Physiological studies have shown that the VP movement is a complex combination of the movements of the velum, the rear pharyngeal wall and the lateral pharyngeal walls (Skolnick *et al.*, 1975). It is known that several muscles coordinate in controlling VP movements; and some of these muscles are also involved in the movements of other articulators, such as the tongue and the larynx. It is a natural capability of normal speakers to control their VP port during speech. However, the resonance process can be disturbed due to various reasons such as physiological defects, lack of auditory feedback, or motor control disorders.

Before moving on to further discussions, we would like to explicitly clarify the meanings of the following related terms that will be used in the context of this thesis.

Nasalization refers to the process of speech production in which the velopharyngeal port is opened and the nasal cavity is connected with the pharyngeal-oral tract. As a consequence of nasalization, significant amounts of airflows and acoustic vibrations propagate through the nasal tract during speech.

Nasal resonance refers to the voiced vibrations of sound pressure transmitted out of the nasal cavity through the nose during speech.

Nasal turbulence refers to the fricative noise sounds caused by the airflow passing through the nasal tract.

Nasal emission refers to the leakage of the nasal airflow during the production of phonologically "oral" phonemes.

Nasality is the perceptual quality of nasal resonance.

Hypernasality is the perceptual quality of excessive nasal resonance.

Hyponasality is the perceptual quality of insufficient nasal resonance.

In summary, *nasalization* is a general concept of speech production related to the VP opening; *nasal resonance* and *nasal turbulence* are acoustic descriptions of certain characteristics of nasalization; *nasal emission* is the aerodynamic description of an undesired event during speech production; *nasality*, *hypernasality*, and *hyponasality* are perceptual attributes of sounds.

This thesis will discuss *measurement*, *analysis* and *detection* techniques for studying the nasalization process. In the context of this thesis, **measurement** refers to any method that converts certain physical properties of the nasalization process into quantitative representations, e.g. electrical signals; **analysis** refers to a procedure or algorithm that extracts useful information of the nasalization process from raw signals; **detection** refers to an artificial approach through which different phases of the dynamic nasalization process can be automatically inferred.

1.1.2 Studies of nasalization

Nasalization can be a natural process of speech production. In the phonetic systems of languages in the world, [+/-nasal] is one of the discriminative features that groups speech sounds into different categories of phonemes (Ladefoged, 1993). Nasal consonants are achieved by opening the VP port and simultaneously closing the oral tract at a certain point, so that airflows and acoustic vibrations from the vocal fold propagate only through the nasal tract. On the other hand, the closure of the VP port is a prerequisite for oral "pressure" phonemes, such as plosives, affricates and fricatives, because sufficient intra-oral air pressure is needed to produce these phonemes. In addition, while vowel production generally requires an open passage through the oral tract, there are nasal vowels that are phonemically contrastive to oral vowels in some languages such as French and Portuguese, in which the VP port is open at the same time as the oral tract. Though there are no phonemic nasal vowels in languages such as American English, vowels and consonants can be nasalized in the context of nasal consonants due to coarticulation effects.

Appropriate manipulation of the VP port is important for normal speech production. However, any defects in accomplishing the right degree of openness or timing of the VP opening and closure may degrade the intelligibility of speech, thereby reducing the communicative ability of a speaker. Several groups of speech disorders are characterized by inappropriate nasalization (Baken and Orlikoff, 2000, Chapter 11). For example, a cleft palate or other anatomical defect that changes the normal physiological structure of the VP port or the nasal tract may cause excessive or insufficient nasalization in speech. Deaf speakers may exhibit excessive nasalization due to the lack of auditory feedback. Inappropriate nasalization is also a prevalent characteristic of several groups of dysarthric speech, due to the motor control impairments of some neurologically affected individuals.

A basic scientific question concerning the VP function (or behavior) is how the VP port works appropriately during normal speech. A thorough answer to this question is one of the basic goals of phonetic and phonological studies. It is also of interest to understand the relationship between VP movements and the corresponding characteristics of speech sounds. A deep understanding of the aerodynamic, acoustic and perceptual consequences

of nasalization may benefit engineering applications such as automatic speech recognition and speech synthesis, when such knowledge can be properly applied. In addition, the understanding would help the diagnosis, assessment, and treatment of disordered speakers with nasalization problems. The research work presented in this thesis may not answer the ultimate question about the control mechanism of the nasalization process; however, we are trying to develop practical techniques that may help to analyze the relationship between the physiological condition and the aerodynamic or acoustic consequences of nasalization.

1.2 Research motivations

A long term research goal at the Center for Spoken Language Understanding (CSLU) has been to develop assistive techniques that can enhance the intelligibility of a certain group of disordered speech, *dysarthria*, in an *analysis-transform-synthesis* framework (Hosom *et al.*, 2003; Kain *et al.*, 2004, 2007). Dysarthria results from disturbances in muscular control over the speech mechanism due to damage to the central or peripheral nervous system (Darley *et al.*, 1969). It has been observed that inappropriate nasalization is one of the reasons that causes reduced intelligibility of dysarthric speech. For example, when voiced sounds are contaminated by inappropriate acoustic coupling of the nasal tract, the spectral patterns of these sounds are "blurred" and the intelligibility is reduced (e.g. Weismer and Martin, 1992; Forrest and Weismer, 1997); when significant nasal emission occurs during oral "pressure" phonemes, such as plosives, affricates and fricatives, the distinguishing features of these sounds are weakened, thus reducing the intelligibility. It is reasonable to attribute the nasalization problems in dysarthric speech to inappropriate degree and timing of VP movements.[1] However, informative and objective approaches to analyze the nasalization process over time are still lacking, mainly due to measurement and analysis difficulties. Therefore, we focus on developing feasible objective measurement and analysis approaches to extract the information of the nasalization process, so that the resulting techniques may be used in, but not limited to, the applications of disordered

[1] It is to be noted that the speech of hearing-impaired speakers can exhibit similar nasalization problems due to lack of auditory feedback.

speech enhancement.

There are some special challenges in studying the nasalization process. First, the position of the VP port is "hidden", which makes any direct measurements of VP movements difficult. Though some measurement techniques have been developed to study the VP function (see a review in Section 2.1), none of them can provide the whole spectrum of information under investigation. Second, the VP port is not just a two-state valve that opens for nasals and closes for non-nasals instantaneously; it varies continuously due to either anticipatory or carry-over movements of the velum and the pharyngeal walls. These coarticulation effects complicate the timing pattern of VP movements (Moll and Daniloff, 1971; Kent *et al.*, 1974) in both normal and disordered speech. Third, the aerodynamic and acoustic consequences of nasalization are not independent of other features of speech. For example, the different degree of oral opening for various vowels highly influences the amount of airflow and acoustic energy emitted through the nasal tract (Curtis, 1970; Laine *et al.*, 1988). Complex aerodynamic and acoustic techniques are needed to reduce the adverse effects resulting from these dependencies. Fourth, speech components other than the resonance process, such as phonation (Imatomi *et al.*, 1999; Imatomi and Arai, 2002) and articulation (Maeda, 1993), may compensate for the perceptual resonance effects. This may undermine the reliability of perceptual measurements of nasalization.

From the speech analysis point of view, the *specific objectives* of the research work presented in this thesis are 1) to choose and/or develop instrumental measurement techniques for nasalization analysis, 2) to better understand the aerodynamic and acoustic consequences of VP movements, and 3) to develop reliable approaches for automatic extraction of nasalization information and detection of the speech segments that are nasalized. Explorations and solutions leading to these aims should be beneficial to both understanding the normal nasalization process and assessing disordered nasalization.

1.3 Proposed approaches

In general, the existing techniques for the measurement of the VP function can be grouped as perceptual, physiological, and indirect approaches. Perceptual approaches include designing and conducting perceptual evaluations of nasality. Audio-perceptual rating procedures have been used extensively in clinical assessments, but they often suffer from reliability problems and cannot provide a detailed description of nasalization over time (e.g. Kent, 1994, 1996). Physiological approaches include measurements of muscle electrical activities, and imaging and tracking of VP movements. These approaches have been used in both phonetic studies and clinical practices. Indirect approaches include measurements of airflows and acoustic vibrations. Due to the accessibility of the instrumental resources and the inherent invasiveness, physiological approaches are not so appropriate for practical engineering applications such as assistive devices. In our research, we are mainly interested in developing indirect approaches for nasalization analysis, and more of the research emphasis is on acoustic approaches.

It is hypothesized that multiple-channel measurements can overcome some inherent limitations of conventional single-channel acoustic measurement in studying the nasalization process. In the thesis, we first investigate, develop and choose appropriate indirect measurement techniques, either aerodynamic or acoustic, that aim at revealing as much information about nasalization as possible. We then explore and analyze possible advantages and limitations of using certain multiple-channel techniques. It is also hypothesized that an acoustic model concerning the effects of multiple channels can explain the inherent ambiguities in conventional single-channel acoustic signals. Based on a simulation analysis of the process of nasal resonance, we aim to develop a quantitative acoustic model that describes the relationship between nasal and oral acoustic signals. It is then hypothesized that the acoustic properties of nasal resonance presented in the dual-channel acoustic model can be used as distinguishing features for automatic detection of nasalization conditions. We further investigate the problem of automatic detection of nasalized segments in running speech, developing a feature extraction algorithm based on the dual-channel analysis approach for a practical nasalization detection system.

It is to be noted that the analysis techniques in our studies are initially developed from the model of normal speech production and validated with normal speech. Considering the possible applications to disordered speech, we have made one assumption about the targeting group of disordered speakers, which is that their physiological speech apparatuses remain unimpaired and the nasalization problems are not caused by physiological defects. This assumption will hold for most dysarthric and hearing-impaired speakers, but not for speakers with cleft palate. The assumption allows normal and dysarthric speech to be analyzed on the same physiological basis of speech production. When the resulting methods are to be generalized to the nasalization problems caused by physiological defects, speaker-dependent factors must be taken into account.

1.4 Thesis organization

The research concerning the measurement, analysis, and detection of nasalization in speech will be presented in the following five chapters in rest of this thesis.

Chapter 2 provides an extensive summary of a wide range of relevant background information about the research on the VP function in speech. First, instrumental techniques for the measurement of the VP function are reviewed. Then, different methodologies of nasalization analysis are summarized and compared from a variety of viewpoints. The discussion will consequently lead to our original methodology of nasalization measurement and analysis.

In Chapter 3, some schemes of multi-channel measurements are investigated and analyzed. From the data acquisition point of view, a new aerodynamic device is designed and evaluated as one possible channel of measurement. Both advantages and limitations of the new device are discussed. Then a simple dual-channel acoustic device is evaluated. The aerodynamic-acoustic and the dual-channel acoustic schemes are compared to each other.

In Chapter 4, a dual-channel acoustic model is investigated theoretically. Then simulation experiments are conducted to test the derived properties of the model. A novel dual-channel analysis approach is presented and validated by simulated speech signals.

In Chapter 5, based on the analysis approach developed in Chapter 4, a dual-channel

acoustic feature for nasalization detection is presented; a practical algorithm is designed to extract this feature from dual-channel acoustic speech signals; a nasalization detector is then built and its performance is evaluated and reported.

Chapter 6 summarizes the contributions of the thesis work and discusses future work and possible applications.

Chapter 2

Related work on nasalization measurement and analysis

2.1 Measurement techniques

Many instrumental techniques have been developed to investigate the VP function in speech. Krakow and Huffman (1993) have presented a summary of most methods for the purpose of phonetic and phonological research. From the speech production point of view, they categorized these methods into three groups, including those for studying 1) the sources of VP movements, 2) the patterns of VP movements, and 3) the effects of VP movements. With an emphasis on clinical purposes, Baken and Orlikoff (2000) also had a thorough review of existing instrumental techniques for the assessment of the VP function. In this section, we will briefly examine these measurement techniques according to Krakow and Huffman's categorization in order to choose appropriate ones for our research goals. Typical examples will be elicited to highlight certain characteristics of each technique.

1. The sources of VP movements refer to muscle activities involved in the control of the VP port. Electromyography (EMG) is the technique that measures electrical activities associated with muscle contraction. EMG signals are picked up by pairs of electrodes that are either glued on the surface over muscles or inserted into muscles. Due to the complex layout of VP muscles, inserted electrodes are usually needed. EMG data are useful for studying the relationship between muscle activities and movements (e.g. Bell-Berti, 1976). However, because of the inherent invasiveness, this technique is not widely used for engineering and clinical purposes.

2. The patterns of VP movements can be studied with the help of imaging or tracking techniques. Through these techniques, researchers are able to observe the "hidden" structures and movements directly. The following techniques belong to this category.

   (a) The imaging techniques include fiberoptic endoscopy, radiography (X-ray), magnetic resonance imaging (MRI), and ultrasound.

      i. The most direct but relatively invasive way to record the VP port images is to insert a fiberoptic endoscope into the nasal cavity (e.g. Karnell *et al.*, 1988). A tiny video camera is often mounted at the probing end of the endoscope, so that the motion images of the VP port can be captured. The endoscopic technique is widely used by clinicians to study the physiological structure in the nasal cavity and to make qualitative diagnoses of VP movements during speech. However, it is difficult to make quantitative measurements, because the images obtained in this way may be unstable and distorted.

      ii. The radiographic technique has been used in different ways. It can provide detailed static two-dimensional images of the vocal tract (e.g. Fant, 1960). It can also be used in computed tomography (CT) scans, from which a three-dimensional structure of the vocal tract is constructed. In order to record the dynamic movement, the X-ray images can be projected onto a fluoroscope and then filmed (cinefluorography) or video-taped (videofluorography). By comparing velum positions in a sequence of frames, the time-varying movement can be studied (e.g. Moll and Daniloff, 1971). Non-invasive as it is, the radiographic technique has a potential health risk to subjects. This risk prevents it from being widely used in both recent and future speech research.

      iii. The MRI technique is another non-invasive way to obtain multi-dimensional images of structures inside the vocal tract. It does not cause the hazardous ionizing radiation that the radiographic technique does, while providing excellent discrimination of soft tissues. It has been used to measure the

shapes of pharyngeal and oral tracts during sustained vowels (e.g. Baer *et al.*, 1991; Moore, 1992), and the structure of the nasal and para-nasal cavities during sustained nasal consonants (e.g. Dang and Honda, 1994b). Recently, MRI data were also used to construct a three-dimensional articulatory model of the velum (Serrurier and Badin, 2005). Due to its safety and non-invasiveness, the MRI technique is expected to be used more in future speech research. However, MRI units currently used in medicine are expensive (million-dollar per unit with several hundred thousand dollars per year for maintenance). The lying position of subjects may cause some gravity effects on the soft tissues. Loud noises and vibrations during the scan have to be canceled out when acoustic data are needed in parallel studies. The conventional MRI scan is not fast enough to obtain dynamic information of speech. These limitations are to be overcome before the MRI technique can be used widely in speech research. Recent improvements have increased the acquisition rate to 8-9 images per second (Narayanan *et al.*, 2004). It is expected that the research on the VP function will benefit more from further development of the MRI technique.

iv. The ultrasound imaging technique is another non-invasive way to observe the hidden body tissues. It has been used to study the tongue shapes during speech (e.g. Stone *et al.*, 1988). Because the ultrasound waves cannot pass through a tissue-air boundary, it is hard to obtain images of the velum. Therefore, there are few studies using ultrasound images to observe VP movements.

(b) The tracking techniques include point tracking and aperture tracking. Point tracking techniques are designed to monitor the positions of one or more points on articulators, such as the lips, the tongue, and the velum. Aperture tracking techniques are designed to measure the relative opening of the VP port. These tracking techniques usually need less processing effort than imaging techniques. They are often used to extract dynamic information of the VP port. Various principles have been adopted in the design of these techniques. Examples

include the following,

i. The velotrace is a mechanical point tracking device (e.g. Horiguchi and Bell-Berti, 1987). A lever is inserted through the nasal cavity and placed on the upper surface of the velum. The movement of the velum is transmitted by a set of levers that are connected with the internal one. Stain gauges were used in another mechanical method (e.g. Moller *et al.*, 1971). The movement of the velum is transduced by a spring that is attached to the lower surface of the velum inside the oral tract. The placements of these specially designed mechanical sensors are all too invasive for them to be widely used in engineering applications.

ii. Less invasive techniques of point tracking make use of X-ray microbeams (e.g. Fujimura *et al.*, 1977) or magnetometry (e.g. Perkell *et al.*, 1992). Pellets or transducer coils are attached to the velum to track its movements. Besides the possible health risk, the X-ray microbeam systems are costly and only a few research sites have them. On the other hand, there is no known health risk to use the magnetometry technique; it is currently a promising one to study the dynamics and coordination of most articulators. The magnetometry system is also costly and bulky. It is better for research purposes than for daily applications. Also, there are usually some difficulties in gluing a transducer coil on the velum when studying VP movements.

iii. The photodetection technique can be used in aperture tracking systems. For example, a device called the nasograph has been used to measure the intensity of light transmitted through the VP port during speech (Ohala, 1971). It was made of a light source and a light detector that are placed at different sides of the VP port. The relative opening of the VP port is measured from the intensity of light passing through it. Due to its invasiveness, this technique is also rarely used nowadays.

3. VP movements result in both aerodynamic and acoustic effects. In studies of phonetic science, the *aerodynamic effect* usually refers to the static or slowly varying air

pressure or velocity of the airflow that goes through the vocal tract during speech; the *acoustic effect* refers to the rapid vibrations of air pressure as waves that propagate from the sound source to a receiver through the medium of air. Aerodynamic and acoustic measurements are both indirect methods when studying the VP function. They are important measurements because air is the most common medium for speech communication. Moreover, the devices for such measurements are relatively inexpensive to construct and easy to obtain, and the usage of these devices does not cause any health risks nor too much discomfort to subjects. They are ideal for studies that need large-scale data collection and portable applications. However, because the measurements are indirect, it is important to understand the relationship between "hidden" movements and physical observations of the aerodynamic or acoustic signals. Theories or models are usually needed to make such connections. The following briefly introduces instruments that have been built to make these measurements.

(a) Aerodynamic signals are usually acquired by pressure or airflow sensors. Both static structure and dynamic changes of the VP port can be reflected in aerodynamic measurements. Under a static condition, the area of the VP opening can be estimated by simultaneously measuring the nasal airflow and the intra-oral pressure (Warren, 1964b,a), in which the physical relationship between pressure and flow is adopted to calculate the resistance of the orifice and then to estimate its dimensions. In order to study the dynamic changes of the airflow from the nostrils, a pair of differential pressure transducers (e.g. Warren, 1967) or a hot-wire anemometer (e.g. Hutters and Brrndsted, 1992) can be built into a mask that covers the nose. Then the average amount of air or the velocity of the airflow can be measured. The mask conducts all the airflow from the nose to a passage with sensors inside, but it may sometimes interfere with articulation during speech.

(b) Acoustic signals are acquired by microphones. Conventional microphones transduce vibrations of air pressure into electrical signals. Some special type of microphones, such as contact microphones (or accelerometers), that pick up vibrations of solid surfaces instead, can be used to measure tissue vibrations. Since microphones are the most developed devices that can accurately and reliably measure the fast changes of sound pressure, they have been used extensively in speech research and daily applications. However, because the acoustic signal acquired by a microphone is a mixing effect of all the sources and articulators, specific analysis techniques are needed for extracting nasalization information from acoustic signals.

Among all these available techniques, each single measurement may answer a certain aspect of the research questions concerning the VP function. Due to the accessibility of resources and our targeted applications (e.g. assistive devices), we are mainly interested in the indirect approaches to nasalization measurement. The acoustic analysis is the key focus of our research, because any modifications of speech have to be applied to acoustic signals and the final natural communicative medium also has to be acoustic. In addition, aerodynamic techniques are also explored in our studies as supplemental information sources to acoustic analysis. The following sections will present a detailed review of the methodologies of acoustic analysis of nasalization in both normal and disordered speech.

2.2 Analysis of nasalization in normal speech

Research on the acoustic characteristics of nasalization in normal speech spreads in different areas of speech science. In the area of speech synthesis or simulation, researchers have presented production models for nasal consonants and nasalized vowels. In the area of speech signal analysis, various models and algorithms have been developed to estimate nasal tract parameters from acoustic signals. In the area of automatic speech recognition, the search for acoustic correlates of nasalization is one of the important aspects for building an automatic system that imitates the recognition process of humans. This section is organized according to these three related areas.

2.2.1 Production models

The acoustic models of speech production can be either parametric or articulatory by nature. Most parametric models stem from the *source-filter theory* (Fant, 1960), in which the sound pressure at the lips is the result of one or more excitation signals (sources) modulated by the vocal resonator (filter). For example, the source of a vowel sound is modeled as a quasi-periodic volume velocity signal passing through the vocal folds at a certain fundamental frequency; the relatively stable vocal tract for the vowel serves as a short-time invariant linear filter; and the radiation characteristic is modeled as another linear filter. Therefore, the frequency-domain representation of the sound pressure, $S(f)$, is the product of a source function, $G(f)$, a vocal tract filter, $T(f)$, and a radiation function, $R(f)$, that is

$$S(f) = G(f) \cdot T(f) \cdot R(f). \tag{2.1}$$

The configuration of the vocal tract determines the transfer function of the vocal tract filter, $T(f)$. The vocal tract configuration of a vowel can be simplified as a non-branched acoustic tube (assuming the nasal tract is completely isolated by the velum), whose property is characterized by a series of resonators. Each resonator can be parametrized by a second-order band-pass filter, whose natural frequency is called the *formant frequency* and the energy-loss property is characterized by the *formant bandwidth*. Thus, the transfer function of the vocal tract filter is fully parametrized by a series of complex *poles* corresponding to the formants. When the source-filter model is applied to nasal sounds, the pharyngeal and nasal tracts are considered as a tube, and the oral cavity is modeled as a close-ended tube branch. The vocal tract filter of this type of configuration has not only poles (or formants) but also *zeros* (or *anti-formants*) that are modeled as band-stop filters. The source-filter model of a nasalized vowel is even more complex, not only because the output sound pressure is a mixture of outputs from both the mouth and the nose, but also because the oral and nasal cavities work as an open-ended branched tract to each other. In a parametric synthesis system, the pole-zero pattern of a certain phoneme can be empirically found through the trial-and-error approach (e.g. Allen *et al.*, 1987). Using a parametric synthesizer whose pole-zero patterns can be specified, researchers have tried to

explore the acoustic and perceptual correlates of nasalization (e.g. Hawkins and Stevens, 1985).

Articulatory models of speech production are derived from the equations that govern acoustic waves propagating in a tube (Flanagan, 1972; Rabiner and Shafer, 1978). When the acoustic wave in the vocal tract is assumed to be a linear plane wave (for frequencies below 4 kHz), and the energy loss is neglected, the physical properties of the wave can be approximately represented by the following equations,

$$\begin{aligned} -\frac{\partial U(x,t)}{\partial x} &= \frac{1}{\rho c^2}\frac{\partial\left(P(x,t)\ A(x,t)\right)}{\partial t} + \frac{\partial A(x,t)}{\partial t} \\ -\frac{\partial P(x,t)}{\partial x} &= \rho\frac{\partial\left(U(x,t)/A(x,t)\right)}{\partial t}, \end{aligned} \tag{2.2}$$

where $P(x,t)$ is the sound pressure at position x in the vocal tract at time t, $U(x,t)$ is the volume velocity function, $A(x,t)$ is the area function of cross sections of the vocal tract, ρ is the density of air, and c is the velocity of the sound wave. When the area function is considered stable, i.e. $\frac{\partial A(x,t)}{\partial t} = 0$, within a short period of time, the wave equations are simplified as the following,

$$\begin{aligned} -\frac{\partial U(x,t)}{\partial x} &= \frac{A(x)}{\rho c^2}\frac{\partial P(x,t)}{\partial t} \\ -\frac{\partial P(x,t)}{\partial x} &= \frac{\rho}{A(x)}\frac{\partial U(x,t)}{\partial t}. \end{aligned} \tag{2.3}$$

One approach to solve these equations is through the Kelly-Lochbaum (K-L) method (Kelly and Lochbaum, 1962; Rabiner and Shafer, 1978). When the vocal tract is further simplified to be a series of short lossless uniform sections of equal-length tubes, $A(x)$ becomes a piecewise step function. For the i-th section of tubes with a constant area of A_i, there is an analytical solution to the wave equations (2.3),

$$\begin{aligned} U_i(x,t) &= U_i^+\left(t-\frac{x}{c}\right) - U_i^-\left(t+\frac{x}{c}\right) \\ P_i(x,t) &= \frac{\rho c}{A_i}\left[U_i^+\left(t-\frac{x}{c}\right) + U_i^-\left(t+\frac{x}{c}\right)\right], \end{aligned} \tag{2.4}$$

where $U_i^+\left(t-\frac{x}{c}\right)$ and $U_i^-\left(t+\frac{x}{c}\right)$ denote forward (transmitted) and backward (reflected) traveling wave functions, respectively. At the junction of two sections with different cross-sectional areas, the volume velocity and the sound pressure must be continuous in both

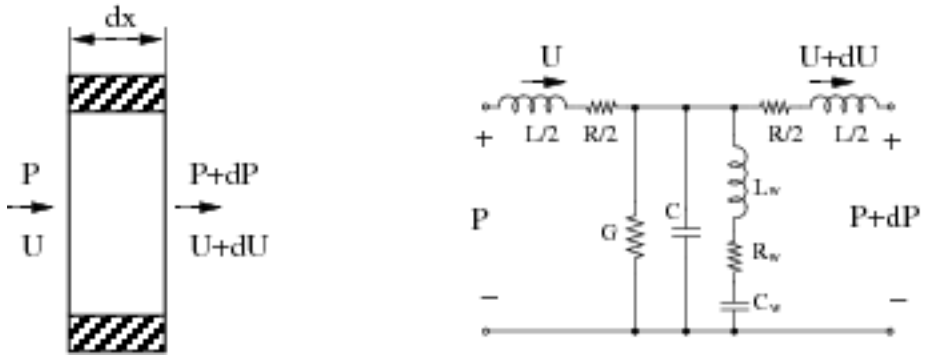


Figure 2.1: Transmission line model of a lossy cylindrical acoustic tube.

time and space. The special form of the forward and backward functions indicates that a traveling distance of the transmitted or reflected wave in space is equivalent to a delay in time. Therefore, the traveling wave in a series of uniform lossless tubes can be represented by a discrete time system that is characterized by a set of reflection coefficients and propagation delays. To simulate the nasal tract, a branch of tubes can be added at the junction of the pharyngeal, oral, and nasal tracts (i.e. the VP port). A special continuity condition of the sound pressure and the volume velocity at the junction can be modeled as a three-port-adaptor (Lim and Lee, 1993, 1996; Liu and Lacroix, 1996), which enables us to calculate the transfer function of the linear discrete system.

Another approach to simulate the acoustic wave propagating in short uniform tubes is the transmission-line method (Flanagan, 1972). The acoustic wave equations (2.3) have the same structure as the equations of an electrical wave that travels through a lossless transmission line, in which the sound pressure and the volume velocity are analogous to voltage and current, and the parameters $\frac{A(x)}{\rho c^2}$ and $\frac{\rho}{A(x)}$ are analogous to the capacitance (C) and the inductance (L) per unit length of the transmission line, respectively. When the vocal tract is modeled as a series of short uniform tubes that are connected together, the transmission properties of each section of the tubes can be represented by a circuit unit (see Figure 2.1). The additional resistance (R) and conductance (G) in Figure 2.1 are analogous to the viscous friction loss and the heat loss, respectively. The shunt impedance (R_w, L_w, and C_w) in the figure represents the yielding wall effect of the tube (Childers, 2000). These extra parameters extend the model capacity to represent lossy tubes. All the

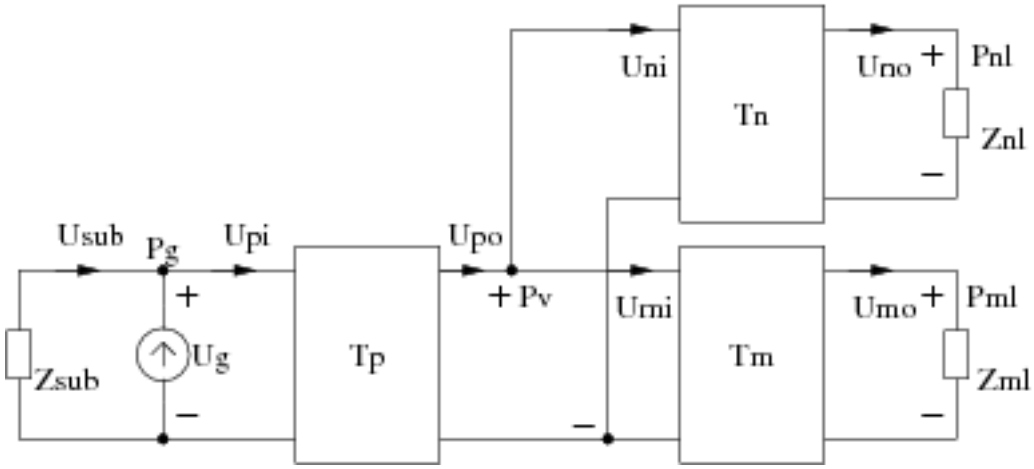


Figure 2.2: Circuit network representation of the production of nasalized voiced sounds.

parameters in such a unit are determined by the dimensions of the vocal tract sections, the physical properties of air, and the physical properties of the vocal tract walls. Given an area function of the vocal tract, along with a proper model of the excitation signal and the radiation resistance, a serial connection of such units can simulate the production of a certain vowel (Maeda, 1982a). To simulate nasal consonants or nasal vowels, the transmission-line model can be extended by adding a branched circuit network that represents the nasal tract (see Figure 2.2). Due to the complicated structure of the nasal cavity (e.g. asymmetric branches, irregular internal surfaces, and various sinuses), it is difficult to directly convert the physiological data into this representation. An empirically simplified area function can be specified, and the sinuses can be modeled as shunt resonators (Maeda, 1982b).

Since the articulatory production models tend to set up a relatively explicit relation between the articulatory configuration and the acoustic signal, such a simulation system is a very useful tool to test hypotheses of the relation between the production and perception of nasalized sounds (e.g. Maeda, 1993; Pruthi and Espy-Wilson, 2005). The methodology of using either a parametric or an articulatory synthesizer to study the acoustic properties of nasalization belongs to analysis-by-synthesis approaches.

2.2.2 Analysis of acoustic signals

The field of acoustic speech signal analysis has been developing for decades. The key purpose of analysis is to extract useful information from acoustic speech signals. What types of information are extracted depends on applications, such as coding, modification, enhancement, and speaker/speech/language recognition (Quatieri, 2002). Some of the analysis methods are closely related to speech production models, while others emphasize more on certain mathematical merits for easy manipulations. Among these methods, we are interested in those that are potentially related to the nasalization process.

The linear prediction (LP) analysis technique has been prevalent in the field of speech analysis, based on the source-filter theory of speech production (Fant, 1960). In LP analysis, the spectral effects of the source volume velocity, the vocal tract, and the radiation are combined and represented by an all-pole (or autoregression, AR) filter, and the excitation is either a sequence of impulses (for voiced phones), a single impulse (for plosives), white noise (for unvoiced fricatives), or a combination of them. The LP model can be represented in the z-domain as follows,

$$\begin{aligned} S(z) &= E(z)\,\frac{1}{A(z)} \\ &= E(z)\,\frac{1}{1-\sum_{i=1}^{N} a_i\, z^{-i}}, \end{aligned} \tag{2.5}$$

where $E(z)$ is the z-transform of an impulse sequence or a zero-mean white noise, and $A(z)$, a polynomial of order N, is called the *inverse filter*. In the discrete time domain, the model is equivalent to

$$\begin{aligned} e[n] &= s[n] - \sum_{i=1}^{N} a_i\, s[n-1] \\ &= s[n] - \hat{s}[n], \end{aligned} \tag{2.6}$$

where $s[n]$ and $e[n]$ are the speech signal and the excitation signal, respectively, and $\hat{s}[n]$ is defined as the predicted value of $s[n]$ using the previous N samples. Thus, $e[n]$ can also be interpreted as the error between the signal and the predicted value. The importance of LP analysis is that the coefficients of the all-pole filter can be estimated efficiently from a short segment of the speech signal by applying a least-squares criterion

to the error. These AR coefficients can be computed in a variety of ways (Markel and Gray, 1976). Among them, the partial-autocorrelation-reflection-coefficient (PARCOR) formulation (Itakura and Saito, 1971) reveals a lattice structure of the inverse filter. This lattice structure is related to the lossless uniform tube (K-L) model of speech production. When the effect of the glottal source and the lip radiation is canceled out by a pre-emphasis filtering process, the area function of the vocal tract can be approximately estimated by Levinson-Durbin's recursive solution to the Yule-Walker function (e.g. Wakita, 1972, 1973b,a) or by Burg's method (Gray *et al.*, 1977). Another approach to estimate the tube parameters is to extend the lattice structure by incorporating the glottal model and lip load, and then to optimize the reflection coefficients iteratively (Schnell and Lacroix, 2001a). Since the relationship between the articulatory configuration and the acoustic signal can be a many-to-one mapping, the resulting lattice parameters do not necessarily correspond to the real vocal tract structure.

The all-pole model is widely used in speech analysis because 1) the poles in a filter are appropriate to represent the spectral prominences in different groups of phonemes including vowels, plosives, and fricatives; 2) the spectral tilt caused by the glottal source and the lip radiation can be approximately canceled out by pre-emphasis; 3) each zero in the filter can be approximated by a series of poles; 4) the solution to parameter estimation is linear and mathematically tractable. Therefore, with a sufficient order, an all-pole model is a practical approximation of speech signals. However, since zeros exist in speech signals under most conditions, such as nasals, nasalized vowels, liquids with a branched oral tract, fricatives or plosives with a back cavity before their noise sources, it is intuitively more plausible to model speech signals with a pole-zero (or autoregression moving average, ARMA) model,

$$\begin{aligned} S(z) &= E(z)\,\frac{B(z)}{A(z)} \\ &= E(z)\,\frac{\sum_{j=0}^{M} b_j\, z^{-j}}{1-\sum_{i=1}^{N} a_i\, z^{-i}}, \end{aligned} \tag{2.7}$$

in which zeros are explicitly modeled by the numerator polynomial, $B(z)$.

Unlike the multitude of methods available for AR estimation, no comparable general algorithms exist for ARMA estimation currently. Directly applying the least-squares criterion to the error between the model and the signal leads to a set of non-linear functions.

Iterative searching approaches to the solution of these non-linear functions are computationally complex (Kay, 1988). Some solutions are based on the linearization of the problem. For example, one group of approaches is to estimate the denominator parameters first with the covariance LP analysis by assuming the impulse response of the ARMA filter is known, then the numerator parameters are estimated by solving another set of linear equations. The estimation procedure can be performed in either a two-stage or an iterative way (e.g. Kopec *et al.*, 1977; Steiglitz, 1978). This group of approaches require a preliminary deconvolution of the speech signal. Possible methods include pitch-synchronous processing only on the glottal closure portions of speech signals, or the homomorphic deconvolution processing. Another method of linearization is to convert the problem into a system identification problem by assuming that the excitation signal of the ARMA filter can be modeled and estimated, then all the parameters are estimated by solving the linear equations derived from $S(z)A(z) - E(z)B(z) = 0$ (e.g. Konvalinka and Matausek, 1979; Fujisaki and Ljungqvist, 1987). A third method of linearization is to estimate a high-order all-pole model first, then to decompose the resulting model into a pole predictor and a zero predictor (Song and Un, 1983).

Most of the ARMA algorithms above are "signal-based", that is, they are designed to accomplish a better spectral match to the speech signal than the AR model can achieve. However, there are no explicit constraints that connect these acoustic parameters with the physiological configurations of the vocal tract. On the other hand, there have been a few "model-based" approaches that focused on estimating the ARMA transfer function from the branched tube model of nasals or nasalized sounds either from single-channel acoustic data (Lim and Lee, 1993, 1996; Liu and Lacroix, 1996; Schnell and Lacroix, 2002) or from dual-channel acoustic data (Schnell and Lacroix, 2001b). The validity of each "model-based" approach was usually examined by the merit of spectral match as well, because it is difficult to obtain the real vocal tract configuration data.

2.2.3 Spectral features of vowel nasalization

Because the vocal tract is an inter-dependent acoustic system, it is difficult to find an invariant acoustic feature of nasalization that is independent of other parts of the vocal

tract. Moreover, because of individual variations in vocal tract anatomy, the spectral representation of nasalization may also differ from one speaker to another. However, since nasalization is a perceivable event in speech, researchers have always been expecting to find acoustic correlates to it. These correlates may be useful to improve the performance of an automatic speech recognition system, for example. Below is a summary of some acoustic features that have been found to be associated with nasalization.

Compared with its non-nasalized counterparts, a nasalized vowel usually has a reduced amplitude (or increased bandwidth) of the first formant, and the frequency of its first formant shifts upward (House and Stevens, 1956; Fant, 1960). The introduction of a pole-zero pair in the region of the first formant (700-2000 Hz) may result in a secondary spectral peak, which can be another feature of nasalized vowels (House and Stevens, 1956; Hawkins and Stevens, 1985). Due to the paranasal sinuses, an extra nasal resonance or pole-zero pair may occur between 200 and 500 Hz during vowel nasalization (Lindqvist-Gauffin and Sundberg, 1976; Maeda, 1982b). There are other spectral features observed by different researchers, including changes in the amplitude of the second and/or third formant and shifts in their frequencies, an extra pole-zero pair in the third formant region, and the reduction in overall amplitude of the vowel, although these higher frequency effects are considered not very stable across observations.

While these spectral observations are qualitative in nature, researchers have also tried to formulate some quantitative features to measure or detect nasalization. Huffman (1990) proposed to use the amplitude difference between the first formant ($A1$) and the first harmonic ($H1$) as a correlate to the perception of nasality. The idea was to capture the relative amplitude decrease of the first formant. Maeda (1993) introduced another measure of the spectral "flatness" in the low frequency region as a feature related to the perception of nasalized vowels. Two spectral peaks in the low frequency range of the spectrum were chosen, and the amplitude difference between them was calculated. They both found the degree of spectral flatness is highly correlated to the perception of nasality. However, a drawback of both measures is that the spectral peaks have to be manually chosen by visual inspection of the spectrum. Chen (1995; 1997) proposed to use two amplitude differences, $A1-P1$ and $A1-P0$, for quantifying nasalization, where $A1$ was the amplitude of the first

formant in dB, $P1$ was the amplitude of the spectral peak between the first two formants, and $P0$ was the amplitude of the spectral peak below the first formant. The first parameter was to capture the amplitude decrease or the bandwidth increase of the first formant; the second parameter was to capture the low frequency peak due to sinuses. In practice, the values of $P1$ and $P0$ were estimated at peak harmonics around 950 Hz and within low frequencies, respectively.

These spectral features of nasalization may be used in an automatic speech recognition system, not only because there are nasal vowels in some languages, but also because an accurate detection of nasalized vowels may improve the recognition rate of their neighboring nasals. Glass and Zue (1985) developed a set of acoustic parameters for automatic distinction of non-nasal and nasal vowels. The parameters include 1) the center of gravity in the 0-1000 Hz region, 2) the standard deviation of the center of gravity, 3) the maximum and minimum percentage of time there is an extra resonance in the low frequency region, 4) the maximum value of the average dip between the first and the extra resonance, and 5) the minimum value of the average difference between the first and the extra resonance. They obtained a 74% classification accuracy rate on their test set. In Chen's work (2000), the two amplitude distance features of nasalization ($A1 - P1$ and $A1 - P0$) described above (Chen, 1997) were adopted to improve the recognition of nasals in a knowledge-based speech recognition system. In their experiment, when nasal murmurs were missing or the nasals were syllabic, the recognition of nasals was improved by the detection of vowel nasalization. In the report of a JHU workshop (Hasegawa-Johnson *et al.*, 2005), a combination of features were used to build two versions of nasalization detectors. One of them was vowel independent, and the other was vowel dependent. The features included Mel-Frequency Cepstral Coefficients (MFCCs), phonological distinctive features, rate-scale auditory cortical parameters and formant parameters. Support vector machines (SVMs) were trained as the classifiers. In their test, vowels in the nasal context was regarded as nasalized. The vowel-dependent classifier achieved the following accuracy rates on four vowels: /ey/ (81% accuracy, in a test set with 50% nasalized vowels), /iy/ (76%), /ae/

(75%), and /ao/ (73%) [1].

2.3 Analysis approaches to resonance disorders

"Resonance disorder" refers to either excessive or insufficient nasal resonance when producing vowels, voiced consonants or both, therefore resulting in speech that is perceived as hypernasalized or hyponasalized. Because resonance disorder could be a symptom of many groups of speech disorders (e.g. cleft palate speech, motor speech disorders, hearing-impaired speech, and so on), the assessment of nasality is an important aspect during the diagnosis of these speech disorders. Due to the poor reliability of perceptual judgments, it is of great interest to develop objective measures as an informative supplement in the diagnostic process. The existing approaches to the objective measurement of nasality can be grouped into two categories according to their methodologies. In one category, like in the majority of normal speech analysis, the relevant information is extracted only from the acoustic signal picked up by a conventional microphone; while in the other category, more complicated instrumental techniques are used in order to acquire more reliable information.

2.3.1 Single acoustic channel analysis

Kataoka (1988) originally proposed a one-third-octave measure to quantify hypernasality by analyzing acoustic signals directly. This measure was obtained by filtering the acoustic signal of a sustained vowel /iy/ with a set of band-pass filters whose central frequencies are located at one-third-octave intervals. The amplitude of each band was calculated and averaged across the steady portion of the vowel to obtain an averaged one-third-octave spectrum. In their later studies, Kataoka *et al.* (1996, 2001) adopted this measure in the analysis of speech from adults and children with cleft palate or velopharyngeal incompetence. It was found that the spectral characteristics of vowel segments perceived as hypernasalized were consistent with the spectral features of vowel nasalization (see Section 2.2.3). The multiple regression analysis in their experiments showed a high correlation between the amplitudes of certain one-third-octave bands and the perceptual rating scores

[1] In this thesis, all phonemes are denoted in OGIbet (see Appendix A)

of nasality. The resulting equation could be used to predict hypernasality scores from the acoustic parameters in a clinical setting, for the purpose of speech evaluation of patients with resonance disorders resulting from physiological deficits such as cleft palate. The one-third-octave measure is very similar to other frequency-band-energy analysis such as the MFCC feature that is widely used in speech recognition applications. The choice of a single sustained vowel (/iy/) in the evaluation procedure helps to limit the variations of the feature as a correlate to the perception of nasality. There was still no evidence about whether this analysis could be extended to the sentence level, where other sonorants might be nasalized as well.

Cairns *et al.* (1994; 1996a; 1996b) proposed a method of using the Teager Energy operator (TEO) to detect hypernasality from the acoustic signals. The TEO was introduced by Kaiser (1990) based on a special definition of energy developed by the Teagers (1985; 1990), in which the total energy needed for a system to generate oscillating motions was calculated from the signal. It has been shown that the Teager energy required to generate a harmonic motion is proportional to the product of the squared amplitude and the squared frequency of the oscillation. Thus, the TEO can be used to separate the time-varying amplitude envelope and the instantaneous frequency of a signal with a single narrow-band frequency component (Maragos *et al.*, 1993). It has also been observed that the Teager Energy profile of a multi-component signal is composed of not only the Teager Energy of each component but also the cross-correlated terms between components. The assumption behind the hypernasality detection approach is that the spectrum of a nasalized speech segment is composed of not only oral formants but also nasal formants and antiformants in the low frequency region, whereas a normal vowel only has the first formant in this region. Therefore, when a low-pass filter with a cutoff frequency just above the first formant is applied, the filtered signal of a nasalized vowel is a multi-frequency-component signal, while that of a normal vowel has only one component. However, a band-pass filter that is centered at the first formant will only retain the first formant component in the signal no matter whether it is nasalized or not. Since the TEO is sensitive to multi-component signals, the shape of the Teager Energy profile of a multi-component signal is very different from that of a single component one. Therefore, the correlation coefficient between the

Teager Energy profiles of low-pass filtered and band-pass filtered signals may be used to discriminate the nasalized and normal vowels. In their research, a likelihood ratio detector based on the correlation coefficient was built to make the decisions. Automatic as it is, the approach has some drawbacks. First, the algorithm relies on a formant tracker to locate the first formant, and a pitch detector to avoid instability of the TEO. These two tasks are both difficult problems in the field of speech analysis. The robustness of the approach may be undermined due to the inaccuracy of either task. Second, it has been observed that the glottal flow excitation signal contains a resonance-like peak, called the glottal formant, in the low frequency region. This spectral influence of the glottal flow may be filtered out by the band-pass filter of the TEO, but retained in the low-pass filtered signal, which possibly results in different Teager Energy profiles even for normal vowels. Third, the back vowels were not tested in the reported experiments because their first and second formants are close to each other and it is difficult to filter out the effect of the second formant on the Teager Energy profiles. This also raises a question about its application across all the sonorants in a sentence.

In a recent study, Vijayalakshmi and Reddy (2005a; 2005b) adopted a modified group-delay method to extract the nasal formant from the acoustic signal, when it exists close to the first formant. The modified group-delay function (Murthy and Gadde, 2003) has a better frequency resolution property than conventional formant extraction techniques. It was used to distinguish the extra formant adjacent to the first formant, which may be introduced due to nasalization. The signals were low-pass filtered with a cutoff frequency of 800 Hz to improve the resolution of the group-delay function. They proposed to use three parameters, including the frequencies of the first two highest peaks in the modified group delay spectrum and the ratio of the group delay of these frequencies, as a feature vector for hypernasality detection. They trained two classifiers with the feature vectors extracted from the speech of cleft palate and normal speakers. The classifier achieved a correct rate of 85% for hypernasal/normal decision of the phonemes /aa/, /iy/ and /uw/.

2.3.2 Multi-channel measurements

It has been acknowledged that separate measurements of oral and nasal sound pressures could have advantages over a single channel of acoustic measurement for the nasalization analysis, because the mixing of nasal and oral acoustic waves may conceal the information that is conveyed by either one separately. More sources of information may help to resolve the inevitable ambiguities in the single-channel acoustic analysis.

Fletcher (1970) first introduced an electronic system named TONAR (The Oral Nasal Acoustic Ratio). It is the ancestor of the Nasometer (by KayPENTAX, Lincoln Park, NJ), which has been used extensively by speech clinicians to measure nasalance scores. The term *nasalance* refers to the ratio of the nasal acoustic energy to the total acoustic energy radiated from the nose and the mouth in total. In the Nasometer system, the oral and nasal sound pressures are transduced by two microphones mounted on the two sides of a sound-separating plate. The two signals are individually amplified and filtered with band-pass filters (350-650 Hz). The filtered signals are then converted into nasal and oral energy signals, $E_N(t)$ and $E_O(t)$, and fed into a computer. Then the nasalance score is calculated as $\frac{E_N(t)}{E_N(t)+E_O(t)}$. In clinical settings, the average nasalance score over certain standard passages, which is called *Nasalance Index*, has often been used as a quantitative measure of nasality for the purposes of patient assessment and management (Dalston *et al.*, 1991; Daltson *et al.*, 1991; Karnell, 1995, etc.). The time course of the nasalance score may provide useful information about velopharyngeal timing and coarticulation (Seaver III and Dalston, 1990).

One drawback of the Nasometer system is that it does not record the original nasal and oral signals for further analysis. This undermines its value as a research tool. Another drawback is the relatively high cost due to its hardware implementation. Recently, an alternate lower-cost system (Awan, 1996), called the NasalView (by Tiger Electronics Inc., Seattle, WA), has become commercially available. The microphones, sound-separating plate, and headset of the NasalView are similar to those of the Nasometer, but the nasal and oral signals are directly fed to a computer after being amplified. Unlike the Nasometer, it does not band-pass filter the acoustic signals, and the nasalance score is calculated from

the root-mean-square energy of the two signals. Since the hardware only takes care of the data acquisition task, and all the computations are done by the software, it is more flexible and inexpensive. One can simulate the nasalance calculation of the Nasometer by manipulating the raw signals, and can also design and test other sophisticated analyses from them.

Multi-channel measurements may also include measurements other than acoustics in air. For example, another measure of the nasal coupling effect, called the Horii Oral Nasal Coupling (HONC) index (Horii, 1983), is derived from two channels of signals, one measured by an accelerometer attached to the outside of a nostril, another measured by a microphone positioned in front of the mouth. A high correlation of the HONC index with the perceived nasality was found. The advantage of the accelerometer measurement is that it picks up vibrations through tissues instead of airborne pressure waves; while the disadvantage is that the position sensitivity and speaker variations may hinder it from being used in more detailed analyses. Other multi-channel studies include photodetection and aerodynamic measurements that were usually conducted with parallel recordings of acoustic data, so that the information is supplementary across channels.

Chapter 3

Dual-channel measurements of nasalization

The acoustic signal resulting from nasalization is complicated because 1) the resonant cavities (pharyngeal, nasal and oral) are inter-dependent on one another, 2) the anatomic structures of the nasal tract and the paranasal sinuses are irregular and vary among speakers, and are therefore difficult to model, 3) the sound pressure measurement is actually a mixture of radiation from the lips, nostrils and soft tissues, and 4) the inverse mapping from the acoustical signal to the articulatory configuration is one-to-many. When reliable information about the VP status is the goal of the analysis, more sources of measurement are often helpful. The studies presented in this chapter investigate the feasibility of extra channels of non-invasive measurements in addition to conventional acoustic signals. The following sections will first present the studies of an additional aerodynamic channel, then an additional acoustic channel. The advantages and limitations will be discussed at the end.

3.1 Additional aerodynamic channel[1]

Aerodynamics in speech refers to the study of the moving airstream during speech production. Though acoustics in the air is actually a branch of aerodynamics in this sense, people usually prefer to call the non-acoustic part of the study "aerodynamics" in the speech field. We will also follow this convention in the context of this thesis.

While acoustic studies mainly concern the propagation of sound waves, in which the

[1] **Parts of the research in this section have been published in a conference paper (Niu *et al.*, 2006).**

air is assumed to be steady but compressible, aerodynamic studies in speech mainly concern the generation of sounds from the airstream. Aerodynamic measurements include the measures of air pressures and airflow velocities; and aerodynamic analyses study the relationships between these measurements and the physical structures of sound sources (Shade, 1999). The idea of parallel studies of the aerodynamic and acoustic signals of speech is not new. Most of the studies emphasized the extraction and analysis of aerodynamic properties, and used the acoustic signals as a reference (e.g. Warren, 1964b,a, 1967; Hutters and Brrndsted, 1992). Due to the complexity of the devices needed for accurate aerodynamic measurements from speech (see Section 2.1), the corresponding acoustic signals may be distorted, or even worse, the devices for aerodynamic measurement can be intrusive or directly interfere with acoustic signals. The distortion may undermine the analysis and further processing of the acoustic signal itself. Our studies explore an opposite direction that emphasizes the analysis of acoustic signals and using the aerodynamic measurement as an extra information source for nasalization. We first need a way to simultaneously record both nasal airflow and acoustic signals without interference with articulation and degradation of the acoustic signal; then we can investigate their relationships. In this section, the design of such a device will be first described; then the specific signal processing algorithms will be presented; and at last, the usage of such a device in nasalization analysis will be discussed with sample data from the experiments.

3.1.1 Device design

3.1.1.1 Measurement principles

The purpose of designing such a device is to monitor the nasal airflow while simultaneously recording the acoustic speech signal. In order to avoid interference with articulation and distortion of the acoustic signal, we choose not to use nasal masks that were often used by other researchers to hold the transducer and seal the airflow from the nose. Instead, the velocity of the open airflow out of a nostril will be picked up during speech.

Airflow velocity can be measured indirectly according to the principle of the Bernoulli effect. Given a tube that is open at one end and closed at another, the air pressure at the closing end, P_t, is related to the velocity of the airflow in which the open end is placed and

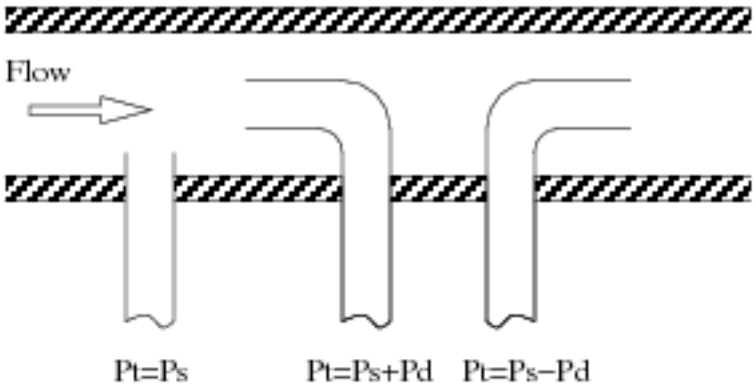


Figure 3.1: Principle of the airflow velocity measurement. Bernoulli's equations are applied under three conditions with respect to different orientations of the open end of the tube related to the airflow direction.

its relative orientation to the direction of the flow (see Figure 3.1). Assuming that air is compressible, when the open end of the tube is perpendicular to the flow direction, P_t is equal to the static pressure of the air, P_s; when the open end faces the oncoming airflow, P_t is the summation of the static pressure P_s and the dynamic pressure P_d; when the air flows in the opposite direction of the opening, P_t is the difference between P_s and P_d. The dynamic pressure P_d is proportional to the square of the absolute velocity of the air, so the later two conditions of the tube above can be used in the measurement of airflow velocity according to Bernoulli's equations,

$$P_t = P_s \pm P_d = P_s \pm \frac{1}{2}\rho V^2, \tag{3.1}$$

where ρ is the density of air, and V is its velocity.

In our design, a small segment of an L-shaped glass tube is to be placed 5 mm under a nostril of the speaker, with one open end pointing up to the nostril. The other end of the tube is connected to one of the two air-intakes of a differential pressure sensor through a 25-cm-long surgical tube. The differential pressure sensor can transduce the pressure differences between its two air-intakes into voltage signals. With its one air-intake connected to the tube and the other open to the static air, it is used to measure the pressure difference, $P_t - P_s$. According to Equation 3.1, the sign of this measure represents the direction of the airflow, and its absolute value is the dynamic pressure that is proportional to the square of the velocity of the airflow from the nostril.

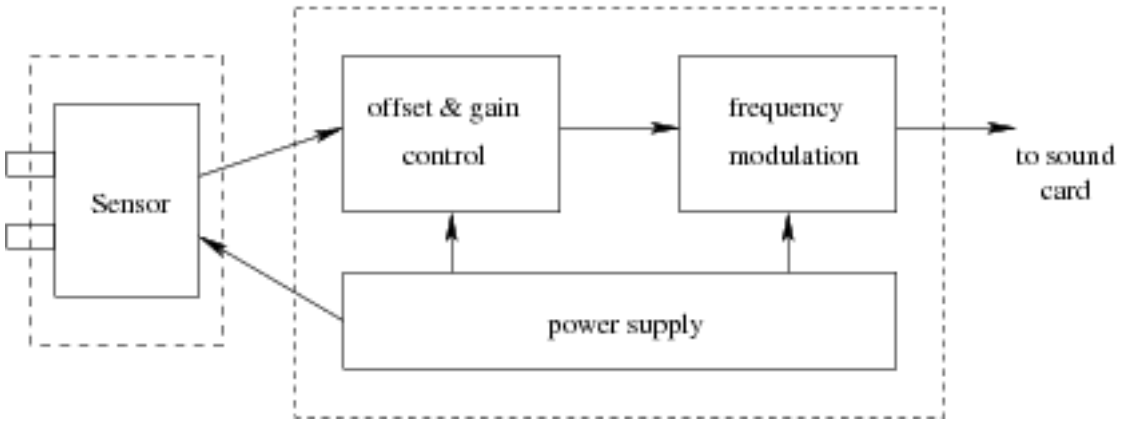


Figure 3.2: Structural design of the measuring device. The sensor has two air intakes to measure differential pressures. Arrows indicate the flow of signals between modules.

3.1.1.2 Data acquisition

The usage of personal computers (PC) has made it inexpensive and easy to acquire acoustic signals of speech through sound cards. A generic sound card in a computer usually uses AC coupling for the input signal, thus filtering out DC and low-frequency information of the signal. However, the DC and low-frequency components in airflow signals are important, so a sound card can not be used directly for the acquisition of airflow signals. Due to the high cost of using a special data acquisition card, we design a pre-processing module in the device, by which the airflow signal from the differential pressure sensor is frequency modulated (FM) with a carrier frequency within the audio frequency range. Then the FM signal can be recorded through one input channel of the sound card, while the acoustic signal is recorded through another input channel simultaneously. The recorded FM signal, once captured, can be demodulated to recover the original airflow signal.

Figure 3.2 shows the modular structure of the measuring device, and the detailed schematic plot of the design is shown in Figure 3.3. A low pressure sensor (1-MBAR-D-4V, All Sensors, CA) is encapsulated in a small plastic box that can be attached to a microphone headset. The operating pressure range of the sensor is ± 1 mbar. The sensor is connected to a processing box that contains the power supply, offset and gain control, and frequency modulation modules. The power supply module includes batteries and regulators. It supplies stable voltage for the sensor and other components. The module

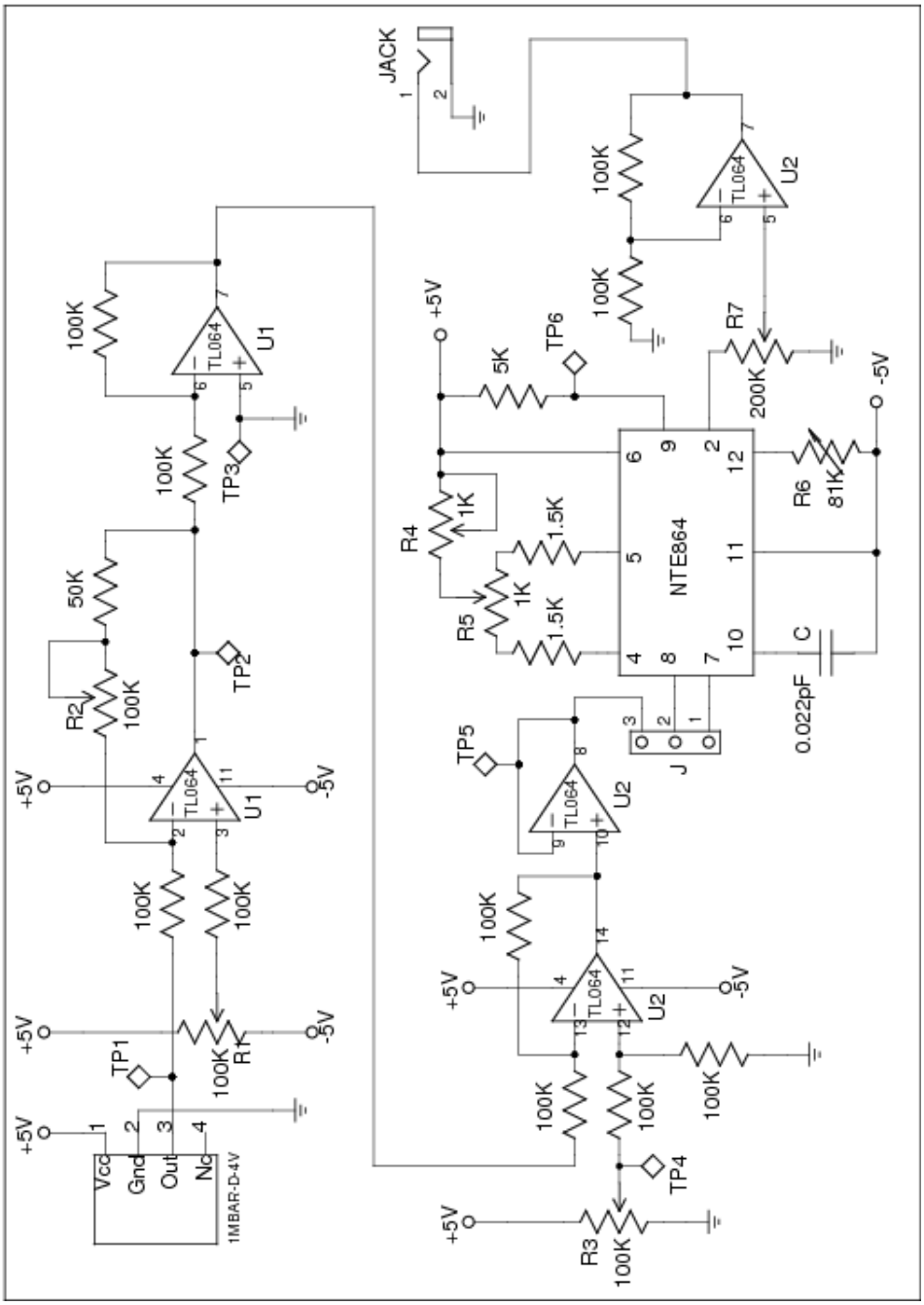
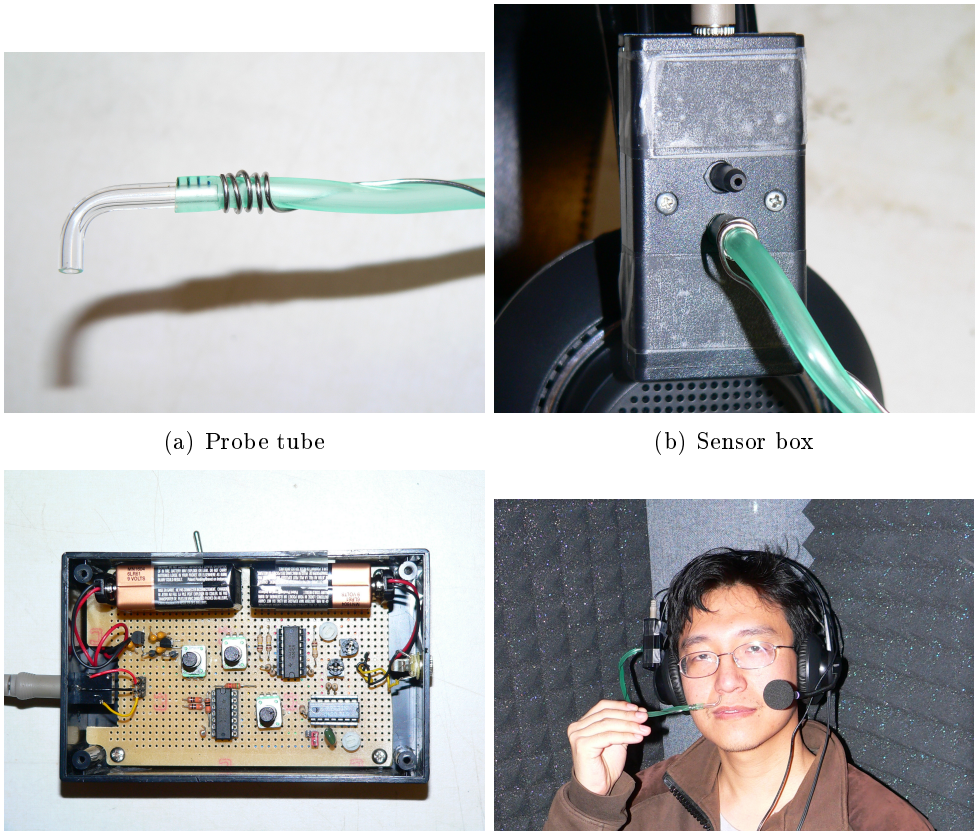


Figure 3.3: Detailed schematic of the measuring device. It includes the pressure sensor, the offset and gain module, and the frequency-modulation module. The power unit is not included.



(a) Probe tube

(b) Sensor box

(c) Processing box with circuit in it

(d) Usage during recording

Figure 3.4: Assembly and usage of the device as an additional measuring channel to the acoustic channel. Sub-figures include the probe tube, the sensor that is encapsulated in a plastic box attached on a headset microphone, and the signal processing box with circuits in it. A speaker is showing the usage of the device in a recording booth.

for offset and gain control simply includes op-amps and potentiometers. They are tuned to convert the output signal of the sensor within the operating range of the frequency modulation module. A waveform generator chip (NTE-864, NTE Electronics, INC., NJ) is used as a voltage-controlled oscillator (VCO). Several resistors and one capacitor are chosen and tuned in order to set the highest frequency of the VCO to about 5 kHz. The sine wave output of the VCO is then amplified to standard audio device line levels. This signal is sent to the line input of the sound card of a PC. As a last step, the offset of the VCO input is tuned when zero differential pressure is applied to the sensor (i.e. when the sensor is inert), so that the frequency of corresponding output signal is about 3 kHz.

Photos in Figure 3.4 show how the sensor is attached to the headset of a microphone, and how a subject can make multi-channel recordings with the whole setup. It can be seen that the extra sensor only occupies slightly more space, and the processing circuits are contained in a small box that is simply connected between the sensor and the sound card of the computer.

3.1.2 Signal processing algorithms

3.1.2.1 Demodulation

The demodulation algorithm of FM signals is based on the property of analytical signals (Oberg, 2001). An FM signal can be represented in the following form,

$$y(t) = K\cos\left[2\pi f_c t + \phi(t)\right], \tag{3.2}$$

where f_c is the carrier frequency, and K is a constant. The relationships among the phase signal $\phi(t)$, the instantaneous frequency $\omega(t)$, and the original input signal $x(t)$ can be represented by the following equations,

$$\omega(t) = 2\pi f_c + \frac{d}{dt}\phi(t) = 2\pi[f_c + f_d x(t)], \tag{3.3}$$

where f_d is a frequency deviation. When $x(t)$ is a narrow-band signal, the Hilbert transform of $y(t)$ is

$$\hat{y}(t) = K\sin\left[2\pi f_c t + \phi(t)\right]. \tag{3.4}$$

Then an analytical signal of $y(t)$ can be constructed as

$$y_a(t) = y(t) + j\hat{y}(t) = Ke^{j[2\pi f_c t + \phi(t)]}. \tag{3.5}$$

Equation 3.3 and 3.5 indicate that $x(t)$ can be calculated from the derivative of the phase signal of $y_a(t)$. Since all signals collected from the sound card of a computer are sampled, the following will introduce the corresponding demodulation algorithm for discrete signals.

Suppose the FM signal $y(t)$ is sampled at a frequency of F_s ($F_s \gg f_c$), and the resulting discrete signal is $y[n]$. First, its corresponding analytical signal $y_a[n]$ is obtained through the discrete Hilbert transform (DHT),

$$y_a[n] = y[n] + j \cdot DHT\left\{y[n]\right\}, \tag{3.6}$$

Second, given the carrier frequency f_c, an auxiliary signal $y_x[n]$ is obtained by

$$y_x[n] = y_a[n] \cdot e^{-j2\pi f_c n/F_s} = Ke^{j\phi[n]}. \tag{3.7}$$

Third, the phase signal $\phi[n]$ is obtained from the unwrapped angle of the signal $y_x[n]$. Fourth, the derivative of the phase is approximated by differencing operations, which is proportional to the input signal $x[n]$. Because we are only interested in the relative changes of $x[n]$, the constant factor f_d can be ignored. It is to be noted that a similar algorithm has been implemented in the Matlab signal processing toolbox (function "dmod").

3.1.2.2 Zero calibration

Though the carrier frequency, f_c, has been tuned to about 3 kHz during the construction of the device, it can drift due to temperature changes and other environmental conditions. An accurate value of f_c is needed in the demodulation algorithm. It can be calculated from a section of the FM signal, $y_0[n]$, that is recorded under the zero-airflow condition, that is when no airflow applies to the sensor. The frequency of $y_0[n]$ is the actual carrier frequency f_c. The demodulated signal of $y_0[n]$ must be zero when the carrier frequency is exactly the same as the device setting, so we can search an optimal f_c by calculating how the demodulated signal is close to zero. The calibration algorithm is as follows.

First, the demodulation procedure described above can be represented as a transform from the signal $y_0[n]$ to the signal $x_0[n]$, given the parameter f_c,

$$x_0[n] = D\left\{y_0[n];\ f_c\right\}. \tag{3.8}$$

Then the optimal value f_c^* can be defined as the f_c that minimizes the root mean square (RMS) of $x_0[n]$,

$$f_c^* = \arg\min_{f_c} \mathrm{RMS}\left\{\mathcal{D}\left\{y_0[n];\ f_c\right\}\right\}. \tag{3.9}$$

A simple line-search algorithm can be used to find f_c^*, starting from an initial guess of 3 kHz. The resulting optimal carrier frequency is then used in the demodulation operations of other signals recorded during the same session.

3.1.3 Preliminary data and discussions

3.1.3.1 System validation

The first usage of the device we made is to examine some sample recordings under certain controlled conditions so as to verify the feasibility of the acoustic-aerodynamic setup and the signal processing algorithms. Recordings were made by a male adult speaker, uttering a sample sentence, "Say /d ae n/ please," repeated three times in an acoustically dampened booth. A head-mounted AKG HSC-200 condenser microphone was used to record the acoustic signal. The microphone was placed off-axis, 5 cm away from the edge of the speaker's mouth. The sensor box was attached to the frame of the headset. The open end of the L-shape glass tube was placed under the opening of one nostril of the speaker. The FM airflow signal from the processing box and the acoustic signal from the microphone were recorded simultaneously through two channels of an MAudio Delta-1010 sound card in a computer, and they were stored digitally on the hard drive. The two waveforms were both sampled at 44.1 kHz and quantized in 16-bit PCM format. This sampling rate is far greater than the carrier frequency of the FM signal. A 20-second of zero-airflow signal was recorded in the booth after the sentences were recorded. A zero-phase low-pass filter was applied to each FM signal. The cutoff frequency was set to 5 kHz in order to filter out higher-order harmonics. The accurate carrier frequency is calculated from the filtered zero-airflow signal using the search algorithm presented in Section 3.1.2.2. The resulting f_c^* was used to demodulate all other filtered FM signals, using the demodulation algorithm described in Section 3.1.2.1. Besides the signal processing, phoneme boundaries in all sentences were manually labeled by examining the waveforms of the acoustic signals.

Figure 3.5 shows the acoustic and nasal airflow waveforms of one sample sentence in parallel. In the top and middle panels of the figure are the acoustic signal and the demodulated airflow signal, respectively. It can be noticed that the demodulated airflow signal contains strong harmonic components within the sections of voiced speech sounds. These components are caused by acoustic vibrations of the air that propagate from the nostril. This information is redundant since the acoustic signal has been picked up by the other channel. To eliminate these harmonic components, the demodulated airflow signals

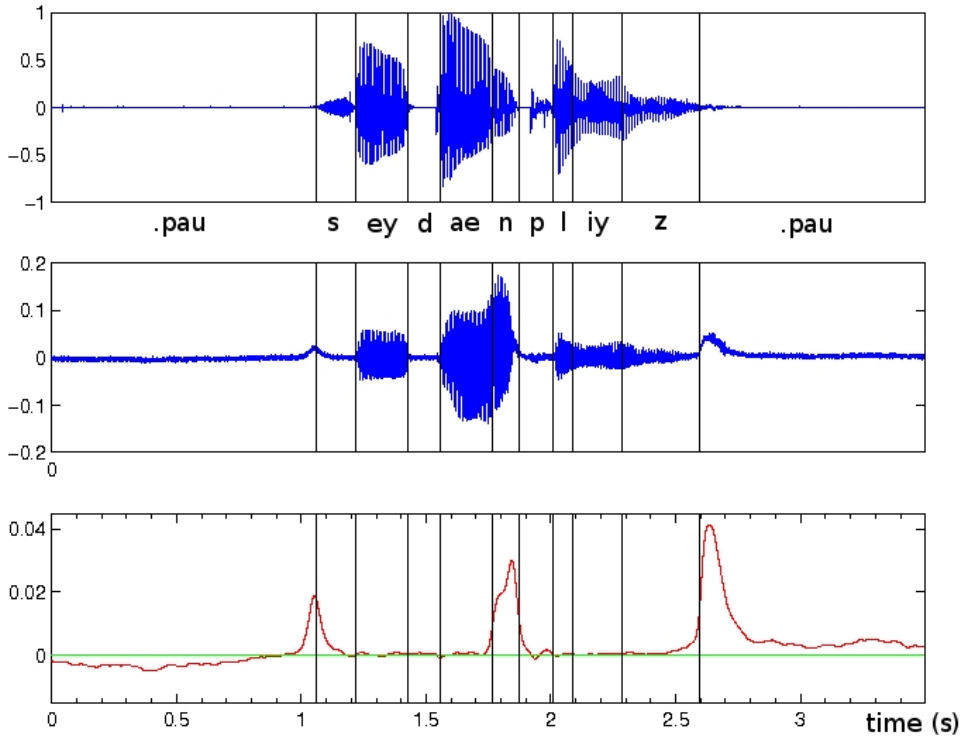


Figure 3.5: Acoustic and nasal airflow signals of the sentence "Say /d ae n/ please". The acoustic signal is plotted in the top panel; the demodulated airflow signal in the middle; and the static airflow in the bottom panel. The vertical bars represent phoneme boundaries.

were convolved with a 30 ms normalized Hamming window. This operation is equivalent to a weighted average of the signal within the window length, effectively low-pass filtering the signal with a cutoff frequency of about 20 Hz. The DC and low-frequency components in each demodulated airflow signal are left, representing the *static airflow* as it moves in and out of the nostril during speech. In the bottom panel of the figure is the resulting static airflow signal that contains the aerodynamic information.

It is informative to compare the acoustic signal and the static airflow signal in the figure. The nonsense word /d ae n/ in the sentence was designed to trigger a VP opening and then closing movement during the utterance. This event can be clearly identified from the prominent peak at about 1.8 seconds in the static airflow signal. According to the phoneme boundaries obtained from the acoustic signal, this nasal airflow peak occurs in the middle of the nasal segment /n/. Slight amounts of out-going nasal airflow start

before and end after the boundaries of the nasal segment, which indicate a mild degree of coarticulation effect. Similar nasal airflow peaks corresponding to nasal phonemes can be consistently observed in all the sample sentences. It is implied that the device we built has successfully picked up the information of certain aerodynamic events caused by VP movements during speech.

It is to be noted that two other prominent peaks can also be seen in the static airflow signal. One is located just before the beginning of the sentence, and the other is located right after the end of the sentence. Similar peaks are observed in all the static airflow signals of all the recorded sentences. Since all the sentences begin with a consonant /s/, whose target VP status is closure, the velum has to move from its rest position to close the VP port while the oral pressure increases at the beginning of an utterance. This action may rapidly push a certain amount of air out of the nose at the beginning of each utterance, thus causing the former peak. At the end of each utterance, the velum always returns to its rest position. When the sentence ends with a pressure consonant (such as /z/ in the recordings) that requires a closing VP port, extra air may be suddenly released from the nose after the utterance, thus causing the latter airflow peak. In the acoustic channel, there are no noticeable signals corresponding to these two peaks.

In addition, a slow nasal inhalation can be observed as a negative section of signal before the sample sentence in the figure, and a slow nasal exhalation is observed after the sentence. The negative nasal inhalation signal before an utterance may or may not be observed in other samples, because the speaker sometimes inhaled through the mouth instead of the nose. The nasal exhalation after an utterance is always observed in each sample, but varies in both scale and slope of change among different sentences. This variation indicates the speaker could control the release of breath at the end of a sentence in different ways. It can also be seen that no acoustic signals correspond to the nasal inhalation and exhalation process. Because some aerodynamic events do not necessarily cause acoustic sounds during speech, these events can not be detect from acoustic signals.

CVN	NVC	NVN	CVC
/d ae n/	/n ae d/	/n ae n/	/d ae d/
/s aa n/	/n aa s/	/n aa n/	/s aa s/
/t iy n/	/n iy t/	/n iy n/	/t iy d/
/z uw n/	/n uw z/	/n uw n/	/z uw z/

Table 3.1: Four groups of words in carrier sentences to study the patterns of nasal resonance.

3.1.3.2 Airflow patterns of nasal resonance

With the aid of the above two-channel setup, we can study the relationship between the nasal airflow and the acoustic signal when the nasal resonance is achieved during normal speech. For the purpose of systematic comparisons, four groups of monosyllable nonsense words were designed in the forms of CVN, NVC, NVN, and CVC. In these groups, C is a plosive of fricative chosen from /t/, /d/, /s/, /z/, V is a vowel chosen from /iy/, /ae/, /aa/, /uw/, and N is the alveolar nasal /n/. These groups of words are supposed to represent the opening process, the closing process, the complete opening, and the complete closing of the VP port within a monosyllable, respectively. Table 3.1 lists the set of 16 words that were used in data collection. During the recording procedure, each word was inserted in the carrier sentence, "Say ___ please." The recordings were made by a male speaker. The system setup was the same as that in the previous section. The four groups of sentences made up one recording session. Three sessions of recording were made by the same speaker with a 5-minute rest between consequent sessions. At the end of each session, a 20-second portion of zero-airflow signal was recorded for the calculation of an accurate carrier frequency of that session. Then all FM airflow signals were processed in the same way as in the previous section, obtaining the static airflow signals. Phoneme boundaries were also labeled manually for each sentence according to the acoustic signals.

In order to study the detailed airflow patterns during the production of different words, we compared the static nasal airflow signal corresponding to the phoneme sequence of each word and its adjacent phoneme contexts. Figure 3.6 displays one session of the static nasal airflow signals. Each panel in the figure represents the word in the corresponding cell of Table 3.1. Each column of the panels belongs to one group of the words. It can been seen

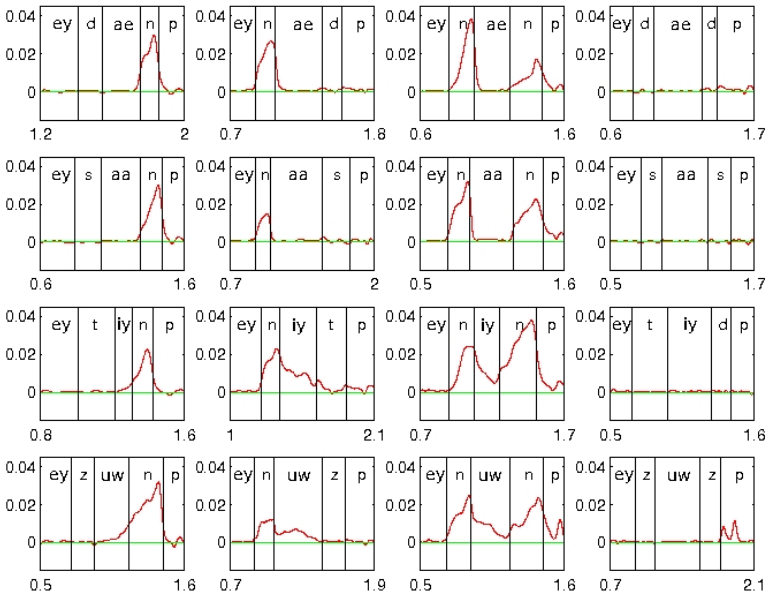


Figure 3.6: Static nasal airflow signals of recorded words. Each section of the signal is cut from the carrier sentence. Only the phonemes that are adjacent to the word on both sides are shown. The phoneme boundaries are marked manually according to the simultaneously recorded acoustic signals.

that the patterns of the static nasal airflow signals within each column are similar to each other. Without any exceptions, airflow peaks occur during all nasal segments, while the static airflow signals are almost zero during the CVC segments.

In the CVN group, the positive rise of the airflow can start early in the vowel segment due to anticipatory coarticulation. In the NVC group, the positive airflow drops after the nasal but can extend into the following vowel, presenting the effect of carry-on coarticulation. In the NVN group, the positive static airflow signals can extend into the vowels from both directions. However, in all of these three groups, it can be observed that the extension of the positive static nasal airflow is more prominent for the high vowels (/iy/ and /uw/) than for the low vowels (/ae/ and /aa/). This difference can be explained by the fact that the oral cavity of a high vowel has greater airflow resistance than that of a low vowel. When the oral cavity is widely opened, like in cases of low vowels, less static airflow passes through the nasal tract. It can be confirmed perceptually that the recorded vowels in the NVN condition were all nasalized, though few static airflow signals were measurable for some low vowels. This observation indicates that the airflow measure of

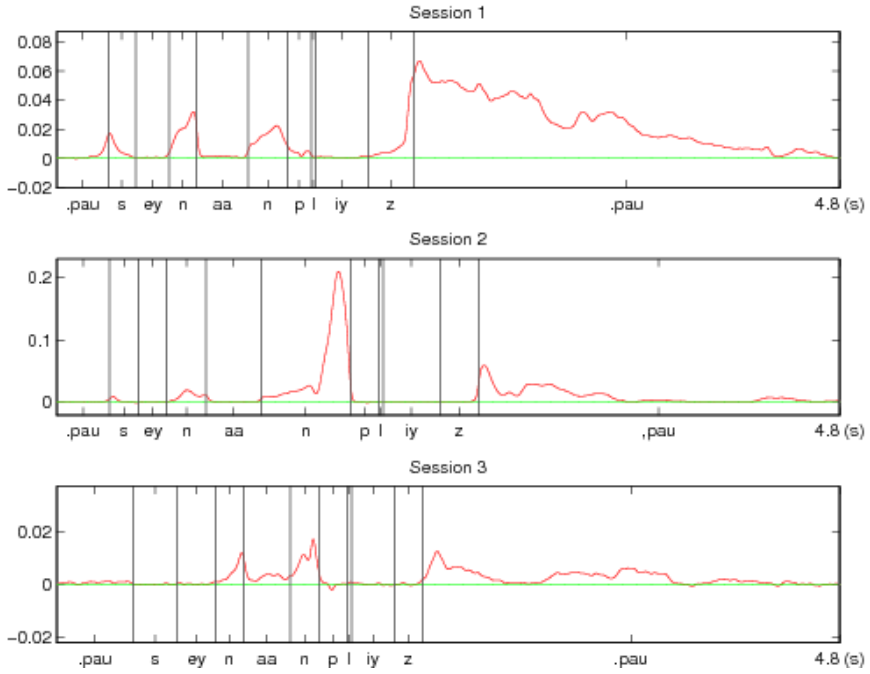


Figure 3.7: Three recording sessions of static nasal airflow signals of the sentence "Say /n aa n/ please", produced by the same speaker.

nasal resonance is highly dependent on oral configurations, which is consistent with the findings of other researchers (Moll and Daniloff, 1971; Dang and Honda, 1994a).

In some panels of Figure 3.6, a slight amount of nasal airflow emissions can be observed between /n/ and /p/ segments of the signals, which can also be explained as coarticulation effects. In addition, it is to be noted that there are two adjacent airflow peaks during the /p/ segment after the last word /z uw z/ in the figure, which does not happen in every session. It can be explained that the speaker may have added a pause before the word "please" during the recording of this particular sentence. The two peaks indicate the end and then the beginning of two adjacent utterances.

3.1.3.3 Variability of the measurement

By repeating the recording sessions, we studied the intra-speaker variability of the measurement with the new device. Figure 3.7 shows three renditions of the static airflow signals

of the same sentence, "Say /n aa n/ please", produced in different recording sessions. Airflow peaks corresponding to nasal phonemes and nasal exhalations can be identified in all samples. It can be seen that the nasal exhalation curves at the end of sentences are quite different from one to another, which exhibits the speaker's flexibility of exhaling inaudible airflow. The airflow curves during nasal phonemes also vary in both amplitude and shape. It is to be noted that the value scales of the three sessions in the figure are different, possibly due to the measurement sensibility to the change of the prob-tube position in different sessions. This variation sometimes caused the signals to run out of the range of the device during data collection.

3.1.3.4 Discussions

In the experiments presented in this section, we first successfully tested the functionality of a new device that serves as an additional channel to the conventional microphone setup for nasalization analysis. Then we designed controlled speech materials and collected preliminary data with the proposed aerodynamic-acoustic setup. Nasal emission and nasal resonance effects were investigated by comparison studies of the two-channel signals. At last, the variability of the airflow measurement was examined.

The first experiment (3.1.3.1) shows that certain nasal airflow events, such as nasal phonemes, nasal emissions, and nasal exhalation and inhalation can be detected from the aerodynamic signal, even though sometimes there may be little information in the corresponding acoustic signal. Negative nasal air (inhalation) signals are often observed before an utterance, and positive nasal exhalation signals usually appear after an utterance. The observed nasal inhalation and exhalation signals vary in both scale and slope of change. The variation indicates that though the buildup of air pressure for an utterance and the release of the extra air after it are necessary, a normal speaker can freely do this in various ways. These observations indicate that, for the purpose of speech production, while some of the nasal airflow events are accompanied with proper sounds with nasal resonance, others may be controlled within an inaudible range (i.e. very low level of acoustic radiations). It has been observed that in disordered speech there are sometimes nasal airflow leakages that may weaken the acoustic energy of non-nasal sounds. The complementary information

of nasal airflow events detected from the aerodynamic signal can be especially useful in detecting or diagnosing nasal emission problems in disordered speech.

The aerodynamic data collected in the second experiment (3.1.3.2) reveal relatively complex patterns of nasal resonance. It turns out that the nasalized vowel segments are not necessarily marked by significant amounts of static airflow velocity released through the nasal tract, especially for low vowels. There can be measurable acoustic vibrations radiated out of the nasal tract, even though the net nasal airflow velocity is almost zero. So it may be unreliable to determine the nasal resonance status only from the aerodynamic signal.

The multiple recording sessions of data in the third experiment (3.1.3.3) reveal some practical issues of the new device. First, the measurement turns out to be sensitive to the position of the probe-tube. Second, the variations of the airflow signals corresponding to the same acoustic event may make this aerodynamic measurement a more qualitative representation.

3.2 Additional acoustic channel

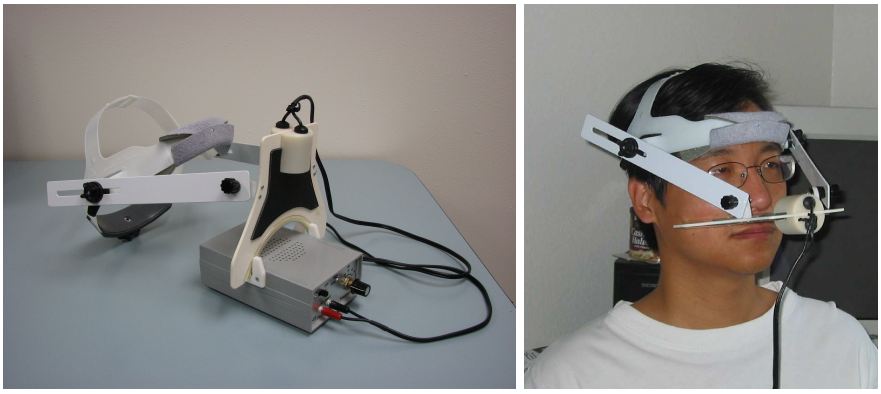
The key motivation of introducing an additional acoustic channel for nasalization analysis is to reduce the mixing effects of acoustic radiations from the nasal and the oral tracts. Consequently, the inherent information ambiguity in the mixed acoustic signal may be alleviated, so that it may be easier to extract the information about the VP status during speech. However, it is difficult to separate acoustic radiations perfectly. There has to be a trade-off between acoustic separation and the ease of data collection. This section introduces existing techniques of multi-channel acoustic measurements. A simple device is chosen for further investigations. Preliminary data are collected to examine the extra information captured by the additional channel.

3.2.1 Measuring devices

Multi-channel acoustic measurements have only been used in a few studies that aim at understanding detailed acoustic characteristics of nasalization. In these studies, a huge

amount of effort was spent to obtain separate acoustic signals from the mouth and the nose. For example, Suzuki *et al.* (1990) have designed a sound proof box with two vertical isolated chambers to achieve the mouth-nose separation. The speaker can place the face against a sealed mask that that is attached to one side of the box. Two microphones mounted in the upper and lower chambers can pick up nasal and oral signals separately. From the data collected with this device, they have found the cross-velar vibration phenomena during the production of vowels and voiced plosives when the VP port is closed. This finding indicates that sound waves can transmit from the oral to the nasal cavity through vibrations of the closed velum. So, strictly speaking, there are rarely "pure" oral voiced phonemes in natural speech. In Schell and Lacroix's (2003) studies, they built a special recording booth that was divided into two parts with a large horizontal sound-separating plate. A hole was opened on the plate, through which a speaker could place their head. The acoustic signals from the nose and the mouth were then recorded by two microphones in the two separate parts of the booth. These dual-channel acoustic data were used to jointly estimate the parameters of a branched tube model of the oral and nasal tracts. Recently, Feng and Kotenkoff (2006) made a device similar to what has been used by Suzuki et al., and they collected dual-channel acoustic data to test a special acoustic model on spectral patterns of nasal vowels in French (Feng and Castelli, 1996). In the model, a nasal vowel is supposed to be in a transition status from an oral-only configuration to a pharyngeal-nasal configuration, and its spectrum is modeled as an intermediate representation of the spectra of the two extreme configurations. All the above special designs for acquiring almost "perfectly" separated multi-channel acoustic signals may not be widely used due to the inconvenience of data collection. This difficulty may hinder deep investigations into detailed acoustic characteristics of nasalization.

In the last chapter, we introduced some simple multi-channel acoustic devices that have been used in studies of resonance disorders (see Section 2.3.2). Those clinical studies mainly focused on the correlation between the instrumental measurement and the perceptual assessment of nasality. So the purpose of using multi-channel acoustic devices, such as those in the Nasometer or the NasalView systems, was to find empirical "index" scores of nasality. These devices are relatively easy to make and convenient to use, though they



(a) Calibrating the channel balance. (b) Recording.

Figure 3.8: The NasalView device in use. The sub-figure on the left shows how the two channels are calibrated with a mono tone generator. The sub-figure on the right shows a speaker wearing the headset to make recordings, with the sound-isolation plate placed between the nose and the mouth.

may not achieve perfect separation of nasal and oral acoustic signals as the more complex devices do. There still lack enough theories and corresponding analysis approaches that may help to extract more detailed information about nasalization with these simple devices. We expect to extract more spectral and temporal information with such devices by introducing new signal processing techniques. Before that, in the following section, we will first study some dual-channel acoustic data collected with the simple device that is used in the NasalView system, and then study the basic information that can be directly extracted from the data.

As shown in Figure 3.8, the hardware of the NasalView system is composed of a headset, a sound-isolation plate, and two microphones. The speaker wears the headset during data collection. The position of the sound-isolation plate is adjusted so that the nose and the mouth are separated horizontally from the upper lip. Signals picked up by the microphones are first fed to a circuit unit that pre-amplifies them with appropriate gains, then they are sampled by the sound card of a computer through the stereo channel. The circuit unit also contains a tone-generator and a speaker that are used to calibrate the balance between the nasal and oral channels. At the beginning of each recording session, the plate can be mounted above the circuit unit, with the microphones to the same distance from

	/aa/	/iy/	/m/	/n/
unblocked	-6.14	3.62	22.00	23.86
blocked	-16.75	-12.23	22.57	25.59

Table 3.2: Magnitude difference of RMS energy energies (in dB) between the nasal and oral signals of sustained phoneme under two conditions.

the speaker. The tone-generator generates a mono tone, while a calibration program is sampling signals from the sound card and displaying short-time RMS energies of the two signals. The two energy gains are tuned to be approximately equal to each other. For the flexibility of signal analysis, we only use the NasalView device to collect the raw acoustic data. Since the raw data can be directly acquired through two channels of a sound card, it is convenient to design our own programs independently for various purposes such as channel calibration, data recording and visualization, and other sophisticated analyses.

3.2.2 Preliminary data and discussions

3.2.2.1 Acoustic leakage test

Our first experiment with the NasalView device was to test its acoustic leakage characteristics. A male speaker wore the NasalView device to make recordings in the same booth as in the previous experiments (Section 3.1.3). Signals were sampled through two channels of the MAudio Delta-1010 sound card and stored in 22,050 Hz, 16-bit PCM format. The channel balance was tuned to be less than 0.01 dB of the difference in RMS energy with the calibration signal before the recording session. A simple experiment was designed to record four sustained phonemes, /aa/, /iy/, /m/, /n/, under two conditions. One was the "unblocked" condition in which the speaker produced each phoneme with the regular setup; the other was the "blocked" condition in which the speaker's nose was covered with a piece of foam when /aa/ and /iy/ were produced, and then the mouth was covered when /m/ and /n/ were produced. Each phoneme was produced continuously for about 3 seconds. After recording, the central 1 second portion of each phoneme was analyzed. The magnitude difference of RMS energies ($10 \log\left(\bar{E}_n/\bar{E}_m\right)$) between the nasal and oral signals was calculated from the central section.

Table 3.2 lists the magnitude difference of each phoneme under each condition. It

can be seen that the difference is always greater than 22 dB during the production of the nasal phonemes, /m/ and /n/. This order of acoustic attenuation is consistent with the sound-separation characteristic (20 dB) described in the manual of the NasalView device. Since a nasal sound is mainly radiated from the nasal tract, the signal captured by the oral microphone must be a mixture of the acoustic leakage across the sound-separation plate and some radiation from the tissue of the closing mouth. Under the blocked condition of nasal phonemes, less radiation from the mouth was captured, so the energy differences become greater for /m/ and /n/.

Because vowels are presumably oral-only sounds, it was expected that the energy differences of vowels would be close to those of nasals except the opposite sign. However, the absolute values of the energy differences of vowels are far less than those of nasals under both conditions. The unblocked vowel /iy/ even caused more energies in the nasal channel than in the oral channel. This amount of acoustic energies in the nasal channel for vowels can not be explained only by the acoustic leakage cross the sound-separation plate of the NasalView device. It can be better explained by acknowledging the cross-velum transmission of acoustic vibrations during vowel production. These vibrations can propagate through the unblocked nasal tract and be picked up by the nasal microphone. When the nose is blocked, this portion of energies are mainly attenuated, which reduces the value of $\bar{E}_n/\bar{E}_m$, and therefore the magnitude differences. It can be inferred from the unblocked vowel data that there is a greater proportion of total acoustic energies transmitted through the nasal tract during /iy/ than during /aa/. This difference may account for the higher acoustic resistance of the oral tract for the high vowel than that for the low vowel.

3.2.2.2 Nasalance patterns of nasalization

The NasalView system has been conventionally used to measure the nasalance score that is the ratio of the energy from the nasal channel to the total energy from both channels. In our implementation of nasalance calculation, nasal and oral signals were first filtered with a pass-band of 350-650 Hz, then the short-time RMS energy of each signal was calculated and the ratio between the nasal energy and the total energy was obtained. The window length and frame shift for the calculation were set to 20 ms and 5 ms, respectively. The resulting

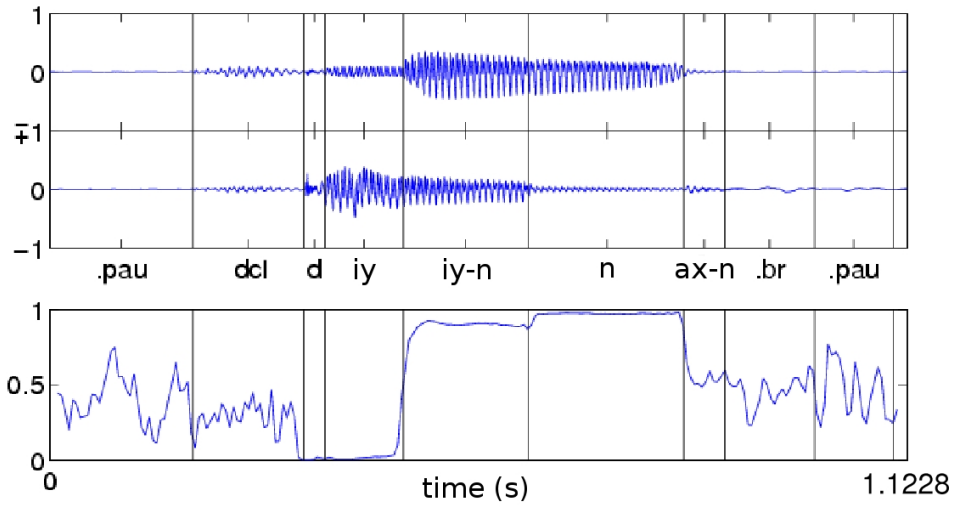


Figure 3.9: Dual-channel acoustic signals of an isolated word "dean". From the top panel to the bottom panel are the nasal acoustic signal, oral acoustic signal and the smoothed nasalance signal, respectively. The vertical bars represent phoneme boundaries.

nasalance signal was then optionally smoothed with a median filter. In this experiment, we looked for patterns of nasalance scores during normal speech.

In the experiment, several samples of isolated words were first recorded and processed to test the system setup. These words were uttered by a male adult speaker. The channel calibration and recording procedure were the same as those in the previous experiment. Nasalance signals were then calculated from the recorded data.

Figure 3.9 shows one typical example of the nasalization process when an isolated word "dean" is produced. In the top panel, the nasal and oral waveforms are displayed in parallel: the upper one is the nasal signal and the lower one is the oral signal. A detailed phonetic transcription is labeled under the waveforms, and the corresponding segments are manually marked. It is easy to identify the nasal phoneme /n/ by comparing the signal energy between the two channels. It is clear that the vowel switches from non-nasalized (/iy/) to nasalized status (/iy-n/) within the utterance. Note that not all nasalized vowels in the samples had such a distinct switch of nasalization status, in stead, the coarticulation effect of nasalization often presented across the whole phoneme. These status changes can be reflected in the smoothed nasalance curve in the bottom panel. As a general trend, the

nasalance scores tend to be low in non-nasalized vowel sections, higher in nasalized vowel sections, and the highest in nasal sections. It can also be seen that the nasalance scores are relatively noisy during the pauses (/.pau/ and /.br/), the voiced bar of the stop closure (/dcl/), and the short release (/ax-n/) of the nasal phoneme, though the curve has been smoothed. The noisy nasalance signals usually happens when the acoustic signals of two channels are both weak.

In order to study the patterns of nasalance curves in different contexts, we reused the speech materials in the airflow experiment described in Section 3.1.3.2. The four groups of words in Table 3.1 were recorded in the same carrier sentence by the same speaker as in the experiment in Section 3.1.3.2. The difference was that the airflow and microphone setup was replaced by the NasalView device. The channel calibration and recording procedure were the same as in the previous experiment. After recording, all the sentences were manually transcribed with phonetic labels by examining the waveforms of the dual-channel signals. Then nasalance curves were calculated for all the sentences with the same parameter settings as above. No smoothing operations were applied to the nasalance curves.

Figure 3.10 shows the nasalance curve of each word within its adjacent phoneme neighbours. The four columns of panels correspond to the four word groups, CVN, NVC, NVN, and CVC, respectively. In all panels, the nasalance curves are relatively smoother in sonorant segments than in non-sonorant segments, such as plosives and fricatives. All the nasal phonemes in the first three columns represent a continuous high level of nasalance scores that are very closed to one. The nasalance scores of a vowel within the non-nasalized context (CVC) are always lower than those of its counterpart within the nasalized context (NVN). In the CVC context, the nasalance scores of vowels /iy/ and /uw/ are higher than those of vowels /aa/ and /ae/. This contrast can also be easily identified between /iy/ and /aa/ in the NVN, NVC, and CVN contexts.

3.2.2.3 Variability of nasalance

We have found that the NasalView system is robust enough to collect separated channels of acoustic signals. In this experiment, we wanted to study both intra- and inter-speaker

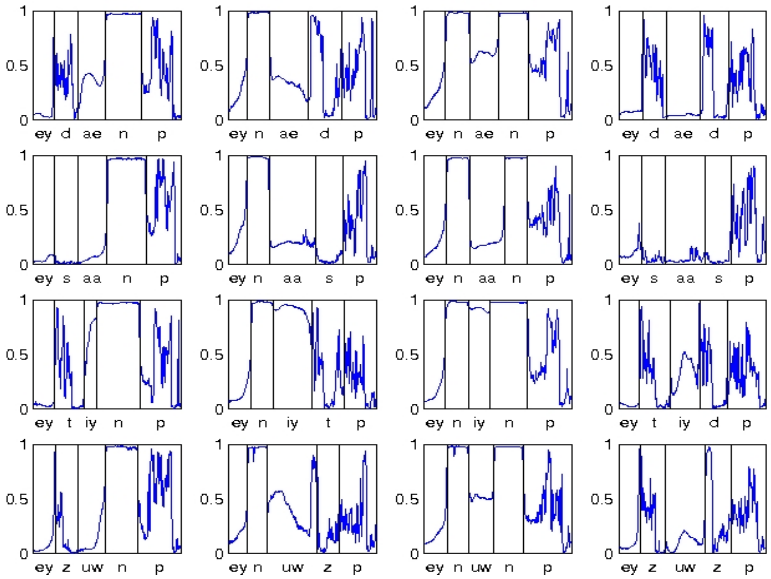
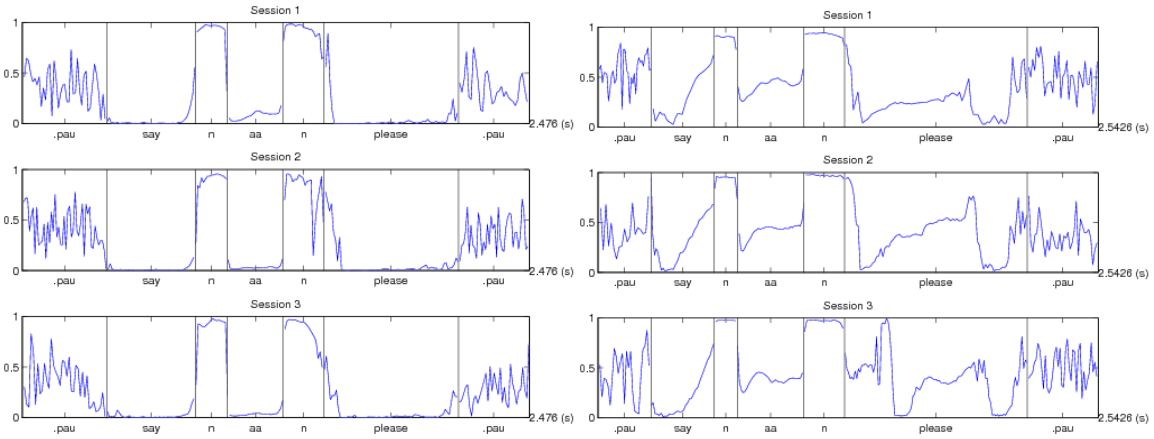


Figure 3.10: Nasalance curves of recorded words. Each section of the signal is cut from the carrier sentence. Only the phonemes that are adjacent to the word on both sides are shown. The phoneme boundaries are marked manually by examining the acoustic signals of both channels.

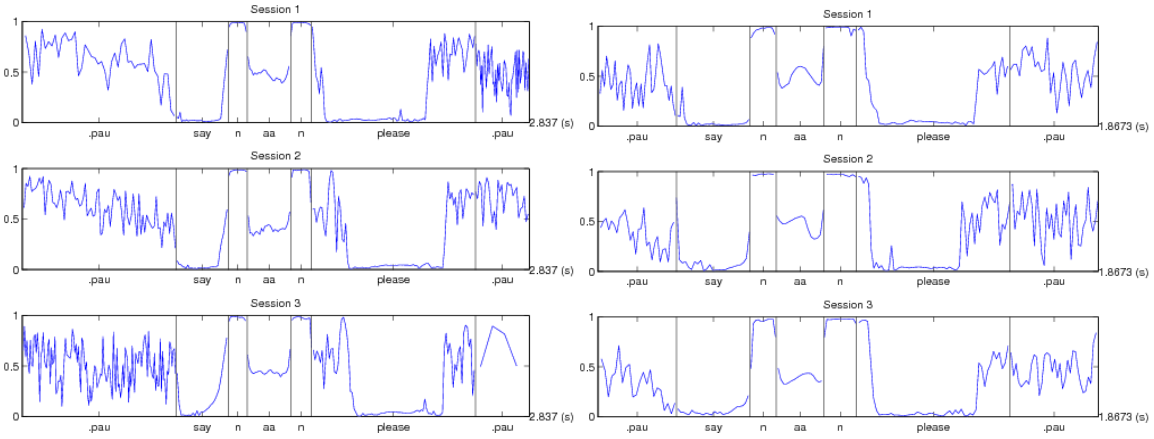
variability of nasalance during the production of nasal resonance.

In the experiment, six normal adult speakers (three male and three female) were asked to make recordings of the sentences in previous experiment. Each speaker made three sessions of recordings. The channels were calibrated before each recording session and the recording procedures were the same as those in the previous experiment. All waveforms were manually segmented. Nasalance signals were then calculated from the recorded data.

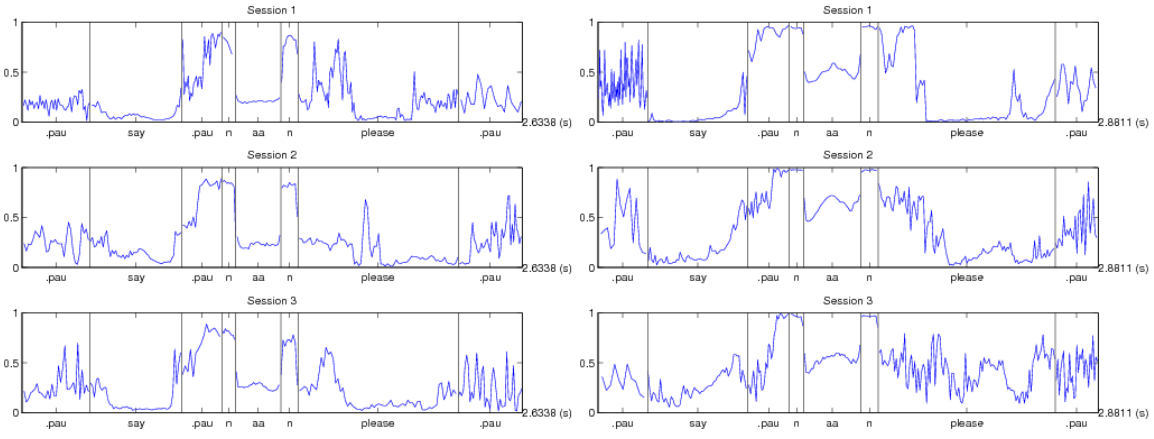
In order to study the variability, we plotted different renditions of the nasalance signals of each word in the carrier sentences. Figure 3.11 shows an example sentence, "Say /n aa n/ please", in which the vowel /aa/ is supposed to be nasalized due to the context effect. In the figure, the nasalance trajectories are time-warped within each segment according to the average segmental durations of each speaker. It is easy to identify the similarity of the nasalance trajectories of the nasalized vowel among the recording sessions of each speaker; while different speakers tend to produce trajectories with different stylized shapes.



(a) Speaker 1 (b) Speaker 2



(c) Speaker 3 (d) Speaker 4



(e) Speaker 5 (f) Speaker 6

Figure 3.11: Intra- and inter-speaker comparison of nasalance trajectories. The same sentence, "Say /n aa n/ please", is recorded by six speakers. Each speaker has three recording sessions.

3.2.2.4 Discussions

In the three experiments presented in this section, we first tested the NasalView device that serves as a simple dual-channel acoustic setup for nasalization analysis. Then we collected recording with controlled speech materials in order to study the basic functionality of the NasalView device. Finally we had a qualitative study of the intra- and inter-speaker variability of the nasalance trajectories during nasalization.

The first experiment evaluated the cross-channel acoustic leakage effects of the NasalView device. During vowel production, the excessive amount of acoustic energy in the nasal channel can be explained as a cross-velum effect, which confirms previous observations by other researchers. The attenuation of average RMS energy across the sound-separation plate is about 20 dB. This degree of leakage due to imperfect sound-separation can be regarded as acceptable for further analyzing the relationship between the two channels of signals. Visual examination of the signals confirmed that the energy changes in two channels are much more easily detected than those in convectional signal-channel signals, and detailed information about the VP status is also more easily identified by examining two signals together.

The second experiment was designed to observe the patterns of nasal resonance in different contexts by displaying the nasalance curves of controlled materials. It is shown that the nasalance curves can reflect the change of the nasal resonance status. Nasalance scores have been used in many clinical applications as an objective measure of nasality in speech. As a simple measure of oral-nasal energy-balance, the nasalance score provides much information about the VP function that may be difficult to extract from a single-channel acoustic signal. The *nasalance index,* as a global measure, is usually obtained by averaging the nasalance scores of a normative passage read by the speaker. However, considering the segment-level information, as can be seen in the speech samples we recorded, the instantaneous nasalance scores within non-sonorant segments are not as stable as those within sonorant segments. The instantaneous nasalance scores within the segments of nasalized vowels also present a dependency on the degree of oral opening. For the purpose of detailed analysis and/or modification of the acoustic signals, more knowledge about the relationship

between the two channels of signals may be needed.

The third experiment showed that the intra-speaker variability of nasalance trajectories during nasal resonance is much less than the intra-speaker variability of the airflow signals. The experiment also showed a greater inter-speaker variability of nasalance trajectories than intra-speaker variability. We expect a more detailed analysis of the dual-channel acoustic signal may reduce this variation.

3.3 Summary

In this chapter, we studied two possible ways to introducing an extra instrumental channel in addition to the conventional single-channel acoustic measurement, so as to obtain more information about the nasalization process. One way was to add an aerodynamic channel, another was to add an additional acoustic channel.

For the additional aerodynamic channel, we designed and made a noninvasive device that can be used to measure the velocity of the nasal airflow out of a nostril during speech. A small, low-cost differential pressure sensor is used to pick up the dynamic pressure of the airflow. The airflow signal is frequency modulated so that it can be recorded by a generic sound card. The airflow signal is recovered by a demodulation operation. A filtering process extracts the static nasal airflow from the demodulated signal. The sensor is light enough to be attached to a headset. The usage of this device neither interferes with the articulatory process during speech, nor causes degradation of the simultaneously recorded acoustic signal, which is critical to the further analysis and enhancement of acoustic signals of disordered speech.

The static airflow signal obtained with the new device provides additional information that is not present in the acoustic signal. The information includes the breathing patterns during speech and the nasal emission events. We have observed that the airflow patterns of breathing during speech can be freely varied. Besides nasal exhalations, the positive peaks in the airflow signals are also associated with major nasal emission events, including the production of a nasal phoneme and the preparation of a oral pressure sound from the velar release state. The new device turns out to be more sensitive to nasal emissions than to

nasal resonance. A potential application of the airflow device is to detect the unnecessary nasal emissions that may weaken the oral acoustic signals during non-sonorant speech segments.

A trade-off we made during the design of the airflow device is that the probe-tube picks up the airflow in an open space in stead of a sealed mask, so that the acoustic signals will not be distorted. A drawback of this design is that the airflow signal can be sensitive to the relative position of the probe-tube in the nasal airflow, which makes the measurement more qualitative in essence. For the purpose of a large scale of data collection and detailed qualitative analysis, the robustness of the current device may still need to be improved.

For the additional acoustic channel, we adopted the simplest existing device (NasalView) and examined its potentials for nasalization analysis. The difference between the two channels of the signals from the NasalView device provides information about the nasal resonance that may be not distinguishable in the mixed signal-channel signals. Compared to the multitude of studies on nasalization using single-channel acoustic data, there have been surprisingly few studies to understand the acoustic characteristics of nasalization using dual-channel acoustic data. Although the major usage of the similar devices is limited to obtaining the index of nasality, it is expected that more detailed information about the oral-nasal articulatory process can be extracted from the dual-channel acoustic signal by introducing more sophisticate analysis techniques.

Although both additional aerodynamic and acoustic signals provide extra articulatory information of nasalization, the focus of our studies in the rest of this thesis will be narrowed down to the dual-channel acoustic analysis only. One practical reason is that the acoustic signals we can currently collect are more reliable in quantity than the airflow signals. Since the airflow device we developed is in the preliminary stage, it still needs to be improved for a wider usage. A more stable placement of the probe tube is necessary for consistent measurements. The processing circuit may also need to be updated to allow a larger range of airflow velocities to be measured. Another practical reason for focusing on acoustic studies is that our knowledge of aerodynamics in speech is still limited (Shade, 1999), compared to that of acoustics. Most of the aerodynamic theories on speech production are still in the state of qualitative descriptions.

However, it is to be noted that acoustic measurements are incapable of providing complete nasal emission information, because airflow is not necessarily accompanied by detectable sounds, and microphones are designed to be insensitive to airflow; thus our remaining studies are essentially limited to nasal resonance problems. Note that nasal emission information is important for determining the weakness of plosive or fricative production in disordered speech, thus we are ignoring some interesting information for the sake of convenience.

Chapter 4

Dual-channel acoustic analysis for nasalization

In this chapter, a dual-channel acoustic model for nasalization is investigated through theoretical derivation and computational simulation.[1] Detailed studies of this model may enable us 1) to understand the inherent relationship between the nasal and oral acoustic signals during the nasalization process under certain assumptions, 2) to explain certain characteristics of the acoustic signals during the nasalization process, and 3) to derive informative acoustic features of nasalization for practical dual-channel acoustic analyses.

4.1 Transmission-line model

As reviewed in Chapter 2 (see Section 2.2.1), the acoustic system of the vocal tract can be approximately described by a one-dimensional wave equation under certain conditions (for frequencies below 4 kHz). The mathematical description of acoustic waves in the vocal tract is identical to that of electrical waves, so acoustic signals of speech can be modeled in terms of electrical signals propagating in certain circuits. When the nasal and oral tracts are simplified as connected tubes, the transmission-line model can be used to represent how acoustic waves propagate through the tracts. In such a model, the sound pressure and volume velocity of acoustic waves transmitted in a small section of the vocal tract are represented as the voltage and current of electrical waves transmitted in a lumped two-port circuit unit (see Figure 2.1). A series of such circuit units can be connected to

[1] Some materials in this chapter have been published in a conference paper (Niu *et al.*, 2005).

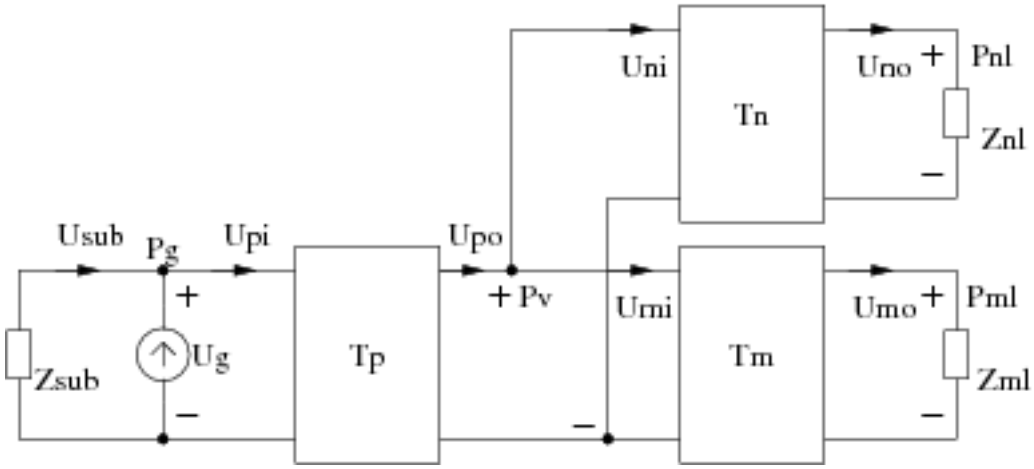


Figure 4.1: Circuit network representation of the production of nasalized voiced sounds (duplication of Figure 2.2).

build an electrical transmission-line network that represents the acoustic process of certain sounds. Acoustic properties of a vocal tract configuration can be derived by analyzing the corresponding circuit network. Acoustic signals can also be synthesized by simulating such a circuit network (Childers, 2000).

4.1.1 Circuit representation

In the circuit representation of an acoustic system, current and voltage signals are substituted by volume velocity and sound pressure signals, respectively. Other electrical parameters, such as impedance and admittance, are also substituted by their acoustic counterparts (Flanagan, 1972). To facilitate circuit analyses, the following notations are use in this chapter. All volume velocity signals are denoted by u, sound pressure signals by p, acoustic impedances by z, and acoustic admittances by y, with distinct subscripts at different parts of a circuit. Lower-case symbols of these variables denote time domain signals or parameters, while higher-case symbols denote their corresponding frequency- or z-domain transforms.

The circuit network in Figure 4.1 is duplicated from Figure 2.2, which represents a simplified model of nasalized voiced sounds. Both ideal vowels and nasals can be regarded as special cases of this model. The whole network is composed of three main sub-nets,

T_p, T_m, and T_n that denote circuit models of the pharyngeal, oral, and nasal tracts, respectively. In the network, Z_{sub}, Z_{ml}, and Z_{nl} denote the acoustic impedance of the subglottal system, the radiation load at the lips, and the radiation load at the nostrils, respectively. P_g, P_v, P_{ml}, and P_{nl} denote the sound pressures at the glottis, the VP port, the lips, and the nostrils, respectively. U_g models the excitation source at the glottis, U_{sub} is the volume velocity out of the trachea, and U_{pi}, U_{po}, U_{mi}, U_{mo}, U_{ni}, and U_{no} denote the volume velocity input and output of each sub-net, respectively. Each sub-net can be regarded as a two-port circuit. According to electrical circuit theory, the transmission property of a two-port circuit is fully characterized by a chain equation (also called the ABCD matrix) in frequency- or z-domain (Balabanian *et al.*, 1969). So the characteristics of the three sub-nets can be described by the following chain equations,

$$\begin{aligned}
\begin{bmatrix} P_g \\ U_{pi} \end{bmatrix} &= \begin{bmatrix} A_p & B_p \\ C_p & D_p \end{bmatrix} \begin{bmatrix} P_v \\ U_{po} \end{bmatrix} \\
\begin{bmatrix} P_v \\ U_{mi} \end{bmatrix} &= \begin{bmatrix} A_m & B_m \\ C_m & D_m \end{bmatrix} \begin{bmatrix} P_{ml} \\ U_{mo} \end{bmatrix} \\
\begin{bmatrix} P_v \\ U_{ni} \end{bmatrix} &= \begin{bmatrix} A_n & B_n \\ C_n & D_n \end{bmatrix} \begin{bmatrix} P_{nl} \\ U_{no} \end{bmatrix}.
\end{aligned} \tag{4.1}$$

Because each sub-net is a serial connection of small two-port units, the ABCD-parameters in each chain equation are locally determined by the physical properties of the corresponding section of the vocal tract. Coupling effects are introduced into the network at the branching points by the following continuous conditions,

$$U_{sub} = U_g + U_{pi}, \tag{4.2}$$

$$U_{po} = U_{mi} + U_{ni}. \tag{4.3}$$

Equation 4.2 introduces the coupling effect of the subglottal tract to the output speech spectrum. Equation 4.3 introduces the coupling effect of the nasal and oral cavities, which accounts for the unique spectral characteristics of nasalized sounds.

4.1.2 Transfer functions

In order to study the spectral properties of nasalized sounds, it is of interest to derive the transfer functions from the glottal volume velocity (U_g) to the output volume velocity of the nasal and oral channels (U_{no} and U_{mo}), respectively. In the following, some useful lemmata are first derived from basic circuit network theory.

1. Transmission characteristic of a two-port network with a shunt impedance

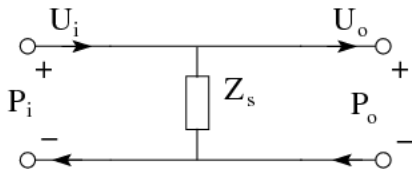


Figure 4.2: A two-port network with a shunt impedance.

Given a two-port network that has a shunt impedance Z_s, as shown in Figure 4.2, when Kirchhoff's Laws are applied, the following equations hold,

$$\begin{aligned} P_i &= P_o \\ U_i &= \frac{P_o}{Z_s} + U_o. \end{aligned}$$

Therefore, *the chain matrix of a two-port network with a shunt impedance Z_s is* $\begin{bmatrix} 1 & 0 \\ \frac{1}{Z_s} & 1 \end{bmatrix}$ (Lemma 1).

2. Input impedance of a two-port network with a load at the output port

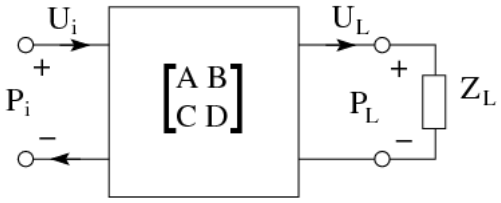


Figure 4.3: A two-port circuit network with a load impedance.

Given a circuit network, as shown in Figure 4.3, in which the chain matrix ($\begin{bmatrix} A & B \\ C & D \end{bmatrix}$)

of the two-port network and the impedance (Z_L) of the load are known, when a constant pressure P_i is applied at the input port, the following equations hold,

$$\begin{aligned} P_i &= AP_L + BU_L \\ U_i &= CP_L + DU_L \\ P_L &= Z_L U_L. \end{aligned}$$

Then P_i can be represented in terms of the input volume velocity U_i as

$$P_i = \frac{AZ_L + B}{CZ_L + D} U_i.$$

Therefore, *the input impedance a two-port network with a load at the output port is* $\frac{AZ_L+B}{CZ_L+D}$ (Lemma 2).

3. Transfer function of a two-port network with a source and a load

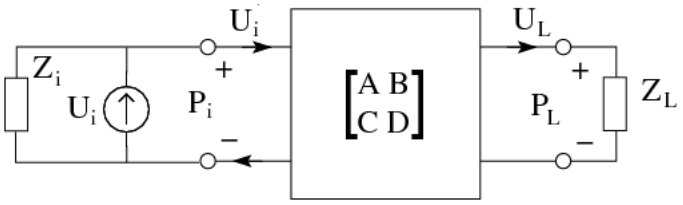


Figure 4.4: A two-port network with a source and a load.

Given a circuit network, as shown in Figure 4.4, in which a two-port network, characterized by the chain matrix $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$, is connected between a volume velocity source U_i and an acoustic load Z_L, and the source has an output impedance Z_i. The transfer function from the volume velocity of the source (U_i) to the volume velocity at the load (U_L) can be derived by applying Lemma 2. Seen from the input port of the two-port network, the input impedance is $Z_o = \frac{AZ_L+B}{CZ_L+D}$. This impedance can be regarded as a load parallel to Z_i, so the volume velocity input to the two-port network is

$$U_i' = \frac{Z_i}{Z_o + Z_i} U_i.$$

U_i' can also be represented with one of the chain matrix equations

$$U_i' = (CZ_L + D)\, U_L.$$

So the two equations can be merged and the transfer function ($\frac{U_L}{U_i}$) is obtained as

$$H = \frac{U_L}{U_i} = \frac{Z_i}{Z_o + Z_i} \cdot \frac{1}{CZ_L + D}.$$

Therefore, *the transfer function from the volume velocity source to the volume velocity at the load is* $\frac{Z_i}{Z_o+Z_i} \cdot \frac{1}{CZ_L+D}$ (Lemma 3).

In order to derive the transfer functions from the circuit network of nasalized voiced sounds, as shown in Figure 4.1, we denote the input impedance of the oral tract with Z_m, and the input impedance of the nasal tract with Z_n, both of which are seen downstream from the junction point of the VP port. Z_m and Z_n can be calculated by applying Lemma 2,

$$Z_m = \frac{A_m Z_{ml} + B_m}{C_m Z_{ml} + D_m}, \tag{4.4}$$

$$Z_n = \frac{A_n Z_{nl} + B_n}{C_n Z_{nl} + D_n}. \tag{4.5}$$

Thus, the circuit network can be converted into an equivalent network as shown in Figure 4.5 (a). It can be seen that Z_m and Z_n are connected in parallel, so their equivalent impedance is

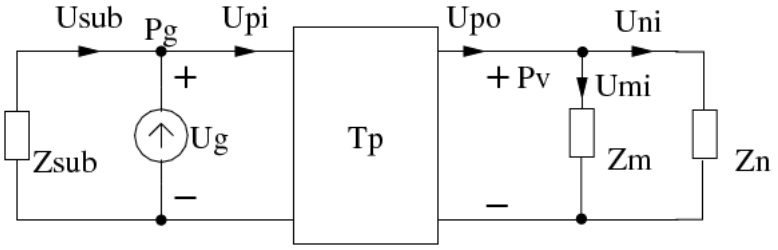
$$Z_{mn} = \frac{Z_m Z_n}{Z_m + Z_n}, \tag{4.6}$$

which can be regarded as the load of the pharyngeal sub-net, seen downstream from the VP port. Then Lemma 2 can be applied again to obtain the input impedance of the whole vocal tract, seen downstream from the glottis,

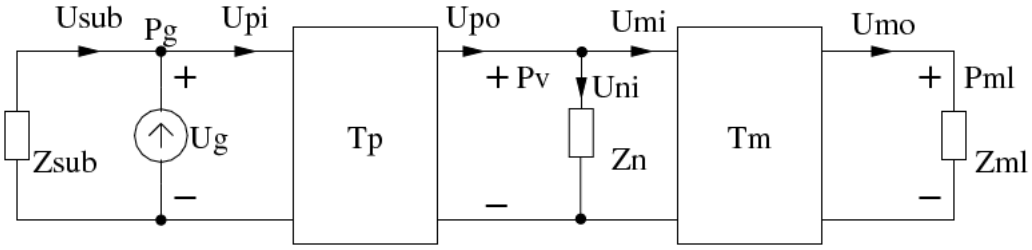
$$\begin{aligned} Z_p &= \frac{A_p Z_{mn} + B_p}{C_p Z_{mn} + D_p} \\ &= \frac{A_p Z_m Z_n + B_p (Z_m + Z_n)}{C_p Z_m Z_n + D_p (Z_m + Z_n)}. \end{aligned} \tag{4.7}$$

To calculate the transfer function from the glottal source U_g to the oral output U_{mo}, the output impedance of the nasal tract can be regarded as a shunt impedance, as shown in Figure 4.5 (b). Then the network between the glottal source and the oral load can be regarded as a serial concatenation of the pharyngeal sub-net, the shunt impedance of the nasal tract, and the oral sub-net. Thus, by applying Lemma 1, the chain matrix of the network is

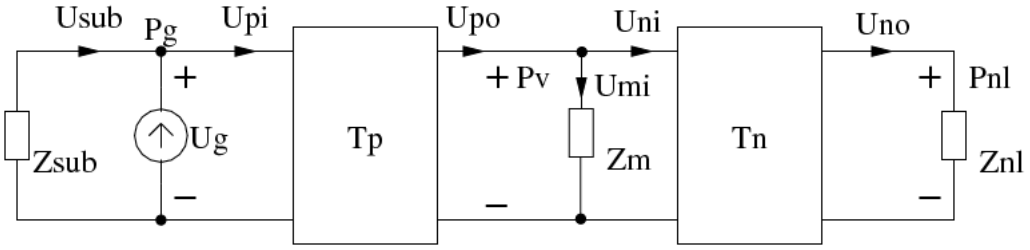
$$\begin{bmatrix} A_{gm} & B_{gm} \\ C_{gm} & D_{gm} \end{bmatrix} = \begin{bmatrix} A_p & B_p \\ C_p & D_p \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \frac{1}{Z_n} & 1 \end{bmatrix} \begin{bmatrix} A_m & B_m \\ C_m & D_m \end{bmatrix}. \tag{4.8}$$



(a) The oral and nasal tracts are represented with their individual equivalent impedance.



(b) The nasal tract is represented with its equivalent impedance.



(c) The oral tract is represented with its equivalent impedance.

Figure 4.5: Equivalent network representations of nasalized sound production. (a) Derivation of the input impedance of the whole vocal tract. (b) Derivation of the oral transfer function. (c) Derivation of the nasal transfer function.

Then Lemma 3 can be applied to obtain the *oral transfer function* $\frac{U_{mo}}{U_g}$,

$$\begin{aligned} H_m &= \frac{Z_{sub}}{Z_p + Z_{sub}} \cdot \frac{1}{C_{gm} Z_{ml} + D_{gm}} \\ &= \frac{Z_{sub}}{Z_p + Z_{sub}} \cdot \frac{Z_n}{(C_m Z_{ml} + D_m)\left[C_p Z_m Z_n + D_p (Z_m + Z_n)\right]} \\ &= \frac{Z_{sub}}{Z_p + Z_{sub}} \cdot \frac{A_n Z_{nl} + B_n}{(C_m Z_{ml} + D_m)(C_n Z_{nl} + D_n)\left[C_p Z_m Z_n + D_p (Z_m + Z_n)\right]}. \end{aligned} \quad (4.9)$$

Note that the oral and nasal tracts are symmetric in the circuit representation in Figure 4.5 (b and c), thus the *nasal transfer function* $\frac{U_{no}}{U_g}$ has a symmetric form to the oral one, which is

$$\begin{aligned} H_n &= \frac{Z_{sub}}{Z_p + Z_{sub}} \cdot \frac{Z_m}{(C_m Z_{ml} + D_m)\left[C_p Z_m Z_n + D_p (Z_m + Z_n)\right]} \\ &= \frac{Z_{sub}}{Z_p + Z_{sub}} \cdot \frac{A_m Z_{ml} + B_m}{(C_m Z_{ml} + D_m)(C_n Z_{nl} + D_n)\left[C_p Z_m Z_n + D_p (Z_m + Z_n)\right]}. \end{aligned} \quad (4.10)$$

Transfer functions of ideal vowels and nasals can also be derived from the circuit network. In an ideal vowel configuration, there is no nasal branch and thus Z_n goes to infinity. The input impedance of the whole vocal tract turns into

$$Z_p^{'} = \frac{A_p Z_m + B_p}{C_p Z_m + D_p}, \quad (4.11)$$

so that the transfer function from the glottal source to the oral output is

$$H_m^{'} = \frac{U_{mo}^{'}}{U_g} = \frac{Z_{sub}}{Z_p^{'} + Z_{sub}} \cdot \frac{1}{(C_m Z_{ml} + D_m)(C_p Z_m + D_p)}. \quad (4.12)$$

Similarly, in an ideal nasal configuration, the oral cavity is closed at a certain point and thus Z_{ml} goes to infinity and the oral sub-net turns into $\begin{bmatrix} A_m^{'} & B_m^{'} \\ C_m^{'} & D_m^{'} \end{bmatrix}$. The impedance of the oral tract and the whole vocal tract are

$$Z_m^{'} = \frac{A_m^{'}}{C_m^{'}}, \quad (4.13)$$

$$Z_p^{''} = \frac{A_p Z_m^{'} Z_n + B_p \left(Z_m^{'} + Z_n\right)}{C_p Z_m^{'} Z_n + D_p (Z_m^{'} + Z_n)}, \quad (4.14)$$

respectively. Then the transfer function of the nasal output is

$$H_n^{'} = \frac{U_{no}^{'}}{U_g} = \frac{Z_{sub}}{Z_p^{''} + Z_{sub}} \cdot \frac{Z_m^{'}}{(C_n Z_{nl} + D_n)\left[C_p Z_m^{'} Z_n + D_p (Z_m^{'} + Z_n)\right]}. \quad (4.15)$$

It can be seen that the transfer functions for ideal vowels and nasals are just special cases of the two transfer functions for nasalized vowels.

4.2 Spectral properties

Spectral properties of voiced sounds can be inferred by analyzing the transfer functions (4.9, 4.10, 4.12, and 4.15) derived from the dual-channel acoustic model. For example, considering the transfer function of an ideal vowel, the effect of the subglottal tract can be assumed to be negligible, since the glottal opening is very small during voiced sounds, and the subglottal impedance Z_{sub} is much greater than the impedance of the vocal tract $Z_p^{'}$. This assumption makes the first factor in transfer function 4.12 approaches to one, so it can be simplified as

$$\begin{aligned} H_m^{'} &\approx \frac{1}{(C_m Z_{ml} + D_m)(C_p Z_m + D_p)} \\ &= \frac{1}{Z_{ml}(C_p A_m + D_p C_m) + C_p B_m + D_p D_m}. \end{aligned} \tag{4.16}$$

When the pharyngeal and oral tracts are considered as a series of acoustic tubes, the above function can be approximated by an all-pole filter in frequency- or z-domain, The form of this transfer function justifies the all-pole model for vowel production that has been used prevalently in speech analysis.

However, the numerators in the transfer functions of nasalized sounds (Equation 4.9 and 4.10) present obvious zero effects in their spectra. It can be seen that the oral-nasal coupling not only introduces zeros into the transfer functions H_m and H_n, but also affects the pattern of their poles. The same denominator in H_m and H_n indicates that the oral and nasal signals share the same poles that are caused by the denominator. The factor, $A_m Z_{ml} + B_m$, in Equation 4.10 indicates that the oral cavity introduces extra zeros to the nasal signal, when the oral cavity is simply considered as a series of tubes; while the factor, $A_n Z_{nl} + B_n$, in Equation 4.9 indicates that the nasal cavity introduces not only zeros but also extra poles to the oral signal, when the shunt effects of sinuses are taken into consideration.

In conventional signal-channel acoustic analysis, signals are transduced with one microphone during nasalization. A physical assumption behind this measurement is that when two waves meet in the same medium, the instantaneous displacement of the medium is

given by the algebraic sum of the instantaneous displacements of individual waves. Therefore, the volume velocity signals from the nose and the mouth are additive. It is often assumed that the addition is equally weighted, so the overall transfer function of the vocal tract, H_v, is equal to the sum of H_m and H_n, that is,

$$H_v = \frac{Z_{sub}}{Z_p + Z_{sub}} \cdot \frac{(A_m Z_{ml} + B_m) + (A_n Z_{nl} + B_n)}{(C_m Z_{ml} + D_m)(C_n Z_{nl} + D_n)[C_p Z_m Z_n + D_p (Z_m + Z_n)]}. \quad (4.17)$$

This equation shows how the zero patterns are be mixed up by the summation operation. However, the equal-weight assumption may not hold in real measurements, because the weights are apparently affected by the different radiation distances from the mouth and the nose to the microphone that is used to pick up the sound pressure instead of the volume velocity. The unknown weights make the zero pattern of H_v almost unpredictable. This partially explains the inconsistent observations of the spectral characteristics of nasalized vowels in the literature. Ideally, separate measurements of the two channels would preclude the uncertainty caused by the mixing of two acoustic channels.

4.3 Dual-channel analysis

Considering the separated transfer functions of the oral and nasal channels in Equation 4.9 and 4.10, their denominators represent complicated inter-dependent coupling effects of the pharyngeal, oral and nasal tracts. Separate analysis of the signal from each channel alone has almost the same degree of complexity as that of the one channel signal.

We have proposed an analysis method (Niu *et al.*, 2005) that makes use of the relationship between the two signals of separated channels. The novel idea behind this method is to cancel out the coupling effects by estimating a special transfer ratio function from the volume velocity of the oral output (U_{mo}) to the volume velocity of the nasal output (U_{no}). This transfer ratio function ($T_{n/m}$) can be directly derived from Equation 4.9 and 4.10 as follows

$$T_{n/m} = \frac{U_{no}}{U_{mo}} = \frac{H_n}{H_m} = \frac{A_m Z_{ml} + B_m}{A_n Z_{nl} + B_n}. \quad (4.18)$$

In the original study, $T_{n/m}$ was derived in a simpler way without explicitly calculating H_m and H_n, since $T_{n/m}$ can also be regarded as the ratio between the two transfer admittance

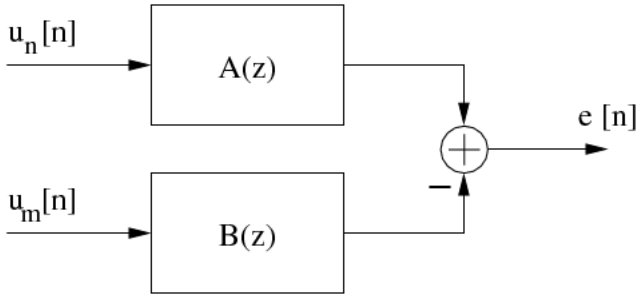


Figure 4.6: System identification method to estimate parameters in $A(z)$ and $B(z)$.

functions from the sound pressure at the VP port (P_v) to the volume velocity at the nose (U_{no}) and at the mouth (U_{mo}), respectively. The oral and nasal transfer admittance functions, $Y_{m/v}$ and $Y_{n/v}$, can be derived directly from the chain matrix equations 4.1 as

$$Y_{m/v} = \frac{U_{mo}}{P_v} = \frac{1}{A_m Z_{ml} + B_m}, \tag{4.19}$$

$$Y_{n/v} = \frac{U_{no}}{P_v} = \frac{1}{A_n Z_{nl} + B_n}. \tag{4.20}$$

Thus the ratio of $Y_{n/v}$ to $Y_{m/v}$ leads to the same result as in Equation 4.18.

The above Oral-Nasal Transfer RatIo Function (ONTRIF, $T_{n/m}$) has the following properties. First, it does not contain any parameters of the pharyngeal tract, the glottal source, and the subglottal configuration, which implies that it is independent of the acoustic system below the VP port. Second, the coupling effect of the oral and nasal tracts is canceled out in $T_{n/m}$, in which all the poles stem from the transfer admittance of the nasal cavity, the oral cavity only introduces zeros, and the effects of sinuses present as pole-zero pairs. Third, there is a mathematically tractable way to estimate $T_{n/m}$, given volume velocity signals from the two separate channels. Assuming $T_{n/m}$ has an ARMA structure in the z-domain, there is a linear solution to estimate the model parameters using a system identification method. The estimation approach is presented as follows.

Suppose we can obtain the volume velocity signals, $u_{no}(t)$ and $u_{mo}(t)$, at the nose and the mouth during the production of a nasalized sound. The signals are sampled at a sampling frequency higher than the Nyquist frequency, so that we have the discrete signals,

$u_{no}[n]$ and $u_{mo}[n]$. Assuming the ONTRIF in z-domain, $T_{n/m}(z)$, has the following form,

$$\begin{aligned} T_{n/m}(z) &= \frac{U_{no}(z)}{U_{mo}(z)} \\ &= \frac{B(z)}{A(z)} \\ &= \frac{b_0 + b_1 z^{-1} + b_2 z^{-2} + \cdots + b_N z^{-N}}{1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_M z^{-M}}, \end{aligned} \tag{4.21}$$

in which the order N and M are known, then $u_{no}[n]$ and $u_{mo}[n]$ can be regarded as the input signals of the system shown in Figure 4.6. The output signal $e[n]$ of such a system should approximates zero. When L samples from signals $u_{no}[n]$ and $u_{mo}[n]$ are known, the parameters in $A(z)$ and $B(z)$ can be estimated by minimizing the mean square of $e[n]$. The solution to this problem is equivalent to the least mean square error solution of the following set of linear equations. For convenience, writing $u_{no}[i]$ as u_{no}^{i}, and $u_{mo}[i]$ as u_{mo}^{i},

$$\begin{bmatrix} u_{no}^{i+1} \\ u_{no}^{i+2} \\ \vdots \\ u_{no}^{i+L} \end{bmatrix} = \begin{bmatrix} A & B \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_M \\ b_0 \\ \vdots \\ b_N \end{bmatrix}, \tag{4.22}$$

in which

$$A = \begin{bmatrix} -u_{no}^{i} & -u_{no}^{i-1} & \cdots & -u_{no}^{i+1-M} \\ -u_{no}^{i+1} & -u_{no}^{i} & \cdots & -u_{no}^{i+2-M} \\ \vdots & \vdots & \vdots & \vdots \\ -u_{no}^{i+L-1} & -u_{no}^{i+L-2} & \cdots & -u_{no}^{i+L-M} \end{bmatrix}, \tag{4.23}$$

and

$$B = \begin{bmatrix} u_{mo}^{i+1} & u_{mo}^{i} & \cdots & u_{mo}^{i+1-N} \\ u_{mo}^{i+2} & u_{mo}^{i+1} & \cdots & u_{mo}^{i+2-N} \\ \vdots & \vdots & \vdots & \vdots \\ u_{mo}^{i+L} & u_{mo}^{i+L-1} & \cdots & u_{mo}^{i+L-N} \end{bmatrix}. \tag{4.24}$$

Poles and zeros can then be determined by calculating the roots of $A(z)$ and $B(z)$, respectively.

It is to be noted that volume velocities may not be measured easily. Instead, sound pressures can be conveniently picked up with microphones. Another transfer function is defined as the ratio of the sound pressure radiated from the nose (P_{nr}) to the sound pressure radiated from the mouth (P_{mr}). This function involves radiation impedances Z_{nr} and Z_{mr} at the positions of measurement. It can be expressed as

$$T^p_{n/m} = \frac{P_{nr}}{P_{mr}} = \frac{U_{no}Z_{nr}}{U_{mo}Z_{mr}} = T_{n/m}\frac{Z_{nr}}{Z_{mr}}. \tag{4.25}$$

Radiation impedances can be modeled as a derivative process with only a spectral zero at zero Hz (Flanagan, 1972), so $T^p_{n/m}$ has the same pole-zero pattern as $T_{n/m}$. The estimation of $T^p_{n/m}$ can be the same as that of $T_{n/m}$, after replacing the volume velocity signals with sound pressure signals, $p_{nr}[n]$ and $p_{mr}[n]$.

The idea of the ONTRIF analysis described above is similar to that of LPC analysis. They both assume a parametrized transfer function for the model, and adopt a linear solution. One difference is that the ONTRIF is in ARMA form, and two separate channels of signals are assumed to be available. The order of the transfer ratio function, M or N, can be chosen in the same way as in LPC analysis. For example, when the sampling frequency is 10 kHz, M and N can be set to 12. L is the window size for the analysis, which can be either fixed or pitch synchronized in real applications.

4.4 Articulatory simulation

4.4.1 Articulatory synthesizer

In order to validate the ONTRIF analysis introduced in the previous section, an articulatory synthesizer was built to simulate the production of voiced sounds. The synthesizer was implemented according to the transmission-line model in Figure 2.2, except that it assumed the subglottal impedance to be infinity for simplicity. The ONTRIF analysis were applied to synthesized signals of nasal and oral volume velocities. Since the articulatory configuration is known before synthesis, the real ONTRIF can be directly calculated from the circuit network. Thus the analysis method can be effectively evaluated by comparing the estimated ONTRIF with the pre-calculated one.

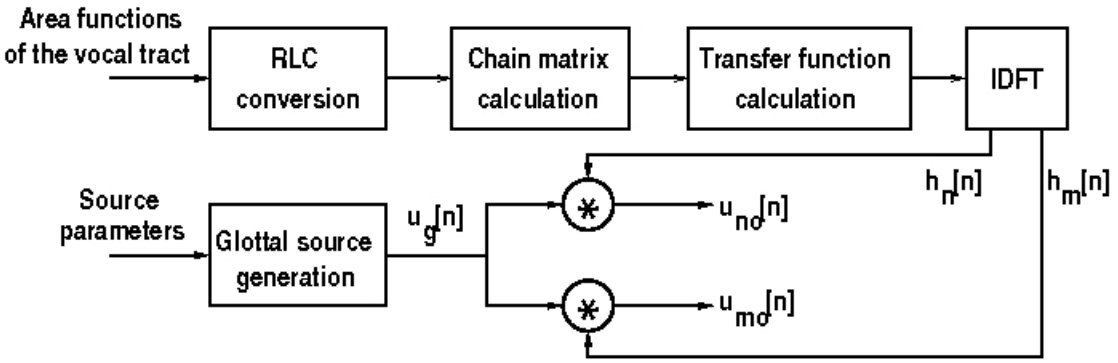


Figure 4.7: Functional modules and I/O flows of the articulatory synthesizer.

Figure 4.7 shows the functional modules and I/O flows among them in the synthesizer. As the input of the synthesizer, parameters of the glottal source and two area functions of the vocal tract are specified. The glottal source signal $u_g[n]$ is generated from the LF model (Fant, 1986). The area functions of the pharyngeal-oral tract are based on existing X-ray data of vowels (Fant, 1960), which specifies the cross-section areas at 0.5 cm intervals from the glottis to the lips. Inside the pharyngeal-oral tract, the VP port is at the 8.5 cm distance from the glottis. The velum can be set open or closed for the simulation of nasalized or non-nasalized vowels. For the nasal tract, an empirical model is adopted (Maeda, 1982b), which assumes the nasal tract to be 11 cm long and sampled at 1 cm intervals. In the nasal tract, The cross-section area of the first three sections from the velum can be changed to simulate different velar positions, while the areas of other sections are constant. When the area of the first section from the velum is specified, the areas of the next two sections are calculated by linear interpolation between the first and fourth sections. One sinus is simulated for simplicity. The sinus cavity is connected with the nasal tract at the 7 cm distance from the velum. It is modeled as a Helmholtz resonator that has a cavity of 20.8 cm^3 in volume and connected to the nasal cavity through a 0.5 cm long tube with the cross area of 0.1 cm^2. The acoustic characteristic frequency of such a resonator is about 550 Hz.

During synthesis, the area functions and the dimension of the sinus are converted into equivalent circuit parameters of tube units in the transmission-line model by the "RLC conversion" module (RLC stands for acoustic resistance, inductance and capacitance).

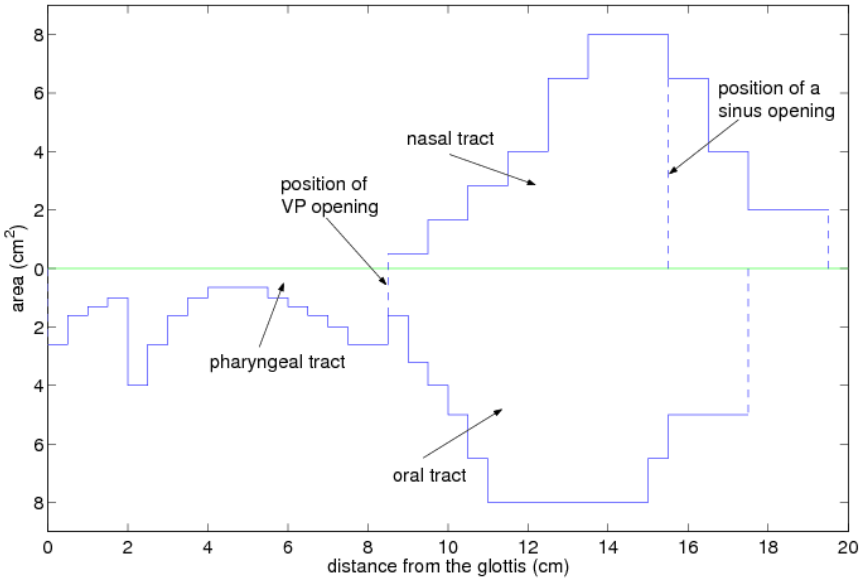


Figure 4.8: Cross-section area functions of vocal tract for synthesizing a pure vowel /aa/ and its nasalized counterpart /aa-n/. The area function of the pharyngeal-oral tract is plotted downward. The nasal tract starts from the VP opening and its area function is plotted upward.

Then the chain matrix of each sub-net is constructed from the units, and the oral and nasal transfer functions (U_{no}/U_g and U_{mo}/U_g) are calculated in frequency-domain. The impulse responses of the two transfer functions, $h_m[n]$ and $h_n[n]$, are obtained by Inverse Discrete Fourier Transform (IDFT), and they are convolved with the glottal source signal, $u_g[n]$. The resulting two signals, $u_{mo}[n]$ and $u_{no}[n]$, are the volume velocity signals out of the mouth and the nose. The final output of volume velocity is $u_{mo}[n]$ for an ideal vowel, $u_{mo}[n]$ for an ideal nasal, and the sum of $u_{mo}[n]$ and $u_{no}[n]$ for a nasalized vowel. The radiated sound pressures signal is obtained by applying a derivative filter to the final volume velocity output.

4.4.2 Experiments and results

4.4.2.1 Evaluation of synthesized speech

In the first experiment, sample waveforms were synthesized to validate the basic functionality of the articulatory synthesizer. An ideal vowel /aa/ and its nasalized counterpart /aa-n/ were synthesized with the synthesizer. The cross-section area function for /aa/, as

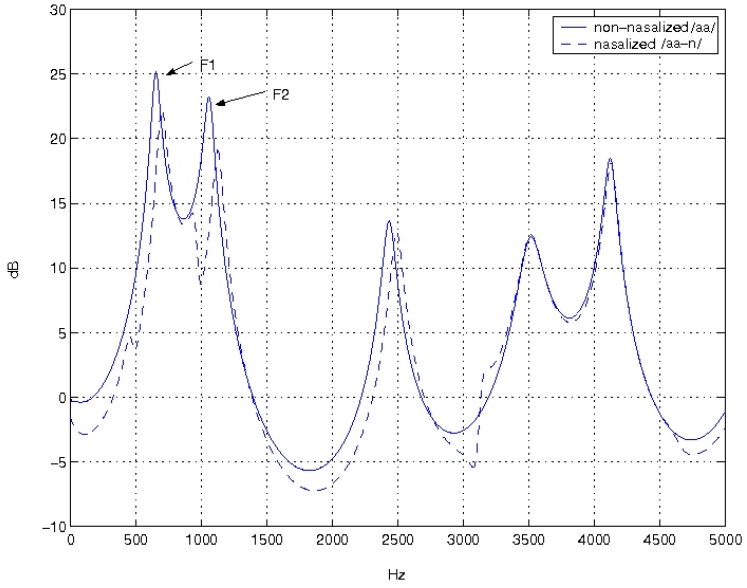


Figure 4.9: Power spectra of the oral transfer functions (U_{mo}/U_g) for synthesizing a pure vowel /aa/ and its nasalized counterpart /aa-n/.

shown in Figure 4.8, was converted from X-ray data (Fant, 1960). There were no available X-ray data of the same speaker to specify the cross-section area function of the pharyngeal-oral tract for /aa-n/, so the same area function for /aa/ was re-used by assuming that the velar movement has relatively little influence on the cross-section of the pharyngeal-oral tract. It is noted that this assumption holds better in a low vowel case than in a high vowel case. The opening area of the VP port were set to $0.5\,\mathrm{cm}^2$ for /aa-n/, and the corresponding area function of the nasal tract is also shown in Figure 4.8. The LF parameters for $u_g\,[n]$ were set to typical modal voice values. The pitch of the impulse was set to 120 Hz, and the sampling frequency of the signal was 10 kHz.

In the above two samples, oral transfer functions (U_{mo}/U_g) can be calculated directly by the synthesizer from the pre-set vocal tract configurations of the two phonemes. Figure 4.9 shows the log-magnitude spectra of U_{mo}/U_g in the two cases, in which the solid curve represents /aa/ and the dashed curve represents /aa-n/. The formant pattern of the pure vowel /aa/ can be easily identified from solid curve, with F_1 around 700 Hz and F_2 around 1,100 Hz. It can be seen on the nasalized spectrum that the first three formants shift

higher from their positions on the non-nasalized spectrum. In addition, extra pole-zero pairs appear at about 500 Hz, 1,000 Hz, and 3,100 Hz on the nasalized spectrum. It can be shown by calculating the transfer admittance function of the nasal tract that the pole-zero pair around 500 Hz is caused by the sinus cavity, and the other two pairs of pole-zeros result from the first and second resonances of the nasal tract, respectively. The figure also shows significant amplitude decreases in the low formant range of the nasalized vowel comparing to the non-nasalized one. These spectral characteristics confirm most of the observations of nasalization effects in real speech (Chen, 1997). The samples indicate that our synthesiser is capable of simulating major acoustic effects of nasalization.

4.4.2.2 ONTRIF analysis

In the second experiment, the nasal-oral transfer ratio function (ONTRIF, U_{no}/U_{mo}) was estimated from the synthesized signals (U_{no} and U_{mo}) of the nasalized sample /aa-n/ in the previous experiment. It was then compared with the pre-set one that was directly calculated from the articulatory configuration. During the calculation, the transfer admittance functions of the nasal cavity (U_{no}/P_v) and of the oral cavity (U_{mo}/P_v) were first obtained from the chain matrices of the nasal and oral tracts; then the ratio between them was calculated.

Figure 4.10 shows the power spectra of the two transfer admittance functions. The solid curve (U_{no}/P_v) in the figure represents the properties of the nasal tract without any influence of other parts of the vocal tract. Each peak on the solid curve is a pole of the nasal admittance, which consequently results in a pole-zero pair on the power spectrum of the oral transfer function of the nasalized vowel (dashed curve in Figure 4.9). The dashed curve (U_{mo}/P_v) in Figure 4.10 represents the spectral properties of the oral tract (from the VP port to the lips). It has spectral peaks higher than $F2$ of the corresponding vowel due to the short length of the oral tract. The pole pattern in the ONTRIF mainly reflects the property of the nasal tract, while the relatively simple influence of the oral tract presents as the zero pattern in relatively high frequency range.

According to the derivation in Section 4.3, the power spectrum of the ONTRIF is the difference between the spectra of nasal and oral admittances, as shown in Figure 4.11 (in

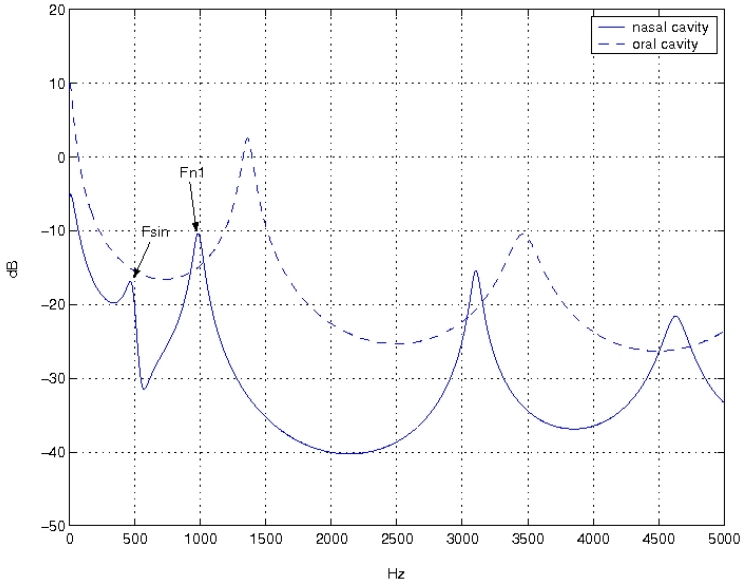


Figure 4.10: Power spectra of the transfer admittance functions of the nasal tract (U_{no}/P_v) and the oral tract (U_{mo}/P_v) during the production of the nasalized vowel /aa-n/.

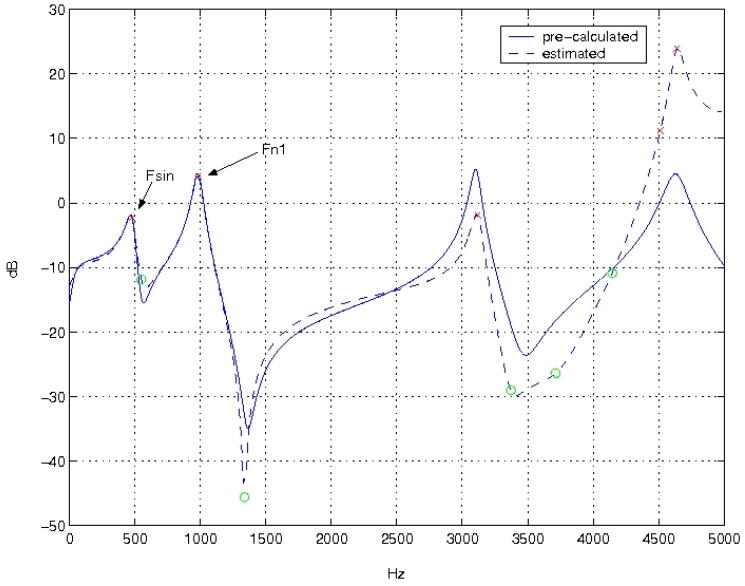


Figure 4.11: Power spectra of the pre-calculated and estimated oral-nasal transfer ratio functions of the nasalized vowel /aa-n/.

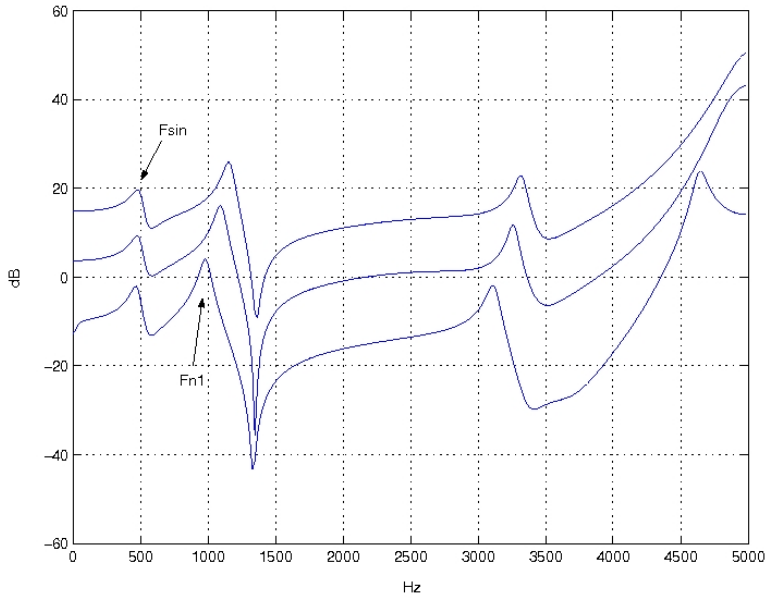


Figure 4.12: The estimated spectra of the ONTRIFs for three nasal configurations. The area of the VP opening increases from the lowest curve up. A shift of 0, 20 and 40 dB has been added to the curves respectively for a better display.

solid curve). Therefore, the spectral peaks of the ONTRIF are the same as those of the nasal admittance. After applying the ONTRIF analysis at the center of the synthesized outputs of $u_{mo}[n]$ and $u_{no}[n]$ for the nasalized sample, the power spectrum of the estimated ONTRIF is also drawn in Figure 4.11 (in dashed curve) for the purpose of comparison. It can be seen that this estimated spectrum matches the pre-calculated one quite well especially in the low frequency range and their peak-valley locations are almost the same.

4.4.2.3 Degree of VP opening

In the third experiment, three different configurations of the velum positions are studied. During synthesis, the area of the VP opening was set to $0.5\,\mathrm{cm}^2$, $1.5\,\mathrm{cm}^2$, and $2.5\,\mathrm{cm}^2$, while the area function of the pharyngeal-oral tract was kept the same as that for /aa/. The output signals of each configuration were then analyzed to estimate their ONTRIFs. The estimated ONTRIF spectra of the three configurations are shown in Figure 4.12. It can be seen that the pole caused by the sinus (F_{sin}) remains at the same position in the three spectral curves, and the first frequency peak caused by the whole nasal cavity (F_{n1})

increases as the area of the VP opening increases. This result implies that the estimated value of F_{n1} could be a possible correlate to the degree of the VP opening.

4.5 Summary

This chapter provides a theoretical base for the dual-channel acoustic analysis of nasalization. The studies were based on an electrical circuit analogue to the acoustic process of speech production. Similar analog approach has been developed by the pioneers in speech sciences (Fant, 1960; Flanagan, 1972). We adopted the circuit model to gain insight into the acoustic process of nasalization. The transmission-line model introduced in this chapter not only helps us to understand the intrinsic complexity of nasalized sounds, but also provides a way to simulate and test the acoustic properties of them.

In order to avoid the ambiguities stemming from the mixture of dual-channel acoustic waves, and to reduce the coupling effects among different parts in the vocal tract, we proposed a novel method to analyze the spectral properties of the nasal cavity by estimating the oral-nasal transfer ratio function (ONTRIF) from the volume velocity output of the oral channel to that of the nasal channel. Theoretical derivation showed that the ONTRIF spectra reflect the characteristics of the nasal tract in a simpler form. A practical algorithm was presented to perform ONTRIF analysis on dual-channel signals. Simulation experiments validated the algorithm and also indicated that the presented method might be used to extract detailed information about the VP opening during speech. This method provides a novel way to perform detailed analysis of nasalization on short-time dual-channel acoustic signals. It paves the way for our further research on applying the analysis method to real dual-channel data and extracting acoustic features of nasalization.

It is to be noted that the real nasal tract is more complicated than what is represented by the model. For example, both the asymmetry of the two passages in the nose and the existence of other sinuses can introduce more pole-zero pairs into the transfer admittance function of the nasal cavity. The trans-velum acoustic vibrations are not modeled, either. In addition, VP movements can change the dimension of the oral cavity. All these effects will be reflected in the ONTRIF spectra. Moreover, because of the difficulty in picking up

volume velocity signals directly, the ONTRIF analysis is more likely to be performed on sound pressure signals in practical usage. As mentioned in Section 4.3, the validity of the same analysis method involves certain assumptions about radiation impedances. Finally, the sound pressure signals radiated from the nose and the mouth may not be perfectly separable from each other by the device to pick up them. Therefore, the pattern of the ONTRIF spectrum estimated from real speech signals may not be as distinct as those obtained from the synthesized signals.

It is to be answered whether the theoretical advantages of dual-channel acoustic analysis presented in this chapter will provide gains over conventional single-channel analysis of nasalization, which is the topic of the following chapter.

Chapter 5

Dual-channel acoustic features for nasalization detection

Based on the theoretical analysis of the dual-channel acoustic model in the last chapter, this chapter presents an algorithm to extract a nasalization feature from real dual-channel acoustic signals that are picked up by a simple two-microphone setup. The feature is then used to build a nasalization detector that automatically classifies sounds into broad categories that reflect different oral-nasal coupling conditions. The performance of the feature is studied by comparing the nasalization detector in parallel with a reference one that adopts the conventional single-channel spectral-based feature.[1]

5.1 Introduction

Different oral-nasal articulatory configurations result in different categories of sounds. Vowel-like sounds have an open oral tract and a closed nasal tract, while nasals are opposite. Nasal vowels or nasalized vowels have both tracts opened. It is desirable to develop indirect analysis techniques that can capture the information about these oral-nasal coupling conditions, since it is difficult to monitor the VP port directly. This information can be used to analyze, assess or enhance certain disordered speech with nasal resonance problems, or to improve the discriminating capacity of recognition systems.

[1] Some materials in this chapter have been published in a conference paper (Niu and van Santen, 2007).

The convectional acoustic features used in current state-of-art automatic speech recognition (ASR) systems are Mel-Frequency Cepstral Coefficients (MFCC) or Perceptual Linear Prediction (PLP) parameters. They are both discrete representations of smooth spectral envelopes of short-time acoustic signals, which take advantage of certain knowledge about human perception mechanisms. These convectional acoustic features do not explicitly characterize specific cues about speech production, but their general discriminating capacities among normal phonetic units in a language are best known for many ASR tasks nowadays. In these systems, statistical models, such as the Hidden Markov Model (HMM), are trained with huge amounts of acoustic data of each phonetic unit in order to capture the accurate distribution of the feature and the temporal dependency.

Some groups of researchers have developed "knowledge-based" speech recognition systems, in which nasalization can be one of the phonological features. Some quantitative nasalization features based on the knowledge of spectral characteristics of nasalization have been tried either as stand-alone features (Glass and Zue, 1985; Chen, 2000; Pruthi and Espy-Wilson, 2006), or as additional features combined with conventional features such as MFCC (Hasegawa-Johnson *et al.*, 2005). However, as reviewed in Chapter 2, because those spectral characteristics of nasalization observed from the acoustic signals of nasalized vowels turn out to be inconsistent and variant, the corresponding features have not exhibited considerable advantages over conventional features in ASR systems. One possible reason for the difficulty, as analyzed in Chapter 4, may be the inherent limitation of single-channel acoustic analysis. However, articulatory-specific features may still be very useful for analyzing disordered speech, because the primary interest is in identifying physiological configurations instead of the average recognition rate of phoneme sequences.

Based on the experimental results and analyses of multi-channel measurement in the previous chapters, our efforts in this chapter are devoted to develop novel analysis approaches for the purpose of automatic detection of nasalization conditions by using information from an extra acoustic channel. Although the dual-microphone setup is rarely used in the current speech recognition community, it has been used in many clinical research works and applications.

5.2 Proposed features

5.2.1 Generalized ONTRIF analysis

In the last chapter, we derived the frequency response of a special transfer function from the volume velocity of the oral output (U_{mo}) to the volume velocity of the nasal output (U_{no}), and called it the Oral-Nasal Transfer RatIo Function (ONTRIF) of volume velocity. We also developed an approach to estimate the ONTRIF with two separate signals of nasal and oral volume velocities. It was also shown that the ONTRIF of sound pressure can be estimated in the same way when the available signals are the separate sound pressures radiated from the nose and the mouth, and the spectral properties of the transfer ratio function approximately remain the same. Since sound pressure can be measured easily in practice, the ONTRIF analysis in this chapter only deals with real dual-channel sound pressure signals that are picked up by a pair of microphones.

The ONTRIF was originally derived under the condition of nasalized vowels when there are output signals from both nasal and oral channels in the model. It is undefined when there is no output in any channel under ideal conditions of oral-only or nasal-only sounds. However, it has been observed consistently in real dual-channel data that a substantial amount of sound pressure signal can be captured by the nasal microphone during the production of "pure" vowels, even when the two channels are well separated. One of our experiments (in Section 3.2) with the NasalView device has confirmed these observations. A reasonable explanation of this phenomenon is the assumption of cross-velum vibrations (Suzuki *et al.*, 1990). This fact indicates that a modification of the dual-channel transmission-line model is needed for "pure" vowel production, in which a special two-port circuit can be adopted to model the transmission characteristics of the closing velum. Therefore, the ONTRIF analysis can still be applied to the dual-channel signals of vowels. The resulting transfer function will be influenced by the front oral tract, the nasal tract closed at the velum end, and the physical property of the velum. Similarly, real oral-channel signals are also measurable during the production of "pure" nasal phonemes. This effect can be represented as a radiation load at the oral-channel output, so that the ONTRIF analysis can also be applied in the "pure" nasal condition. The corresponding

ONTRIF spectrum is determined by the opened nasal tract and the closed oral cavity with the radiation load.

The above argument paves a way for generalizing the ONTRIF analysis method to the dual-channel acoustic data of all different oral-nasal coupling conditions. It is hypothesized that the ONTRIF spectrum discriminably characterizes each category of these conditions, so that nasalization features may be efficiently extracted from the short-time ONTRIF spectrogram of speech.

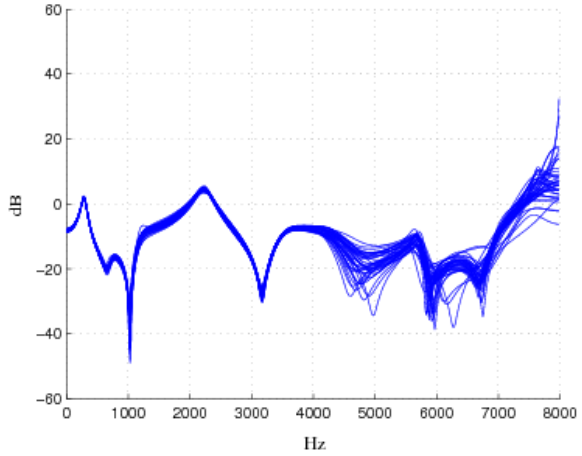
5.2.2 Preliminary studies

In order to study the property of ONTRIF spectra in different nasalization conditions, we analyzed some real dual-channel sound pressure samples recorded by Feng and Kotenkoff (2006), who used microphones in a double-chamber sound proof box to achieve separate recordings of nasal and oral signals.[2] The acoustic attenuation across the two channels was reported as at least 20 dB for all frequencies below 8 kHz.

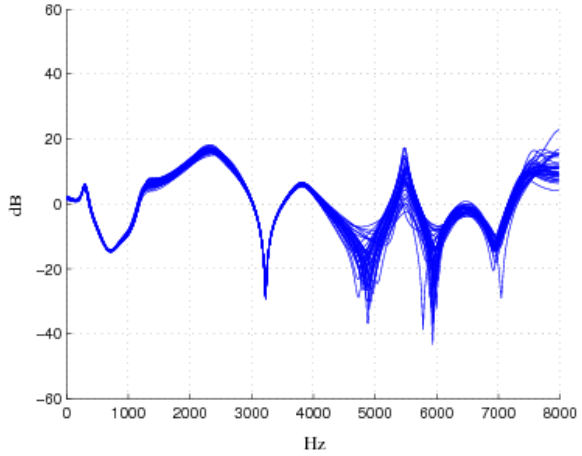
When applying the ONTRIF analysis to these samples, the transfer ratio functions were obtained through a frame-based processing. The dual channel signals were segmented into frames with a fixed length of 20 ms and a frame shift rate of 10 ms. For each frame of oral and nasal signals, the parameters of a ARMA filter was estimated. The order of the filter was set to 18 for both numerator and denominator polynomials while the signals were down-sampled to 16 kHz.

Figure 5.1 shows the log-magnitude frequency responses of the ONTRIFs estimated from the stable sections of a vowel /ao/, a French nasal vowel /aa-n/, and a nasal /ng/. For each phoneme, the transfer ratio functions of 40 successive frames are overlappingly plotted. In each of the sub-figures, the log-magnitude curves vary little among frames within the low frequency range (about 0-4000 Hz). The variances among frames are relatively larger within the high frequency range, especially for the nasal phoneme. A pole-zero pair in the range of 300-700 Hz exists in each sub-figure, which may be introduced by the shunt effect of sinuses to the admittance of the nasal cavity. The zero pattern of the nasal phoneme

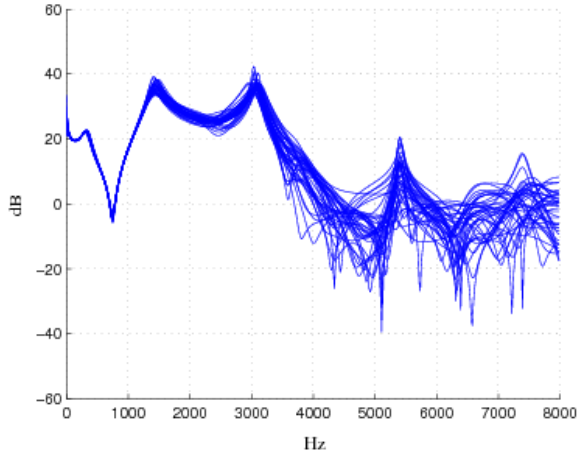
[2] Thanks to the authors for sharing their data samples with us.



(a) Vowel: /ao/



(b) Nasal vowel: /aa-n/



(c) Nasal: /ng/

Figure 5.1: The transfer ratio functions from the oral sound pressure to the nasal sound pressure.

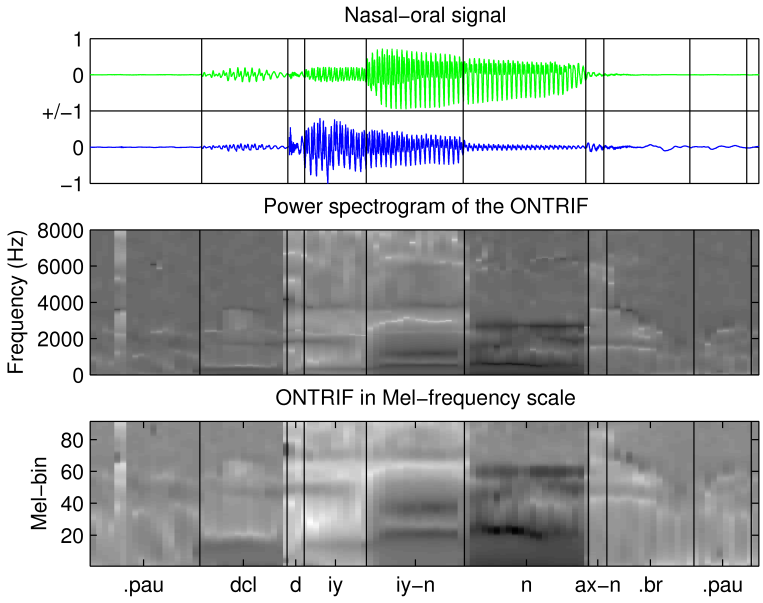


Figure 5.2: ONTRIF spectrogram of a word, "dean". From top to bottom: Nasal signal, oral signal, power spectrogram of the ONTRIF, power spectrogram in Mel-scale. Vertical lines represent manual segmentation.

is quite different from that of the vowel and the nasal vowel because of the closure of the mouth. It can be seen in sub-figures (a) and (b) that the admittance of the vowel cavity causes zeros at about 1000, 3200, 4900 and 6000 Hz, while the opening of the VP port mainly cancels the zero at 1000 Hz. It seems that a feature that represents the shape of the transfer ratio function may be used to discriminate these three different categories of oral-nasal configurations; and the lower frequency range of the frequency response contains more detailed information about nasalization.

There is a question to be answered concerning the practical usage of dual-channel analysis of nasalization. It is whether a simpler device, like the NasalView, that acquires not-perfectly-separated dual-channel data can be used to perform a similar analysis based on the above transfer ratio function approach. We propose to answer this question in an empirical way. That is, we will design a nasalization detector using features extracted from the not-perfectly-separated dual-channel speech signals as input and test how well it performs.

As an example, Figure 5.2 shows the ONTRIF spectrogram of a sample word ("dean")

recorded with the NasalView device. The oral and nasal signals are both down sampled to 16 kHz and segmented into frames of 20 ms length with a frame shift of 10 ms. The ONTRIF of each frame was estimated in the same way as the previous experiment, and the resulting frequency response composes the spectrogram. It can be seen that detailed spectral differences of different oral-nasal articulatory configurations (vowel /iy/, nasalized vowel /iy-n/, and nasal /n/) are visible in the low frequency range, which is more distinct when the spectra are converted into Mel-scale. Visual examination of the spectrograms of other sample signals also indicates that the ONTRIF obtained from the NasalView device contains information that can discriminate different oral-nasal configurations. Therefore, it is expected that robust nasalization features can be extracted from the ONTRIF analysis of the NasalView signals.

5.3 Nasalization detector

5.3.1 Feature extraction algorithm

According to the above observations and analyses of the real dual-channel speech samples, the magnitude responses of ONTRIFs characterize different oral-nasal coupling conditions. In order to build an automatic detector of these nasalization conditions, we propose the following algorithm to extract nasalization features from the signals.

First, the digital sound pressure signals recorded from the oral and nasal channels are high-pass filtered ($F_c = 50\,\text{Hz}$) simultaneously to eliminate any possible DC and low-frequency noise that may be captured during data acquisition. Next, the signals are segmented into equal-length short-time frames with a fixed frame shift. For each pair of oral and nasal frames, a set of parameters of the ONTRIF, $T^p_{n/m}(z)$, is estimated using the method introduced in Section 4.3. Then the frequency response of the ONTRIF is evaluated at frequencies evenly located from 0 Hz to the half of the sampling frequency, resulting in a discrete version of the magnitude response, $\left|T^p_{n/m}[k]\right|^2$, in which k is the index of sampling frequencies. The logarithm of the discrete magnitude response is calculated, and processed with a bank of triangle filters. These filters are generated with Skowronski and Harris' algorithm (2003), as shown in Figure 5.3. The centers of the filters

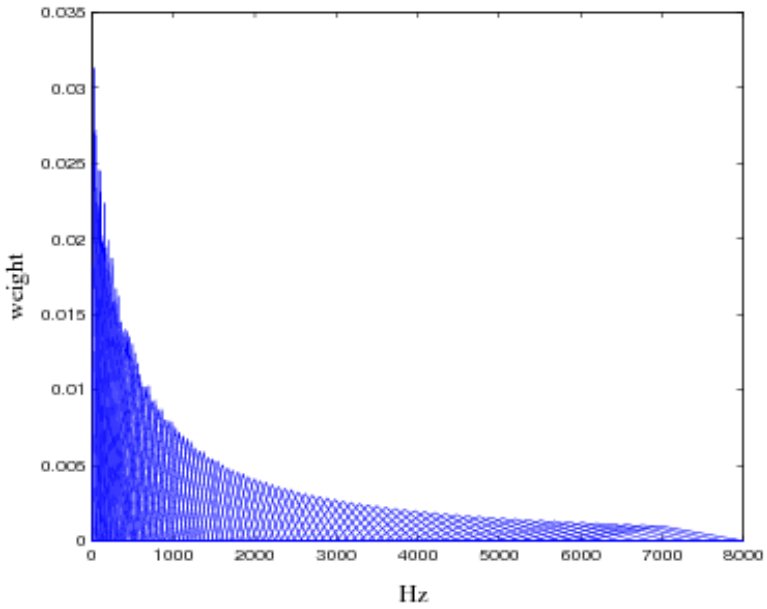


Figure 5.3: Mel-scale ERB filter bank. In this example, there are 91 triangle filters covering the frequency range from 0 to 8,000 Hz.

are evenly spaced along the Mel-scale, and the band of the each filter is determined by the equivalent rectangular bandwidth (ERB) function (Moore and Glasberg, 1983) at its central frequency. The amplitudes of the filters are normalized by scaling them to a constant energy. A set of Mel-scaled features $M[i]$, are obtained by adding the log-magnitude values weighted by each triangle filter, in which i is the index of the Mel-bins. Then the type-II discrete cosine transform (DCT-II) is applied to $M[i]$, obtaining a set of cepstral coefficients, $C[j]$, in which j is the index of the j-th component. The procedure iterates until the cepstral coefficients of all framed are obtained. These coefficients are used as nasalization features of each frame.

The idea behind the above algorithm is to use a small set of coefficients to represent the shape of the magnitude response of the ONTRIF. The procedure is similar to the algorithm for MFCC extraction. The sampling operation discretized the magnitude response. The Mel-scaling operation reduces the frequency resolution unevenly, with higher resolution in the low-frequency range than that in the high-frequency range. The DCT operation compresses the Mel-scaled features to a lower dimensional range, and obtain a set of approximately uncorrelated components.

5.3.2 Nasalization detector

In order to make usage of the above nasalization features, a nasalization detector is designed to discriminate three different oral-nasal coupling conditions during speech, namely vowel-like (Vo), nasal (Ns), and nasalized vowel (Nv) conditions. A simple design is to adopt a Bayesian classifier, in which the conditional probability distribution of nasalization features given each class, $p(x/s)$, is modeled by a Gaussian PDF or a Gaussian mixture model (GMM). The prior of each nasalization condition is assumed to be the same, so that the Bayesian decision rule is simplified as

$$S^* = \arg\max_{s_j} \left[p\left(x/s_j\right) \right], \; s_j \in \{Vo, Ns, Nv\}, \tag{5.1}$$

in which x is a vector of nasalization features of a frame, and s_j is one of the nasalization condition to be decided. When the detector is used to determine the class of a segment of speech, the feature vectors of all the frames in the segment can be simply assumed to be independent of each other, thus the class conditional probability of the whole segment (or token) is the multiplication of each frame's class conditional probability.

The classifier can be trained with labeled vectors of nasalization features. Given the training data, the mean vector and covariance matrix of a Gaussian PDF can be estimated directly, or the parameters of a GMM can be estimated iteratively with the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977).

5.4 Experiments

5.4.1 Speech materials

In order to explicitly distinguish the three nasalization conditions during speech, we design a group of special words in the forms of NVN and CVC. In these words, N is a nasal chosen from /m/, /n/ and /ng/; V is a vowel chosen from /iy/, /ae/, /aa/ and /uw/; and C is a plosive chosen from /t/, /d/, /p/, /b/, /k/ and /g/. Since there are no phonemically contrastive nasal vowels in American English, and nasalized vowels result from context effects, we assume that the vowels in the NVN group are fully nasalized. While in the CVC group, we deliberately insert vowels between plosives, so it is reasonable to assume

that they are not nasalized by typical speakers. There are 48 words in total for a session of recording, which contains 24 vowels and 24 nasalized vowels. During the recording procedure, each word is inserted in the carrier sentence, "Say _ please".

The data were recorded by 3 male and 3 female native American adult speakers. Each speaker made 3 repetitions of the recording sessions. The NasalView device was used to collect the dual-channel acoustic signals. The signals were amplified and recorded simultaneously to the hard drive of a computer through two channels of a M-Audio Duo USB Audio Interface. Waveforms were sampled at 16 kHz and stored in 16-bit PCM format. All recordings were made in a quiet room. At the beginning of each recording session, the gains of two channels were calibrated to the same level. After recording, phoneme boundaries of the words in the sentences were manually marked. The boundaries were decided by visual inspection of energy changes and spectral discontinuities of the signals in both channels.

In order to compare the dual-channel features with conventional single-channel features, we generated a pseudo-single-channel corpus by arithmetically adding up the recorded dual-channel data. Auditory examination of the resulting signals showed no perceptual difference from the acoustic signals recorded with a single microphone.

Nasalization features were extracted from the dual-channel data according to the algorithm presented in Section 5.3.1. The frame length for feature extraction was 20 milliseconds, and the frame shift was 10 milliseconds. For each frame of the signals, a 25-dimensional vector of coefficients was calculated. For the purpose of comparison, a 25-dimensional MFCC vector was also calculated from the corresponding frame of the pseudo-single-channel signal. Both dual-channel and single-channel feature vectors were grouped for each vowel, nasal and nasalized vowel tokens according to manual segmentation labels.

5.4.2 Experimental design

The main purpose of the experiments is to examine whether the proposed feature, extracted from the dual-channel data, can reliably discriminate the three different oral-nasal coupling conditions during speech. Both dual-channel features and single-channel features were used in parallel to train classifiers and to test recognition rates for the purpose of

		Classifier B	
		Correct	Wrong
Classifier A	Correct	(+A, +B)	(+A, -B)
	Wrong	(-A, +B)	(-A, -B)

Table 5.1: Classification results of two classifiers. There are four combinations of the results that two classifiers can make decisions about one data sample.

comparison. Two recognition tasks were designed: one is speaker-dependent, the other is speaker-independent.

In the *speaker-dependent (SD)* task, two sessions of each speaker's data were used to train a classifier and the unused session was left for testing. There were 3 combinations for each speaker, adding up to 18 sub-tasks. Due to the amount of training data of each sub-task, the conditional distribution of each class was simply modeled with a single Gaussian PDF.

In the *speaker-independent (SI)* task, one speaker's data were left out for testing, while the data of other five speakers were used to train the classifier. The speakers were tested successively in 6 sub-tasks. In each sub-task, a GMM was trained for each oral-nasal condition of the classifier. Preliminary validation experiments within training sets suggested 4 components of the GMM performed best on average, given the existing amount of training data, so the number of mixtures of each GMM was fixed to 4.

5.4.3 Results and discussion

In both tasks, the frame classification rate (FCR) and token classification rate (TCR) were calculated by comparing decisions made by the classifiers with manual labels. The classification rate serves as an indicator of how well the feature can discriminate nasalization conditions. McNemar's test, suggested for benchmark tests of speech recognition systems (Gillick and Cox, 1989), was applied to evaluate the significance of the difference between classification results of the classifiers using dual-channel features and single-channel features, respectively. Assuming the classifier using dual-channel features is A and the classifier using single-channel features is B, for each sample of testing data, the classification result will be one of the pairs which is represented in Table 5.1. Essentially,

		Dual			Single			Number
		Vo	Nv	Ns	Vo	Nv	Ns	of samples
FCR (%)	Vo	**97.38**	1.32	1.30	96.37	2.10	1.53	8104
	Nv	1.06	**92.54**	6.41	5.53	85.73	8.74	10610
	Ns	0.17	1.06	**98.77**	1.92	3.14	94.94	11044
Avg. FRA (%)		**96.23***			92.35			total: 29758
TCR (%)	Vo	**98.84**	0.93	0.23	97.77	0.23	0.00	432
	Nv	0.00	96.75	3.25	2.32	**96.98**	0.70	431
	Ns	0.00	0.23	**99.77**	0.93	0.93	97.20	858
Avg. TRA (%)		**98.45**			97.99			total: 1721

Table 5.2: Speaker-dependent (SD) task. Confusion matrices of frame and token classification rates are obtained from SD classifiers trained with dual-channel and single-channel features, respectively. The total numbers of testing samples are also listed. A star (*) means the difference is significant.

McNemar's Test is a Sign-Test. All (+A, +B) and (-B, -B) pairs are ignored and it is tested whether (+A, -B) is as likely as (-B, +A) by labelling the former as + and the latter as - and performing a Sign-Test on the number of + and - labels.

Table 5.2 shows the classification results obtained in the SD task. The testing data samples from all the speakers are presented together. Confusion matrices of FCRs and TCRs are compared between dual-channel features and single-channel features. In the table, the row symbols (Vo/Nv/Ns) are the manually labeled oral-nasal conditions, while the column symbols are classification results decided by the classifiers. The total numbers of frame and token samples for testing are listed for each class. The diagonal element in each confusion matrix is the correct classification rate of a certain oral-nasal condition. Comparing the correct FCRs or TCRs between dual-channel features and single-channel features, the higher one is represented in bold in the table.

The recognition accuracy is calculated as the average of the correct classification rates. For dual-channel features, the average frame recognition accuracy (FRA) is 96.23%, and the average token recognition accuracy (TRA) is 98.45%; while for single-channel features, the average frame recognition accuracy is 92.35%, and the average token recognition accuracy is 97.99%. McNemar's tests show that the difference of frame recognition accuracies between dual and single features is significant at the 0.001 level, while the difference of token classification accuracy is not significant ($p = 0.028$) at the same level.

		Dual			Single		
		Vo	Nv	Ns	Vo	Nv	Ns
FCR (%)	Vo	**92.97**	6.40	0.63	78.88	15.12	6.01
	Nv	5.74	**71.81**	22.45	48.24	43.28	8.48
	Ns	0.55	8.29	**91.16**	28.91	13.57	57.52
Avg. FRA (%)		**85.31***			59.89		
TCR (%)	Vo	**95.83**	3.94	0.23	78.47	13.43	8.10
	Nv	6.96	**84.69**	8.35	42.92	42.00	15.08
	Ns	0.70	24.13	**75.17**	17.25	11.42	71.33
Avg. TRA (%)		**85.23***			63.93		

Table 5.3: Speaker-independent (SI) task. Confusion matrices are comparable with those in Table 5.2. The total numbers of testing samples are the same to those in Table 5.2. A star (*) means the difference is significant.

Table 5.3 shows the classification results of the SI task. It can be seen that all the correct classification rates of dual-channel features are higher than those of single-channel features. For dual-channel features, the average frame recognition accuracy is 85.31%, and the average token recognition accuracy is 85.23%; while for single-channel features, the average frame recognition accuracy is 59.89%, and the average token recognition accuracy is 63.93%. McNemar's tests show that the differences of both frame and token recognition accuracies between dual and single features are significant at the 0.001 level.

The classification results of the SD task show a moderate advantage of the dual-channel features over the conventional single-channel MFCC features. In the SI task, the performances of both features degrade, but the drop in performance is much larger in the single-channel tests than in the dual-channel tests, indicating the dual-channel features are more robust.

To reduce the speaker dependency of the classification results, we also report the testing results of each individual speaker in Table 5.4 to 5.9, respectively. It can be seen that, in the SD task, dual-channel features of almost all speakers (except Speaker #4) perform better than corresponding single channel features; and in the SI task, dual-channel features perform uniformly better. McNemar's tests are performed on each speaker's data. A star mark in the tables indicates the difference is significant at the 0.001 level. The frame classification differences of all speakers except #4 ($p = 0.407$) in the SD task are significant.

The token classification differences in the SD task are all not significant. In the SI task, all frame classification differences are significant; the token classification differences of three speakers (#1, #3 and #6) are significant. These results further confirm that the dual-channel features are more robust to speaker variations.

5.5 Summary

In this chapter, we developed an algorithm of feature extraction for detecting different nasalization conditions from dual-channel acoustic signals. The algorithm is based on the analysis method of the ONTRIF. We presented a simple way to build an automatic nasalization detector based on the ONTRIF features. A special dual-channel corpus was designed and collected in order to test the performance of the proposed features. The proposed features perform better than the conventional single-channel MFCC features in both speaker-dependent and speaker-independent experiments. These experimental results demonstrate the advantage of ONTRIF analysis for such a specific task.

The results lead to a novel approach to automatically detect nasalization conditions using speech signals that are acquired with a simple dual-channel setup. It is also demonstrated that the extracted nasalization features are informative and robust to speaker variations for practical nasalization analysis and detection systems.

		SD task						SI task						Number of samples
		Dual			Single			Dual			Single			
		Vo	Nv	Ns	Vo	Nv	Ns	Vo	Nv	Ns	Vo	Nv	Ns	
FCR (%)	Vo	**99.11**	0.07	0.82	98.98	0.14	0.89	**99.52**	0.20	0.27	90.71	7.99	1.30	1464
	Nv	0.41	**94.99**	4.60	2.97	94.63	2.40	6.96	**61.69**	31.36	61.74	37.75	0.51	1955
	Ns	0.16	0.23	**99.61**	1.13	1.63	97.24	0.78	3.77	**95.45**	42.03	4.39	53.58	2572
Avg. FRA (%)		**97.90***			96.95			**85.55***			60.68			Total: 5991
TCR (%)	Vo	100.0	0.00	0.00	100.0	0.00	0.00	**98.61**	1.39	0.00	90.28	5.56	4.17	72
	Nv	0.00	100.0	0.00	0.00	100.0	0.00	1.39	**72.22**	26.39	61.11	34.72	4.17	72
	Ns	0.00	0.00	100.0	0.00	0.00	100.0	0.00	6.25	**93.75**	24.31	6.94	68.75	144
Avg. TRA (%)		100.0			100.0			**88.19***			64.58			Total: 288

Table 5.4: Speaker #1

		SD task						SI task						Number of samples
		Dual			Single			Dual			Single			
		Vo	Nv	Ns	Vo	Nv	Ns	Vo	Nv	Ns	Vo	Nv	Ns	
FCR (%)	Vo	**95.36**	3.92	0.73	94.18	4.70	1.12	76.16	23.56	0.28	**89.93**	5.65	4.42	1787
	Nv	2.88	**92.18**	4.94	7.12	81.69	11.19	0.62	**78.02**	21.36	61.52	31.48	7.00	2430
	Ns	0.41	1.54	**98.06**	3.89	2.53	93.58	0.00	6.51	**93.49**	11.44	15.87	72.69	2212
Avg. FRA (%)		**95.20***			89.82			**82.56***			64.70			Total: 6429
TCR (%)	Vo	97.22	2.78	0.00	**100.0**	0.00	0.00	81.94	18.06	0.00	81.94	6.94	11.11	72
	Nv	0.00	**100.0**	0.00	1.39	97.22	1.39	0.00	**97.22**	2.78	50.00	22.22	27.78	72
	Ns	0.00	0.69	**99.31**	2.08	2.08	95.83	0.00	27.78	72.22	6.94	9.03	**84.03**	144
Avg. TRA (%)		**98.84**			97.69			**83.80**			62.73			Total: 288

Table 5.5: Speaker #2

		SD task						SI task						Number of samples
		Dual			Single			Dual			Single			
		Vo	Nv	Ns	Vo	Nv	Ns	Vo	Nv	Ns	Vo	Nv	Ns	
FCR (%)	Vo	97.41	1.75	0.84	**97.66**	1.67	0.67	**96.24**	2.42	1.34	87.39	10.61	2.01	1197
	Nv	0.32	**94.81**	4.87	6.15	83.91	9.94	0.19	**83.33**	16.47	52.37	42.31	5.32	1560
	Ns	0.00	0.75	**99.25**	0.81	4.22	94.97	0.06	3.88	**96.06**	32.70	31.31	36.00	1728
Avg. FRA (%)		**97.16***			92.18			**91.88***			55.23			Total: 4485
TCR (%)	Vo	98.61	1.39	0.00	**100.0**	0.00	0.00	**97.22**	2.78	0.00	86.11	8.33	5.56	72
	Nv	0.00	98.61	1.39	4.17	95.83	0.00	0.00	**97.22**	2.78	48.61	43.06	8.33	72
	Ns	0.00	0.00	100.0	0.00	4.86	95.14	0.00	17.36	**82.64**	27.08	28.47	44.44	144
Avg. TRA (%)		**99.07**			96.99			**92.36***			57.87			Total: 288

Table 5.6: Speaker #3

		SD task						SI task						Number of samples
		Dual			Single			Dual			Single			
		Vo	Nv	Ns	Vo	Nv	Ns	Vo	Nv	Ns	Vo	Nv	Ns	
FCR (%)	Vo	96.71	1.28	2.02	**97.24**	2.55	0.21	**96.71**	0.74	2.55	63.02	36.13	0.85	941
	Nv	0.86	87.76	11.38	4.59	**91.30**	4.11	4.49	67.21	28.30	25.24	**71.70**	3.06	1046
	Ns	0.35	1.75	**97.90**	1.69	3.73	94.58	0.82	2.39	**96.80**	15.73	11.24	73.03	1717
Avg. FRA (%)		94.12			**94.37**			**86.90***			69.25			Total: 3704
TCR (%)	Vo	98.61	1.39	0.00	98.61	1.39	0.00	**98.61**	0.00	1.39	65.28	31.94	2.78	72
	Nv	0.00	91.55	8.45	1.41	**98.59**	0.00	5.63	**81.69**	12.68	14.08	77.46	8.45	71
	Ns	0.00	0.00	**100.0**	1.39	1.39	97.22	1.39	5.56	**93.06**	0.00	7.64	92.36	144
Avg. TRA (%)		96.72			**98.14**			**91.12**			78.37			Total: 287

Table 5.7: Speaker #4

		SD task						SI task						Number of samples
		Dual			Single			Dual			Single			
		Vo	Nv	Ns	Vo	Nv	Ns	Vo	Nv	Ns	Vo	Nv	Ns	
FCR (%)	Vo	**98.88**	0.00	1.12	94.85	3.18	1.98	**100.0**	0.00	0.00	95.02	3.35	1.63	1164
	Nv	0.82	**93.86**	5.32	5.39	89.35	5.26	22.25	**71.74**	6.01	74.47	22.39	3.14	1465
	Ns	0.00	1.32	**98.68**	2.36	2.29	95.35	1.60	34.10	**64.31**	55.14	2.50	42.36	1440
Avg. FRA (%)		**97.14***			93.18			**78.68***			53.26			Total: 4069
TCR (%)	Vo	100.0	0.00	0.00	100.0	0.00	0.00	**100.0**	0.00	0.00	93.06	4.17	2.78	72
	Nv	0.00	97.22	2.78	1.39	**98.61**	0.00	31.94	**68.06**	0.00	77.78	19.44	2.78	72
	Ns	0.00	0.00	**100.0**	0.70	1.40	97.90	2.80	68.53	28.67	38.46	2.10	**59.44**	143
Avg. TRA (%)		**99.07**			98.84			**65.58**			57.31			Total: 287

Table 5.8: Speaker #5

		SD task						SI task						Number of samples
		Dual			Single			Dual			Single			
		Vo	Nv	Ns	Vo	Nv	Ns	Vo	Nv	Ns	Vo	Nv	Ns	
FCR (%)	Vo	**97.36**	0.19	2.45	96.07	0.19	3.74	**96.07**	3.80	0.13	45.91	32.30	21.79	1551
	Nv	0.37	**90.48**	9.15	6.17	78.37	15.46	3.81	**67.92**	28.27	11.33	62.72	25.95	2154
	Ns	0.00	1.09	**98.91**	1.45	5.75	92.80	0.22	5.45	**94.33**	16.73	19.27	64.00	1375
Avg. FRA (%)		**95.58***			89.08			**86.10***			57.54			Total: 5080
TCR ()	Vo	98.61	0.00	1.39	**100.0**	0.00	0.00	**98.61**	1.39	0.00	54.17	23.61	22.22	72
	Nv	0.00	**93.06**	6.94	5.56	91.67	2.78	2.78	**91.67**	5.56	5.56	55.56	38.89	72
	Ns	0.00	0.72	**99.28**	1.44	1.44	97.12	0.00	19.42	**80.58**	6.47	14.39	79.14	139
Avg. TRA (%)		**96.98**			96.26			**90.28***			62.95			Total: 283

Table 5.9: Speaker #6

Chapter 6

Conclusion

6.1 Summary

In this thesis, we have investigated non-invasive dual-channel measurement techniques for capturing the nasalization information during speech production, and we have proposed analysis techniques for extracting the nasal resonance features and applied the techniques to automatic detection of nasal resonance conditions during speech.

Though the nasalization process is simply achieved by velopharyngeal (VP) movements during speech, the aerodynamic and acoustic consequences are complicated because of 1) the variations of VP controlling strategies among speakers, 2) the coupling effects of among different cavity of the vocal tract, and 3) the ambiguities caused by the mixing of signals. Since the techniques for direct observations of the VP port are not easily accessible, it is desirable to develop instrumental devices and analysis techniques that capture and extract the VP information from aerodynamic or acoustic signals. The spectral characteristics of nasalization that were extracted from single-channel acoustic signals with conventional spectral analysis methods are either inconsistent or variant. For the purpose of more informative and reliable measurement and analysis, we explored the possibility of using multi-channel signals during the analysis.

The following contributions were made in this thesis:

- Two methods of multi-channel measurement were investigated. One was the combination of an aerodynamic and an acoustic channel, and the other was a dual-channel acoustic device. A novel aerodynamic device was designed and made to collect nasal

airflow signals. The advantages of the multi-channel measurement over the conventional single-channel measurement were justified: the additional aerodynamic signal did capture nasal emission information; the dual-channel acoustic signals may capture more detailed and accurate information about nasal resonance.

- A dual-channel transmission-line acoustic model was analyzed. The coupling effects were shown by the transfer function of each channel's output. The spectral ambiguity caused by mixing two signals was also shown. A special oral-nasal transfer ratio function (ONTRIF) was derived from the dual-channel model. The derivation revealed nice spectral properties of the ONTRIF.

- An analysis method was presented to estimate the parameters of the ONTRIF, given two output signals of the dual-channel model. An articulatory synthesizer was built to simulate the nasalization process, and the analysis method was validated by the simulated signals.

- The ONTRIF analysis method was generalized and applied to naturally recorded dual-channel acoustic signals that were collected by separated nasal and oral microphones. An automatic Bayesian detector of different nasal resonance statuses was trained and successfully tested.

6.2 Future work

The initial motivation of our studies on nasalization problems was to enhance the intelligibility of disordered speech in an analysis-modification-synthesis framework. One of the obstacles in the way is that the conventional speech analysis techniques are not capable of extracting enough articulatory information that may be used to guide the method of modification. The studies in this thesis focus on the nasalization aspect that occurs in both normal and disordered speech. The techniques developed in this thesis provide novel instrumental and analysis tools to possible applications including clinical assessment and enhancement of disordered nasal resonance, phonetic studies of the nasalization process, and special feature extraction for speech recognition.

As a direct usage of the current research results, it is possible to develop an automatic and objective procedure for the assessment of nasal resonance function based on the technique of the dual-channel ONTRIF analysis. The nasalization detection technique can be combined with the existing measurement of nasalance scores to define more robust measures of nasality. With the information of different nasal resonance conditions in speech, not only a more meaningful "energy-balance" based index but also a "temporal" based index can be automatically obtained. One direction of the future research can be to study the correlation of the new measures with the perceptual assessment of nasal resonance.

The results from our studies on the aerodynamic measurement of nasalization indicate that the nasal emission problems would be better analyzed separately from the resonance problems. Though we did not investigate in this direction more deeply, we expect that an aerodynamic measurement of nasal airflow that is more robust to position of the sensor may help us to detect the weakened segments in the acoustic signals in some disordered speech. Future work in this direction is to improve the robustness of the airflow device we have designed.

From the phonetic study point of view, some findings in our research have shown complex nasal channel effects during regular vowel production. There still lacks a satisfactory model for the velum-transmission phenomena in speech. More dual-channel acoustic data and analysis may provide extra knowledge about the process of vowel production. This knowledge may eventually benefit both synthesis and recognition applications.

From the speech recognition point view, the nasalization feature has been regarded as a useful cue but is difficult to be extracted with conventional signal precessing front-ends. The dual-channel analysis approach implies a new way of feature extraction for a speech recognition system. Though used rarely nowadays, multi-channel acoustic front-end for recognition may be used in some adverse environments, such as in the helmet of a pilot. It is worthy to try similar features that are extracted from multiple channels for the recognition purpose.

Bibliography

Allen, J., M. Hunnicutt, and D. Klatt, 1987, *From text to speech: the MITalk system* (Cambridge University Press).

Awan, S. N., 1996, Development of a low-cost nasalance acquisition system, *Pathologies of Speech and Language: Contributions of Clinical Phonetics and Linguistics*, 211–217.

Baer, T., J. Gore, S. Boyce, and P. Nye, 1991, Analysis of vocal tract shape and dimensions using magnetic resonance imaging: vowels, Journal of the Acoustical Society of America **90**, 799-828.

Baken, R. J. and R. F. Orlikoff, 2000, *Clinical measurement of speech and voice* (Thomson Delmar Learning), 2nd edition.

Balabanian, N., T. A. Bickart, and S. Seshu, 1969, *Electrical network theory* (John Wiley and Sons, Inc.).

Bell-Berti, F., 1976, An electromyographic study of velopharyngeal function in speech, Journal of Speech and Hearing Research **19**, 225-240.

Cairns, D., J. Hansen, and J. Kaiser, 1996a, Recent advances in hypernasal speech detection using the nonlinear Teager energy operator, *4th IEEE International Conference on Spoken Language Processing (ICSLP)*, volume 2, 780–783.

Cairns, D., J. Hansen, and J. Riski, 1994, Detection of hypernasal speech using a nonlinear operator, *Conference of the IEEE Engineering in Medicine and Biology Society*, 253–254.

Cairns, D., J. Hansen, and J. Riski, 1996b, A noninvasive technique for detecting hypernasal speech using a nonlinear operator, IEEE Transactions on Biomedical Engineering **43**, 35-45.

Chen, M., 1995, Acoustic parameters of nasalized vowels in hearing-impaired and normal-hearing speakers, Journal of the Acoustical Society of America **98**(5), 2443-2453.

Chen, M., 1997, Acoustic correlates of English and French nasalized vowels, Journal of the Acoustical Society of America **102**(4), 2360-2370.

Chen, M., 2000, Nasal detection module for a knowledge-based speech recognition system, *1st INTERSPEECH/ICSLP*, volume 4, 636–639.

Childers, D. G., 2000, *Speech processing and synthesis toolboxes* (John Wiley and Sons, Inc.).

Curtis, J., 1970, The acoustics of nasalized speech, Cleft Palate Journal **7**, 380-396.

Dalston, R., D. Warren, and E. Dalston, 1991, A preliminary investigation concerning the use of nasometry in identifying patients with hyponasality and/or nasal airway impairment, Journal of Speech and Hearing Research **34**, 11-18.

Daltson, R., D. Warren, and E. Daltson, 1991, Use of nasometry as a diagnostic tool for identifying patients with velopharyngeal impairment, Cleft Palate-Craniofacial Journal **28**, 184-189.

Dang, J. and K. Honda, 1994a, Investigation of the acoustic characteristics of the velum for vowels, *3rd IEEE International Conference on Spoken Language Processing (ICSLP)*, 603–606.

Dang, J. and K. Honda, 1994b, Morphological and acoustical analysis of the nasal and paranasal cavities, Journal of Acoustical Society of America **96**(4), 2088-2100.

Darley, F., A. Aronson, and J. Brown, 1969, Differential diagnostic patterns of dysarthria, Journal of Speech and Hearing Research **12**, 249-269.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society **39**(1), 1-38.

Denes, P. and E. Pinson, 1993, *The speech chain: the physics and biology of spoken language* (New York: W.H. Freeman Co.).

Duffy, J., 1995, *Motor speech disorders* (Boston: Mosby).

Fant, G., 1960, *Acoustic theory of speech production* (Mounton, The Hague), 2nd, (1970) edition.

Fant, G., 1986, Glottal flow: models and interaction, Journal of Phonetics **14**, 393-399.

Feng, G. and E. Castelli, 1996, Some acoustic features of nasal and nasalized vowels: a target for vowel nasalization, Journal of the Acoustical Society of America **99**(6), 3694-3706.

Feng, G. and C. Kotenkoff, 2006, New considerations for vowel nasalization based on separated mouth-nose recording, *7th INTERSPEECH/ICSLP*, 2242–2245.

Flanagan, J. L., 1972, *Speech analysis synthesis and perception* (Springer-Verlag), 2nd, expanded edition.

Fletcher, S., 1970, Theory and instrumentation for quantitative measurement of nasality, Cleft Palate Journal **7**, 601-609.

Forrest, K. and G. Weismer, 1997, Acoustic analysis of dysarthric speech, *Clinical Management of Sensorimotor Speech Disorder*, edited by M. McNeil (Thieme Medical Pub), 63–80.

Fujimura, O., J. Miller, and S. Kiritani, 1977, A computer-controlled x-ray microbeam study of articulatory characteristics of nasal consonants in English and Japanese, *Paper presented at the 9th International Congress on Acoustics.*

Fujisaki, H. and M. Ljungqvist, 1987, Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 12, 637–640.

Gillick, L. and S. Cox, 1989, Some statistical issues in the comparison of speech recognition algorithms, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, 532–535.

Glass, J. and V. Zue, 1985, Detection of nasalized vowels in American English, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 10, 1569–1572.

Gray, A., R. Gray, and J. Markel, 1977, Comparison of optimal quantizations of speech reflection coefficients, IEEE Transactions on Acoustics, Speech, and Signal Processing **25**, 9-23.

Hasegawa-Johnson, M., J. Baker, S. Borys, K. Chen, E. Coogan, S. Greenberg, A. Juneja, K. Kirchhoff, K. Livescu, S. Mohan, J. Muller, K. Sonmez, *et al.*, 2005, Landmark-based speech recognition: report of the 2004 Johns Hopkins summer workshop, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, 213–216.

Hawkins, S. and K. N. Stevens, 1985, Acoustic and perceptual correlates of the non-nasal-nasal distinction for vowels, Journal of the Acoustical Society of America **77**(4), 1560-1575.

Horiguchi, S. and F. Bell-Berti, 1987, The velotrace: a device for monitoring velar position, Cleft Palate Journal **24**, 104-111.

Horii, Y., 1983, An accelerometric measure as a physical correlate of perceived hypernasality in speech, Journal of Speech and Hearing Research **26**, 476-480.

Hosom, J., A. B. Kain, T. Mishra, J. P. H. van Santen, M. Fried-Oken, and J. Staehely, 2003, Intelligibility of modifications to dysarthric speech, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, 924–927.

House, A. and K. Stevens, 1956, Analog study of the nasalization of vowels, Journal of Speech and Hearing Disorders **21**, 218-232.

Huffman, M., 1990, Implementation of nasal: timing and articulatory landmarks, *UCLA Working Papers in Phonetics*, volume 75, 112–143.

Hutters, B. and K. Brrndsted, 1992, A simple nasal anemometer for clinical purposes, European Journal of Disorders of Communication **27**, 101-119.

Imatomi, S. and T. Arai, 2002, The relation between perceived hypernasality of cleft palate speech and its hoarseness, *Forum Acousticum Sevilla.*

Imatomi, S., T. Arai, Y. Mimura, and M. Kato, 1999, Effects of hoarseness on hypernasality ratings, *6th European Conference on Speech Communication and Technology (EUROSPEECH)*, 1075–1078.

Itakura, F. and S. Saito, 1971, On the optimum quantization of feature parameter in the PARCOR speech synthesizer, *IEEE Conference on Speech Communication and Process*, paper L4, 434–437.

Kain, A., J. Hosom, X. Niu, J. van Santen, M. Fried-Oken, and J. Staehely, 2007, Improving the intelligibility of dysarthric speech, Speech Communication **49**(9), 743-759.

Kain, A., X. Niu, J.-P. Hosom, Q. Miao, and J. van Santen, 2004, Formant re-synthesis of dysarthric speech, *5th ISCA Workshop on Speech Synthesis*, 25–30.

Kaiser, J., 1990, On a simple algorithm to calculate the "energy" of a signal, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, 381–384.

Karnell, M., 1995, Nasometric discrimination of hypernasality and turbulent nasal airflow, The Cleft Palate-Craniofacial Journal **32**(2), 145-148.

Karnell, M. P., E. J. Seaver, and R. M. Dalston, 1988, A comparison of photodetector and endoscopic evaluations of velopharyngeal function, Journal of Speech and Hearing Research **31**, 503-510.

Kataoka, R., 1988, Quantitative evaluation of hypernasality - relation between spectral characteristics and perception of hypernasality, Journal of Japanese Cleft Palate Association **13**, 204-216.

Kataoka, R., K. Michi, K. Okabe, T. Miura, and H. Yoshida, 1996, Spectral properties and quantitative evaluation of hypernasality in vowels, The Cleft Palate-Craniofacial Journal **33**, 43-50.

Kataoka, R., D. Warren, D. Zajac, R. Mayo, and R. Lutz, 2001, The relationship between spectral characteristics and perceived hypernasality in children, Journal of the Acoustical Society of America **109**, 2181-2189.

Kay, S., 1988, *Modern spectral estimation: theory and application* (Prentice-Hall).

Kelly, J. and C. Lochbaum, 1962, Speech synthesis, *4th International Congress on Acoustics*, volume G42, 1–4.

Kent, R., 1994, The clinical science of motor speech disorders: a personal assessment, *Motor Speech Disorder: Advances in Assessment and Treatment*, edited by J. Till, K. Yorkston, and D. Beukelman (Baltimore, MD: Paul H. Brooks), chapter 1, 3–18.

Kent, R., 1996, Hearing and believing: some limits to the auditory-perceptual assessment of speech and voice disorders, American Journal of Speech-Language Pathology **5**, 7-23.

Kent, R., P. Carney, and L. Severeid, 1974, Velar movement and timing: evaluation of a model for binary control, Journal of Speech and Hearing Research **17**(3), 470-488.

Konvalinka, I. and M. Matausek, 1979, Simultaneous estimation of poles and zeros in speech analysis and ITIF-iterative inverse filtering algorithm, IEEE Transactions on Acoustics, Speech, and Signal Processing **27**(5), 485-492.

Kopec, G., A. Oppenheim, and J. Tribolet, 1977, Speech analysis by homomorphic prediction, IEEE Transactions on Acoustics, Speech, and Signal Processing **25**, 40-49.

Krakow, R. and M. Huffman, 1993, Instruments and techniques for investigating nasalization and velopharyngeal function in the laboratory: An introduction, *Phonetics and Phonology: Nasals, Nasalization, and the Velum*, edited by M. Huffman and R. Krakow (Academic Press, Inc.), volume 5, 3–59.

Ladefoged, P., 1993, *A course in phonetics* (Harcourt Brace), 3rd edition.

Laine, T., D. Warren, R. Dalston, and K. Mirr, 1988, Intraoral pressure, nasal pressure and airflow rate in cleft palate speech, Journal of Speech and Hearing Research **31**, 432-437.

Lim, I.-T. and B. Lee, 1993, Lossless pole-zero modeling of speech signals, IEEE Transactions on Speech and Audio Processing **1**(3), 269-276.

Lim, I.-T. and B. Lee, 1996, Lossy pole-zero modeling for speech signals, IEEE Transactions on Speech and Audio Processing **4**(2), 81-88.

Lindqvist-Gauffin, J. and J. Sundberg, 1976, Acoustic properties of the nasal tract, Phonetica **33**, 161-168.

Liu, M. and A. Lacroix, 1996, Improved vocal tract model for the analysis of nasal speech sounds, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, 801–804.

Maeda, S., 1982a, A digital simulation method of the vocal tract system, Speech Communication **1**(2-4), 199-229.

Maeda, S., 1982b, The role of the sinus cavities in the production of nasal vowels, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 911–914.

Maeda, S., 1993, Acoustic of vowel nasalization and articulatory shifts in French nasal vowels, *Phonetics and Phonology: Nasals, Nasalization, and the Velum*, edited by M. Huffman and R. Krakow (Academic Press, Inc.), volume 5, 147–167.

Maragos, P., J. Kaiser, and T. Quatieri, 1993, Energy separation in signal modulations with application to speech analysis, IEEE Transactions on Signal Processing **41**(10), 3024-3051.

Markel, J. and A. Gray, 1976, *Linear prediction of speech* (Springer-Verlag, Berlin, Heidelberg, New York).

Moll, K. L. and R. G. Daniloff, 1971, Investigation of timing of velar movement during speech, Journal of the Acoustical Society of America **50**, 678-684.

Moller, K., R. Martin, and R. Christiansen, 1971, A technique for recording velar movement, Cleft Palate Journal **8**, 263-276.

Moore, B. and B. Glasberg, 1983, Suggested formulae for calculating auditory-filter bandwidths and excitation patterns, Journal of the Acoustical Society of America **74**(3), 750-753.

Moore, C., 1992, The correspondence of vocal tract resonance with volumes obtained from magnetic resonance images, Journal of Speech Hearing Research **35**, 1009-1023.

Murthy, H. and V. Gadde, 2003, The modified group delay function and its application to phoneme recognition, *IEEE International Conference Acoustics, Speech, and Signal Processing*, 68–71.

Narayanan, S., K. Nayak, S. Lee, A. Sethy, and D. Byrd, 2004, An approach to real-time magnetic resonance imaging for speech production, Journal of the Acoustical Society of America **115**(4), 1771-1776.

Niu, X., A. Kain, and J. van Santen, 2005, Estimation of the acoustic properties of the nasal tract during the production of nasalized vowels, *6th INTERSPEECH/EUROSPEECH*, 1045–1048.

Niu, X., A. Kain, and J. van Santen, 2006, A noninvasive, low-cost device to study the velopharyngeal port during speech and some preliminary results, *7th INTERSPEECH-ICSLP*, 957–960.

Niu, X. and J. van Santen, 2007, Dual-channel acoustic detection of nasalization states, *8th INTERSPEECH/EUROSPEECH*, 1921–1924.

Oberg, T., 2001, *Modulation, detection and coding* (John Wiley & Sons, LTD).

Ohala, J. J., 1971, Monitoring soft palate movements in speech, *81st Meeting of the Journal of the Acoustical Society of America*, volume 50, 140.

Perkell, K., M. Cohen, M. Svirsky, M. Matthies, I. Garabieta, and M. Jackson, 1992, Electromagnetic midsagittal articulometer (EMMA) system for transducing speech articulatory movements, Journal of the Acoustical Society of America **92**, 3078-3096.

Pruthi, T. and C. Espy-Wilson, 2005, Simulating and understanding the effects of velar coupling area on nasalized vowel spectra, *150th Meeting of the Journal of the Acoustical Society of America*, volume 118, 2024.

Pruthi, T. and C. Espy-Wilson, 2006, Automatic detection of vowel nasalization using knowledge-based acoustic parameters, *152nd Meeting of the Journal of the Acoustical Society of America*, volume 120, 3377.

Quatieri, T. F., 2002, *Discrete-time speech signal processing: principles and practice* (Upper Saddle River, NJ : Prentice Hall).

Rabiner, L. and R. Shafer, 1978, *Digital processing of speech signals* (Prentice-Hall, Englewood Cliffs, NJ).

Schnell, K. and A. Lacroix, 2001a, Inverse filtering of tube models with frequency dependent tube terminations, *2nd INTERSPEECH/EUROSPEECH*, 2467–2470.

Schnell, K. and A. Lacroix, 2001b, Pole zero estimation from speech signals by an iterative procedure, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 109–112.

Schnell, K. and A. Lacroix, 2002, Parameter estimation of branched tube models by iterative inverse filtering, *IEEE International Conference on Digital Signal Processing*, volume 1, 333–336.

Schnell, K. and A. Lacroix, 2003, Generation of nasalized speech sounds based on branched tube models obtained from separate mouth and nose output, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, 156–159.

Seaver III, E. and R. Dalston, 1990, Using simultaneous Nasometry and standard audio recordings to detect the acoustic onsets and offsets of speech, Journal of Speech and Hearing Research **33**, 358-362.

Serrurier, A. and P. Badin, 2005, A three-dimensional linear articulatory model of velum based on MRI data, *6th INTERSPEECH/EUROSPEECH*.

Shade, C., 1999, The aerodynamics of speech, *The Handbook of Phonetic Sciences*, edited by W. Hardcastle and J. Laver (Blackwell Publishers Ltd.), 33–64.

Skolnick, M., R. Shprintzen, G. McCall, and S. Rakoff, 1975, Patterns of velopharyngeal closure in subjects with repaired cleft palate and normal speech: a multi-view videofluoroscopic analysis, Cleft Palate Journal **12**, 369-376.

Skowronski, M. and J. Harris, 2003, Improving the filter bank of a classic speech feature extraction algorithm, *IEEE International Symposium on Circuits and Systems*, volume 4, 25–28.

Song, K. and C. Un, 1983, Pole-zero modeling of speech based on high-order pole model fitting and decomposition method, IEEE Transactions on Acoustics, Speech, Signal Processing **31**, 1556-1565.

Steiglitz, K., 1978, On the simultaneous estimation of poles and zeros in speech analysis, IEEE Transactions on Acoustics, Speech, and Signal Processing **25**, 229-234.

Stone, M., T. Shawker, T. Talbot, and A. Rich, 1988, Cross-sectional tongue shape during vowels, Journal of the Acoustical Society of America **83**, 1586-1596.

Suzuki, H., T. Nakai, J. Dang, and C. Lu, 1990, Speech production model involving subglottal structure and oral-nasal coupling through closed velum, *1st IEEE International Conference on Spoken Language Processing (ICSLP)*, 437–440.

Teager, H. and S. Teager, 1985, A phenomenological model for vowel production in the vocal tract, *Speech science: Recent advances*, edited by R. Daniloff (College-Hill Press, San Diego, CA), 73–109.

Teager, H. and S. Teager, 1990, Evidence for nonlinear sound production mechanisms in the vocal tract, *Speech production and speech modeling*, edited by W. Hardcastle and A. Marchal (Bonas, France; Kluwer Academic Publishers, Boston, MA), D, 241–262.

Vijayalakshmi, P. and M. Reddy, 2005a, The analysis on band-limited hypernasal speech using group delay based formant extraction technique, *6th INTERSPEECH/EUROSPEECH*, 665–668.

Vijayalakshmi, P. and M. Reddy, 2005b, Detection of hypernasality using statistical pattern classifiers, *6th INTERSPEECH/EUROSPEECH*, 701–704.

Wakita, H., 1972, Estimation of the vocal tract shape by optimal inverse filtering and acoustic/articulatory conversion methods, *SCRL Monograph No. 9* (Speech Communications Research Laboratory).

Wakita, H., 1973a, Direct determination of input impedance singularities from speech for obtaining the vocal tract area function, Journal of the Acoustical Society of America **53**(A), 294-295.

Wakita, H., 1973b, Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms, IEEE Transactions on Audio and Electroacoustics **AU-21**(5), 417-427.

Warren, D., 1964a, Velopharyngeal orifice size and upper pharyngeal pressure-flow patterns in cleft palate speech: a preliminary study, Plastic and Reconstructive Surgery **34**, 15-26.

Warren, D., 1964b, Velopharyngeal orifice size and upper pharyngeal pressure-flow patterns in normal speech, Plastic and Reconstructive Surgery **33**, 148-161.

Warren, D. W., 1967, Nasal emission of air and velopharyngeal function, Cleft Palate Journal **16**, 279-285.

Weismer, G. and R. Martin, 1992, Acoustic and perceptual approaches to the study of intelligibility, *Intelligibility in Speech Disorders: Theory, Measurement, and Management*, edited by R. Kent (Philadelphia: John Benjamins), chapter 3, 67–118.

Appendix A

Table of phonetic labels

IPA, Worldbet, and OGIbet English Broad Phonetic Labels

Center for Spoken Language Understanding – Oregon Graduate Institute of Science & Technology

IPA	Worldbet	OGIbet	Example	Category
iː	i:	iy	beet	
ɪ	I	ih	bit	Front
ɛ	E	eh	bet	Vowels
æ	@	ae	bat	
ɨ	I_x	ix	roses	
ʉ	u_x	ux	suit	Central
ə	&	ax	above	Vowels
ɚ	&0		to go	
ɐ	5		pot	(British)
u	u	uw	boot	
ʊ	U	uh	book	Back
ʌ	^	ah	above	Vowels
ɔ	>	ao	caught	
ɑ	A	aa	father	
ɝ	3r	er	bird	Retro-
ɚ	&r	axr	butter	flexes
ei	ei	ey	bay	
aɪ	aI	ay	bye	Diph-
ɔi	>i	oy	boy	thongs
iʊ	iU		few	
aʊ	aU	aw	about	
oʊ	oU	ow	boat	
iə	i&		here	(British)
eə	e&		there	(British)
uə	u&		poor	(British)
p^h	ph	p	pan	Voiceless
t^h	th	t	tan	Plosives
k^h	kh	k	can	
b	b	b	ban	Voiced
d	d	d	dan	Plosives
g	g	g	gander	
m	m	m	me	
n	n	n	knee	Nasals
ŋ	N	ng	sing	
ɾ$_t$	th_(	dx	writer	Flaps
ɾ$_d$	d_(	dx	rider	
f	f	f	fine	
θ	T	th	thigh	Voiceless
s	s	s	sign	Fricatives
ʃ	S	sh	assure	
h	h	hh	hope	
v	v	v	vine	
ð	D	dh	thy	Voiced
z	z	z	resign	Fricatives
ʒ	Z	zh	azure	
ʧ	tS	ch	church	Affricates
ʤ	dZ	jh	judge	
l	l	l	lent	Glides
ɻ	9r	r	rent	
j	j	y	yes	(approxi-
w	w	w	went	mants)
m̩	m=	em	bottom	
n̩	n=	en	button	Syllabics
ŋ̩	N=	eng		
l̩	l=	el	bottle	

IPA	Worldbet	OGIbet	Example	Category
	pc	pcl	_pan	Voiceless
	tc	tcl	_tan	Plosive
	kc	kcl	_can	Closures
	bc	bcl	_ban	Voiced
	dc	dcl	_dan	Plosive
	gc	gcl	_gander	Closures
	tSc	chcl	_church	Affricate
	dZc	jhcl	_judge	Closures
	+	.epi	epinthetic closure	

IPA	Worldbet	OGIbet	Type of Diacritic
t^h	_h	-h	aspirated
	_x		centralized
t̪ d̪	_[		dental
	_(		flapped (consonant)
	_F		fricated stop
	_?*	q	glottal onset
ʔ	_?	-q	glottalized
d'	_l		lateral release
iː	_:	-el	lengthened
d^n	_n		nasal release
ẽ	_~	-n	nasalized
	_NL	.nitl	not in the language
t^j	_j		palatalized
ɚ	_r	-r	retroflexion
ɹ	_i		less rounded
ɻ	_w		more rounded
ɬ	_=		syllabicity
s̬	_v		voiced
n̥ d̥	_0		voiceless
	_*	-	waveform cut off

Worldbet, as modified at OGI			
	_fp	-fp	filled pause
	_ln	-ln	line noise corruption
	_bn		background noise

Worldbet	OGIbet	Non Speech Sound Item
.bn	.bn	background noise
.br	.br	breath noise
.cough	.cough	cough
.ct	.ct	clear throat
.laugh	.laugh	laugh
.ln	.ln	lin noise
.ls	.ls	lip smack
.ns	.ns	human, not speech
.sneeze	.sneeze	sneeze
.tc	.tc	tongue click

Worldbet, as modified at OGI		
.beep	.beep	beep
.burp	.burp	burp
.fp	.fp	filled pause
.pau	.pau	pause or silence
.sniff	.sniff	sniff
.uu	.unk	unintelligible speech
.vs	.vs	squeak, voice crack
.glot	glot	glottalization

Biographical Note

Xiaochuan Niu was born in Jilin, China, in 1973. He received his bachelor degree from Tsinghua University, China, in 1994. He received his master degree in Electrical Engineering from the Institute of Automation, Chinese Academy of Sciences, in 1998, working on speaker adaptation for speech recognition systems. From 1998 to 2001, he was a research member at IBM China Research lab., working on acoustic modeling of speech recognition systems, and the design and implementation of Mandarin and Cantonese text-to-speech (TTS) systems. He entered the OGI School of Science and Engineering, Oregon Health and Science University, in the Fall of 2001. His research interests include speech analysis, synthesis and recognition, models of human speech production, pattern recognition and signal processing. He is a member of the International Speech Communication Association (ISCA) and of the Institute of Electrical and Electronics Engineers (IEEE). At OGI, Xiaochuan has worked on the development of a duration module and a spectral balance module in the Festival speech synthesis system. He has also worked on formant-trajectory analysis of speech, developing a procedure to compare the degree of coarticulation effects between disorded and normal speech. In the summer of 2004, Xiaochuan was an intern with Sensory Inc., working on the implementation and comparison of different synthesis algorithms for a low-resource American English TTS system. In the summer of 2007, he completed an internship with Google Inc., working on the development of a language identification module for the speech recognition engine that transcribes audio data on the Internet. Xiaochuan is the first author on six conference papers, and a coauthor on one published journal article and five conference papers.