AMERICAN VOICE TYPES:
TOWARDS A VOCAL TYPOLOGY FOR AMERICAN ENGLISH

By

TYLER MCPEEK

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2013

To my many colleagues, advisors, family, and friends who have helped me achieve success on this road of intellectual development—too many to mention and too great a contribution to put in words

ACKNOWLEDGMENTS

I would like to acknowledge the members of my committee, Drs. Diana Boxer, Rahul Shrivastav, and Ratree Wayland for their selfless service and dedication to my academic success. Most especially, I want to express my deep appreciation to my advisor, committee chair, academic mentor, and friend, Dr. James D. Harnsberger, without whom this project would have been impossible. The following individuals should be acknowledged for their assistance in data collection: Dot Bourgoise, Shelley Hicks, Anna Rowe, and Christine White.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

A	Average values for database
C	Clear
D	Dynamic (relative to pitch variability)
F	Fast (relative to the average speed in the database)
HCS	Hierarchical Clustering Scheme (Analysis)
ID	Identification
IRB	Institutional Review Board
LDA	Linear Discriminant Analysis
LPC	Linear Predictive Coding
M	Monotone (relative to pitch variability)
MDS	Multidimensional Scaling
ms	milliseconds
N	Nasal
O	Oral
R	Rough
s	seconds
S	Slow (relative to the average speed in the database)
SAUSI	Semi-Automatic Speaker Identification System
SPID	Speaker Identification
SPIN	Speech Perception In Noise (Sentences)
VT	Voice Types
VTID	Voice Type Identification
X	Variable within the voice type for the characteristic in question

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

AMERICAN VOICE TYPES:
TOWARDS A VOCAL TYPOLOGY FOR AMERICAN ENGLISH

By

Tyler McPeek

August 2013

Chair: James D. Harnsberger
Major: Linguistics

Individual voices are not uniformly similar to others, even when factoring out speaker characteristics such as sex, age, dialect, and so on. Some speakers share common features and can cohere into groups based on gross vocal similarity but, to date, no attempt has been made to describe these features systematically or to generate a taxonomy based on such “voice types.” For this purpose, perceived similarity judgments of voice pairs using a database of 100 female and male American English voices were collected and submitted to a hierarchical clustering analysis to generate the initial groupings of individual voices into types, separately for female and male voices. These types, in turn, were labeled based on auditory judgments by expert listeners on nominal scales (e.g., voice quality, mean pitch, pitch variability, and speaking rate) as well as an initial acoustic analysis using automated measures. The new typology revealed a total of 9 female and 10 male voice types, with voice quality, mean pitch, and pitch variability playing the largest roles in determining the taxonomy for both sexes. This new vocal typology of American voices, along with future study and revision, will find utility in academia (phonetics, discourse, sociolinguistics, genetics, and other

fields), forensic linguistics, public and private sector business and marketing, voice acting, and public interest.

CHAPTER 1
INTRODUCTION

Speaker Identification

Individual Speech Characteristics and Natural Class Voice Types

Both linguistic and non-linguistic information are transmitted in the same signal of a spoken voice, and the correct identification of linguistic information can be crucially dependent on the acoustic variability associated with non-linguistic, or *indexical* information. The *indexical* properties of speech specify information about the history and/or current state of the speaker him/herself, such as sex, age, dialect, emotion, fatigue, pathology and, most relevant to this study, speaker identity. Individual speakers can be identified by numerous means, both linguistic and nonlinguistic, but for these purposes, speaker identity refers to all aspects of the speech signal that are independent of other indexical and linguistic properties. These can be a product of the vocal and/or speech anatomy of the individual as well as idiosyncratic physiological patterns. Prior work in the correlates of Speaker Identification (SPID) by human listeners have identified such cues as speaking fundamental frequency (Atal, 1972; Iles, 1972; Jassem, et al., 1973; LaRiviere, 1975; Mead, 1974), mid to high frequency spectral information, such as higher formants F3 – F5 (Goldstein, 1976; Jassem, 1968; Iles, 1972; LaRiviere, 1975), nasality (Glenn & Kleiner, 1976; Su, et al., 1974), temporal speech features (Abberton & Fourcin, 1978; Johnson, et al., 1984), voice quality (Hollien & Majewski, 1977; Johnson, et al., 1977; Zalewski, et al., 1975), fricative articulation (Ingemann, 1968; Schuartz, 1986) -- and others (Hirson & Duckworth, 1995; Lass, et al., 1976; Orchard & Yarmey, 1995; Wolf, 1972; Young & Campbell, 1967). Anatomical and physiological features of the speaker are static to some extent,

determined by factors such as the size and health of the vocal folds, air flow volume from the respiratory system, length, width, and shape of the vocal tract, the sufficiency of the velopharyngeal port, as insufficient closure of the velum may result in more nasality, and dentition. In addition, the way in which we use our articulators can also be unique, exhibiting more or less of various characteristics in speaker styles, such as degree of coarticulation and fast or slow speech (Beigi, 2011; Hollien, 2002; Neustein & Patil, 2012). Interspeaker variability in both speaker characteristics based on articulator usage and anatomically fixed features of speakers are likely to inform the groups in a typology of the type revealed by the present study.

Individual voices are not uniformly similar to others, even when factoring out speaker characteristics such as sex, age, dialect, and so on. Some speakers share common features and thus may naturally form groupings, termed hereafter as “voice types”, that have not been systematically described or labeled to date. In both human perception of speaker identity and in machine-based approaches, error rates are usually above zero, permitting the examination of patterns of confusion *among* individual voices. In all cases, the error patterns do not show a random distribution of errors across competing voices. Rather, a given voice tends to be more confusable with one or more of a limited number of other voices in the test set. For example, in a classic study, Bricker and Pruzansky (1966) examined the effect of stimulus duration and type on the identification of familiar male voices. In the course of doing so, they generated confusion matrices for the ten voices under each experimental condition. In all cases, when voices were misidentified, they were not confused at equal likelihood with all other nine voices in the test set. Instead, each voice was typically confused with 1 – 2 others

consistently (although, asymmetries in the matrices were also observed). A reanalysis of the error rates in the monosyllabic condition of the study shows that the most confusable voice for each of the ten talkers garnered 27% - 53% of the errors recorded, well above a chance level of 11%. In other words, voices in the set were not uniformly similar: some were approximate sound-alikes or showed high degrees of similarity. Such nonrandom error patterns in this and other studies serve as evidence that voices may naturally cohere into vocal similarity groups in the ear and mind of listeners.

That voice types naturally occur is not especially surprising. Speaker identities can be confused over a phone or in other degraded listening conditions. Folk terms exist for vocal qualities that are not necessarily pathological but are distinctive, such as “nasally,” “whiny,” “gravelly,” “droning,” “staccato,” and others. What remains, however, is a systematic approach for identifying the number and type of the most common vocal stereotypes, or types, that speakers cohere into based on human perception. An inventory of voice types should be developed which is independent of other speaker characteristics (e.g., age, sex, dialect, pathology) and which serves to reduce the vast population of speaker identities by voice into a more manageable taxonomy of common types. Such voice types may play a role, as do other indexical properties, as perceptual units that partly influence the processing of linguistic and nonlinguistic information by human listeners. Their existence also points to numerous applications. In the forensic domain, SPID is a very common analysis required of audio evidence in cases and, yet, the duration of the speech samples and their quality can often preclude a highly confident judgment of the match/mismatch to the voice of a defendant or a relevant party in the case. However, such evidence recordings may be of sufficient caliber to

permit a match/mismatch determination on the basis of a more gross category, such as a voice type. The evaluation of voice talent is also a growing field, given the increasing use of digital animation in the entertainment industry. While individual vocal attributes such as “pleasantness” or “authority” have been examined in prior work (Beebe-Center, 1965; Oyer & Trudeau, 1984; Bugental & Lin, 1997; and others), there is currently no rubric or automated procedure for classifying all of the relevant characteristics of a talented voice. Voice talent could be fit into a voice type taxonomy for increased ease of identifying the proper vocal talent for a given commercial application. This would include public service announcements and advertisement narration, where vocal pleasantness correlates such as trustworthiness, sex appeal, and overall pleasantness or friendliness play a significant role in listener impression, attention to message, and overall decision making and effectiveness of the message. The positing of a voice type taxonomy ultimately serves to reduce the vast number of speaker identities within a given sex/age/dialect subpopulation down to a manageable and useful number of categories.

An examination of existing forensic SPID methods will further an understanding of what SPID methods are in use today, and what correlates are measured and applied to individual speaker identity, when forming a speaking voice profile. These methods, in addition to appropriate sample length, recording quality issues, vocal disguise, and other factors relevant to SPID are of help in experimental design for experimentation in voice type identification and labeling, towards the construction of a complete vocal typology of American English.

Forensic Speaker Identification

Identification of the human voice is a challenge that has produced a number of proposed standards, but found no agreement of standard in either the literature or in

practice. It is for this reason that voice identification has been mired in legal challenges from its earliest use in the US justice system more than a century ago (McGehee, 1937). Voice identification's legal standing as evidence in a court of law remains in question today, with outdated methods such as voice prints being presented in courts alongside newer, more reliable analyses. The continual usurping of newer methods aimed at extracting speaker-specific information from speech waves represents both progress and a continuing challenge to the legal validity of contemporary and historical methods (Furui, 1997; Hollien, 2002).

Methods have been proposed that use both machine and human (aural perceptual) analyses. The latter can be seen as a systematic synthesis of the untrained listener's natural approach to the task. Parse-able elements of the speech signal can be fed into a fully automated system ("automatic"), examined by human operators aided by software or machine ("semi-automatic," e.g., use of voice spectrograms (Endres, et al., 1971) and examination of spectral moments of vocalic sounds in a speech sample (Rodman, et al., 2002), or judged on a rating scale by human aural perceptual analysis ("manual," e.g., non- professional and professional historical SPID cases and voice "line-ups"). Thus far, none of these methods can be conclusively shown to be more effective overall than an untrained listener's ability to identify voices by ear in the natural world (Hollien, et al., 1982; Hollien & Harnsberger, 2012). Of course, research continues in a number of promising directions, in an attempt to establish standards and common methods for SPID.

All automated systems involve significant human operator preparation of the sound files (at a minimum). Therefore, thus far, "fully automated" systems are

somewhat of a misnomer. Hollien's successful SAUSI (Semi-Automatic Speaker Identification System) would be considered semi-automatic as it involves manual selection of comparison voices to the two voice samples being compared (Hollien, 2002). Semi-automated systems are generally considered to be the best systematic methods of SPID available to forensic professionals today, as they provide both the advantages of current technological methods of acoustic and phonetic measurement with those of expert listener aural perceptual evaluation and analysis of the evidence recording or speech sample (Hollien, 2002). It is for this reason, that this study's experimental methodology combines these two methods with untrained listener similarity judgment data, to reflect the most current thinking in forensic practitioner and academic researcher methods for speech signal evaluation and identification.

In the case of manual and semi-automatic identification of speakers, creakiness, breathiness, nasality, consonants, vowels, pitch, and intonation are elements that can be examined and rated by a trained or untrained listener. Pitch level, pitch patterns, pitch variability, vocal intensity, dialect, voice and speech quality, timing and melody of the speech (prosody), and articulation can be measures of utility in identifying a particular speaker from an open or closed group of possible candidates. Hollien's approach to aural-perceptual SPID uses a score sheet to rank factors in the categories of "Pitch" (level, variability, patterns), "Voice Quality" (general, vocal fry, other), "Intensity" (variability), "Dialect" (regional, foreign, idiolect), "Articulation" (vowels, consonants, misarticulations, nasality), "Prosody" (rate, speech bursts, other), and "Other" (nonfluencies, speech disorders, other) on a scale of 0-10 (0 = Unknown and Known least alike; 10 = Unknown and Known most alike) by a rater listener (usually

expert). The process can be repeated and averaged amongst a team of SPID experts (Hollien, 2002). Acoustic, machine measurement, if measureable—dependent on the quality of the evidence recordings (often quite degraded in nature, recording quality, or with heavy background noise) and speech samples (frequently also taken from interview recordings that are less than ideal or not of laboratory recording quality), are then used to enhance the aural perceptual judgments and to make determinations, to the degree possible, by the practitioner for issuance of a final report on the given case.

While laboratory quality voice samples, with chosen text (text dependent) would provide the best accuracy, speed, and ease of assessment, anything less than vocalic samples of spoken speech segments would not be useful for this type of analysis, as for other forms of SPID and classification as well. Two glottal pulses of about 25ms in duration is typically the minimum sample for LPC (Linear Predictive Coding) and other types of analyses that would be needed for machine assisted phonetic analysis of any kind.

Any, thus far hypothetical, discussion of the minimum stimulus necessary for accessing vocal type of a voice would need to begin with a general review of the minimum stimulus necessary for standard SPID. The answer to minimum stimulus for SPID will involve at least two factors: size or duration of the sample and acoustic quality. First, duration of the samples (evidence and exemplar) need to be defined in some way. As Hollien (2002) has pointed out, “If you are attempting an analysis, you naturally will want them (the samples) to be long enough to permit you to be successful” (p. 40). The minimum useful sample has been gauged at anywhere from 25ms to 30s, but the problem seems less to do with the length of the sample, as with the phonemic

sampling set (Pollack, et al., 1954). From a proper sampling array of the different phonemes in the language to be analyzed for a particular set of samples, the analyzer can access both the segmental and suprasegmental information necessary for a proper SPID attempt. Bricker and Pruzansky (1966) experimented with the effects of stimulus content and duration on talker identification and found, among other things, that identification accuracy correlated directly with the number of phonemes in the sample, even when duration was controlled. On the segmental side, interspeaker variation in pronunciation of individual sounds can be assessed, as well as intraspeaker variation being accounted for more appropriately with a more lengthy sample. Likewise, suprasegmental information can be gleaned from a sample that includes, at a minimum, full sentences and words for morpho-syntactic, suprasegmental information to have a chance to immerge. In addition, fundamental frequency, vocal intensity, prosody, speech timing, and voice quality will best immerge from a sample containing full sentences (Dommelen, et al., 1987; Kinoshita, et al., 2009; Narayan and Yuan, 2008; Schmidt- Nielsen and Stern, 1984). Lastly, listener familiarity is universally agreed to greatly influence the reliability of manual and untrained SPID and the duration of a sample that would be necessary for an accurate identification in those cases (Goggin, et al., 1981; Holien, 2002).

A second issue regarding minimal stimulus for SPID involves sample quality. In forensic cases, many factors can influence sample quality and make SPID either difficult or impossible. These factors can typically include: channel distortion, cross-talk, low quality microphones or recording equipment (limited signal bandwidth), faint or

otherwise distorted recordings or utterances, vocal disguise, and the presence of noise of all kinds in a recording or real world listening situation (Hollien, 1990).

Most of the above factors are as relevant to voice typing, as they are to SPID, especially when the voice typing is being applied to forensic cases of SPID, where the stakes are high and finite forensic recordings are being compared with exemplars from possible candidates, in an attempt to create a “voice type match” for criminal or civil litigation or for intelligence purposes. Fortunately, many other commercial and research applications can be envisioned for voice typing that would include only one exemplar voice (no evidence recording) and where the length of the sample available is limited only to a reasonable amount of unfettered access to the speaker, under ideal recording conditions with modern equipment in a laboratory setting. Thus, many of the limitations of forensic SPID cases, including quality and length of the recording, would not be an issue for voice typing. In these later cases, a proper phonemic sampling of read speech, analyzed by an expert using a semi-automatic process (eventually to be fully-automated) would be both robust and accurate enough to make a Voice Type Identification (VTID). In the present experiments, 3-5 designed sample sentences (to cover a range of desired and representative phonemes) will be sufficient. Additionally, in most cases, vocal disguise (whispered speech, falsetto, or machine altered speech), which would most certainly pose a countermeasure issue in many SPID cases (Endres, et al., 1971; Reich and Duke, 1979; Reich, 1980; Hollien, 2002), will not be an issue for voice typing—as the samples are given voluntarily, under non-stressed and normal conditions, and are not to do with criminal or civil liability, thus eliminating the incentive to disguise ones voice or deploy similar countermeasures to hinder, constrain, and

confuse the forensic examiner that are typically encountered in forensic SPID cases by field practitioners.

In summary, properties that will motivate the classing of vocal types can be understood to be similar to those used by forensic phoneticians, both phonetically-driven and articulatory (segmental vs. suprasegmental) and having to do with the acoustic consequences of the articulation of speech by the speaker. On the phonetically driven side, properties such as mean pitch, pitch variability, voice quality (such as breathiness and creakiness), consonant usage, and vowel usage emerge. The acoustic consequences of speech may include mean speaking fundamental frequency, standard deviation, harmonics to noise ratio, spectral tilt, and the like. It can be further observed how many of these phonetics and forensic based correlates reoccur, supplemented by layperson and area specific descriptors and scales, in a review of singing voice literature and expert analysis.

The Singing Voice

To date, voice types have been identified only in the realm of singing voices. Singing voice experts tend to hail from the fine arts—musicology and operatically trained academics and professionals in music departments and private voice training schools. While much has been written about singing voice types and natural singing voice types, the singing voice is generally understood to be a product of training, and interviews with singing voice experts revealed no special prediction of an individual's singing voice type, especially after training, based on listening to the singers' spoken voice alone. In addition, a performer's singing voice product has much to do with the auditory perceptual process through training, and not only with the pedagogy of the voice training and physiological factors such as body size and vocal tract anatomical

characteristics and configuration (Shrivastav & Wingate, 2006). For these reasons, a typology of singing voices is necessary to be addressed in a review of the literature relevant to speaking voice types, as it represents the only existing work on typing voices, but is not a predictive measure of a singer's spoken voice type, as outlined in this study. Vocal descriptors specific to the singing arts, however, are especially relevant to the present research into types of spoken voices, as they will provide additional tools for the aural perceptual labels and descriptions of the final natural class voice types by expert evaluators when describing and labeling the final, resulting typology. Limitations in the application of the singing voice literature and method to spoken voice types arise primarily from two factors: trained (singing) vs. untrained (speaking) voices and lack of strict classification standards for singing voice types.

An issue at the forefront of singing voice typology is and has been standardization. From the European Renaissance until the present day, singing voice experts are very critical of the functionality and methodology behind competing singing voice type models (Koopman, 1999). Scientific approaches to singing voice classification, based on anatomy and physiology of the individual singer and acoustically measureable characteristics of the vocal output have generally taken a back seat in the typing process to traditional perceptual judgments of trained voices by singing coaches (Erickson, Perry, & Handel, 2001). Singing voices are judged for both professional quality, according to industry and traditional standards, and also for their natural and trained suitability to a particular singing voice type. In both cases, the judgment criteria is currently understood to be reliant on the subjective perception and interpretation of highly trained and experienced listener judges (Callaghan, 2000). In

either case, ratings of voice quality exhibit low reliable consistency among listener evaluators, due largely to differences in interpretation of rating scales by individual listeners, in addition to lack of standardized rating systems across the field (Kreiman, et al., 1993; Shrivastav, Sapienza, & Nandur, 2005; Shrivastav 2006). Titze (1994), suggests that the utility of existing singing voice judgment methods lies in combining several methods, as a mode of guidance and for finding common ground between a student and a teacher of voice during training and evaluation. Which methods, how they should be combined, and to what degree a standard, mutual criteria should be relied upon even in these individual scenarios is left open and to the discretion of the singing voice training pairs (particular instances of teacher and student)—further enforcing the lack of standardization across the field, academically and in professional settings. For these reasons (lack of standardization, disagreement on methodology, and reliance on subjective and non-quantifiable classification criteria), singing voice typing methods are limited in what they provide as reference towards the construction of a speaking voice typology. However, it does provide useful cues for the labeling and parameter definition of an experiment-based resulting typology of spoken voice.

It is important to this study to summarize the contrasting characteristics and a limited set of vocal descriptors used amongst those in the industry, as singing voice literature is the only existing typology on voice. Also, it is useful to point out those methods (regardless of the standardization of their use in the industry) that are scientific and objective in their application, as a matter of background relevance to the present study.

The singing voice has traditionally been categorized by the assessment of three vectors: range, timbre, and tessitura (Cleveland, 1993). Different assessment methods frequently employ equivalent terms in place of these three. Most, however, acknowledge these parameters as the benchmarks for classification. Scientists over the last five decades have successfully quantified singing voices according to these parameters with the aid of such techniques as long-term spectral analysis, video endoscopy, and electromyographic recordings (Callaghan, 2000). These same three parameters have been used by singing teachers and pedagogues for centuries, gradually developing into the most widely accepted system for voice classification, which categorizes singing voice types by identifying the singer's high fundamental frequency and low frequency boundaries in singing effort. This system renders, in its simplest form, the quintessential voice classes of Bass, Tenor, Alto, and Soprano. These classes originate from the European singing voice classification systems and generally reflect differences in register or pitch. Singers can be classed by their range for choral singing, but also into parallel voice types for solo singers that reflect where the voice feels most comfortable (tessitura). Additional descriptors, standard and nonstandard are further applied to describe elements of an individual singing voice, such as resonance. These sub-types reflect not only range, but also weight, dexterity, and vocal color or timbre. Examples include: soubrette, lyric soprano, and dramatic soprano. The pitch distinction among these four general classes, and their more specific types, such as mezzo-soprano, coloratura, and others, are in some cases sex specific—reflecting the belief among singing voice experts that some ranges can only be covered by either female or male singers.

Despite discrepancies in methodology, different methods frequently employ the use of a common set of descriptive adjectives to describe a particular aspect of singing voice. The prevalence and exchange of these terms with other synonymous descriptors varies in the literature; however, the following four descriptive contrasts were commonly utilized: *dark vs. bright*, *ringing vs. nasal, rich vs. rough*, and *heavy vs. light*. These terms are used frequently, while other, non-standard and difficult to define artistic descriptors (velvet, silky, chocolate, soaring, etc.) also persist in common use and in the literature, but are too numerous and non-standardized to list at length in the current discussion.

These four contrasts, outlined above, may be the closest to a relatable, scientifically-rooted voice classification system that can be found in the singing voice literature, one that will be useful as background to the present study. What follows are some of the measureable characteristics associated with the above contrastive descriptors, as identified by singing voice researchers.

Sundberg (1977) stated that a “darker sound” is associated with singing with the larynx in a “lowered position.” The correlation between laryngeal positioning and the perceptual presence of a darkened sound is furthered by Titze’s view (1993), that for voices of equal larynx positioning, a voice with a longer vocal tract will be perceived as “darker.” According to Callaghan (2000), a “ringing” or “resonant” singing voice is identified by the presence of the ‘singer’s formant,’ measured by spectral energy. The term “singer’s formant” was developed when voice scientists and phoneticians discovered an extra “spectral peak,” created by bringing the normal third (F3) and fourth

(F4) formants close to each other so that there is greater energy around 3kHz in the spectrum of a sung vowel (Sundberg, 1977).

Of the commonly used singing voice terminology, “rich” is among the most subjective. A “rich” singing voice is one with a resonant voice quality that is absent of “non-harmonic noise” (Lundy, et al., 1999). The overall “richness” and “ringing” of a particular voice can partly be attributed, again, to the strong acoustic presence of the “singer’s formant.” Lundy, et al. (1999) explain that the increase in the signal intensity between the third and fourth formants allows for the singer to be heard over sounds and accompanying music without amplification. The distinction between the uses of the term “ringing” versus “rich” (according to lay language used by Lundy and others in the less scientifically-minded singing voice literature) is related to the voice’s vocal quality, or timbre. “Ringing” seems to refer to the overall “robustness” (i.e., the width and distribution of energy between the formant frequencies) of a voice at the supralaryngeal level, whereas the term “rich” is more related to the level of the vocal folds.

The perception of the terms “light” or “heavy” when classifying the singing voice refer generally to the voice’s “register.” Voice register is a popular vector of singing classification, because registers can be identified both perceptually and through acoustic analysis, despite being referred to by inconsistent terminology throughout the field (Titze, 1994). Register is correlated with physiological changes at the level of the larynx. According to Callaghan (2000), registers refer to changes in voice quality at particular pitches due to changes in the action of the interdependent cricothyroid and (later) vocalis muscles in the larynx. A “light” voice has more energy in the fundamental

frequency and the first harmonic overtones, while a “heavy” voice has more energy in the harmonic overtones above 5kHz.

The above singing voice based correlates and scales will find utility in the current study’s eventual description of natural class voice types. In the current study, a series of perception experiments and acoustic and perceptual analysis with an existing database of 100 American English voices was used to determine the number and characteristics of the inventory of voice types. More specifically, perceptual data from untrained listeners was integrated in order to reduce the large set of individual voices to a smaller, workable set of voice types. These voice types and the data itself provide an initial model for vocal typology that will find potential revision through future experimentation and eventual utility in the public, academic, and private sectors.

Proposed Study

Initial development of a voice type inventory will be driven by experimentation on perceptual similarity of voice pairs by listeners. However, once an initial inventory of types can be established, machine analysis in conjunction with statistical and expert perceptual analysis will be necessary for labeling of the type, identification of the signal characteristics common to each type for description, and typological identification of new voices. Thus, for initial experimentation toward establishing voice types, only the amount of read speech necessary for untrained listeners to judge degree of similarity will be necessary—though this may vary from the eventual amount of stimulus necessary for placing new voices within the established taxonomy.

Forensic, clinical, and commercial applications will all require a system for typing incoming voices. However, since forensic, clinical, and commercial applications have a descending order of importance by real world gravity of consequence and ramifications

for identification (criminal and civil litigation for forensic, proper patient pathological diagnosis for clinical, and commercial viability and reliability for commercial applications), expert aural perceptual and machine combination methods are imperative for forensic identification, where as “type range” ID might be acceptable for many commercial applications, such as vocal talent placement agencies and their clients. In these cases, a given incoming voice is classified as belonging to one of a possible range of types, when conditions are not ideal for making a more exact identification and/or classification. Eventually, a coded, semi-automatic or automatic system for typing incoming voices is desirable.

In the present study, the vocal types will ultimately emerge from similarity judgments made by untrained listeners, which are then mapped to natural class categories through the Hierarchical Clustering Scheme (HCS) model. Male and female voices are separated at the onset, before the experimental phase. Male and female voices will be considered separately, because this is one of the indexical properties of the speaker that has been well-researched and is readily detectable by machine, expert, and non-expert listener alike in a majority of cases, which would interfere with the identity property of speech that is sought for isolation in this experiment. Sex constitutes the first two divisions in the hierarchy, and necessarily so, for purposes of gleaning useable types through the listener experiments, and to avoid interference and distraction of properties not the subject of this study of vocal typology. Therefore, N voice types can be hypothetically proposed per sex (though they may correlate by distinguishing features and characteristics in their respective sex context) as a minimum number of useable intermediate categories for application. The final number of

categories will be determined by the strength of the statistical similarity correlation, expert analysis in the labeling phase, and the needs of the user (in the case of a variable model—use of different levels in the hierarchy by application).

Though listeners may have difficulty classifying degree of similarity in spoken voice language samples not native to the listener, it must be expected that identity-based types will have some universal and cross-linguistic application, though duplicate testing in dialect and non-English language settings will be needed to compare the emerging types for degree of similarity by machine and bilingual characteristics. It is expected that the identity property types that emerge will necessarily be at least partially dictated by anatomy of the vocal tract (nasal cavity size, F0, length and thickness of the vocal folds, and so on) and other genetically physiological and anatomically fixed factors, if the environmentally-based indexical properties of speech can be successfully equaled out in the sample population before testing. These types will be at least partially cross-linguistic and universal in nature, though the answer of cultural specificity cannot be known definitively pre-testing.

Considering cultural and sex factors together, it is clear that both linguistic content, affected pitch, prosody, and other factors can hold clues to a person's identity that move beyond physiology and fixed elements of the speech signal. In Japanese, for example, pitch is controlled for both social function and sex or gender identification, which is not typically the same in other languages, such as English. Fast rate speech and high pitch are characteristic culturally, not anatomically for female speech. Furthermore, pitch is affected by social convention, such as speaking in higher pitches when answering the telephone or getting the attention of a waiter in a restaurant

(Washi, 2004). The speech used for stimuli in the present experiment has been filtered, to the extent possible, for indexical properties unrelated to the identity property of the speech signal. These properties include: known speech pathologies, illness, intoxication, obvious or “strong” dialect, age, and sex (by pre-sorting in the first level in the hierarchy for sex). Cultures with specific sex roles by nature constrain the voice types by sex. For this and previously stated reasons, sex must be separated at the onset, to be able to isolate the identity-based properties of the speech signal in judgment criteria towards eventual natural class type inclusion.

When asking what contributes to a “voice type,” factors that contribute to prosody will be very important. These include speaking fundamental frequency, degree of nasality, F0 range, and intonation. Most of these would be expected to fall under the purview of fixed factors in the speech signal, including size of the nasal cavity for nasality and issues related to fundamental frequency and its range by length and shape of the vocal tract and size and thickness of the vocal folds. In addition to these fixed, anatomically-motivated factors, there is speaking rate. Speaking rate is one of the features that has diminished relevance with the use of read speech in this experiment. However, using read speech solves problems related to sex and culturally specific interference in finding natural classes by identity in the speech signal.

A list of vocal descriptors has been compiled which can be utilized by expert analysts in the final labeling of the voice types yielded from the matrix (after acoustic analysis for label). These descriptors have been selected from a compilation that draws primarily from five over-lapping groups: descriptors used by speech pathologists and speech experts, linguistic descriptors, descriptors used to type and specify voices by

singing voice experts and voice talent specialists, common-use descriptors used by non-experts to describe the qualities of a voice, and common-use descriptors used by non-experts to describe the emotional response of listeners to a given voice. A representative sample of descriptors and descriptive vectors/scales from each group are shown in Table 1-1.

Descriptors used to describe emotional responses to certain voice types are useful; however, they constitute a different method of labeling a voice type that are most efficiently yielded from further testing on the established voice types by un-trained, non-expert listeners, utilizing an emotional response rating system similar to those used in previous vocal pleasantness experimentation.

Table 1-1. Vocal descriptors by category

Speech Pathology	Linguistic	Singing Voice	Popular/Common Use	
			General Characteristics/Voice Quality	Emotional Response
rough	creaky	timbre	gravely	whiney
breathy	laryngeal	register	staccato	droning
strain	breathy	baritone	scratchy	commanding
hoarse	murmured	chocolate	deep	nerdy
clear		velvet	high	meek
flexibility		soaring	low	strong
asthenia		range	melodic	weak
nasal resonance		dark	monotone	authoritative
loudness		bright	child-like	masculine
pitch range		ringing	musical	feminine
pitch		nasal	clear	flighty
quality		rich	rough	annoying
rate		rough	low	bright
prosody		light	high	upbeat
glottal fry		heavy	smooth	creepy
diplophonia		bass	creaky	unsure
		tenor	whispered	golden
		alto	halting	sexy
		soprano	full	reassuring

Five hypothetical voice types (sex specific, one set for each sex) might look like those outlined in Tables 1-2, 1-3, and 1-4.

Table 1-2. Perception based descriptor labels

Scratchy-Nasally
Deep Baritone
High Melodic
Low Breathy
Rapid Staccato
Dark Ringing
Bright Heavy
Rich Deep
Low Monotone
Rough Halting

The labels in Table 1-2 would be considered the “descriptor labels.” These labels could be assigned corresponding labels by arbitrary coding, avoiding emotional, social, or pathology based assumptions about a voice type or speaker by label (until such associations can be made and labeled to each type by further vocal pleasantness studies), as shown in Table 1-3.

Table 1-3. Arbitrary coding type labels

Type 2A
Type 4C
Type 1D
Type 7S
Type 3R
Type 6T
Type 7Y
Type 8D
Type 5B
Type 0Y

Celebrity monikers judged by experts to be included in and able to represent a given type could then be assigned to make the labels more interesting, relevant, and easier to remember for the general public, as shown in Table 1-4. This type of labeling system also risks emotional and social associations with each type that might be better

avoided for scientific labeling, but nonetheless carry commercial and general interest appeal for general use.

Table 1-4. Celebrity moniker labels

“Gilbert Gottfried Type”
“George Clooney Type”
“Shirley Temple Type”
“Marilyn Monroe Type”
“Lou Costello Type”
“Casey Kasem Type”
“Barry White Type”
“Julia Childs Type”
“James Earl Jones Type”
“Woody Allen Type”

General public labeling can also utilize emotional response labels in a personality type labeling system of the type shown in Table 1-5, if appropriate for the given application of the typology. These types of labels also will be enhanced by future research done on the typology and VT generally by researchers in the field of vocal pleasantness and sociolinguistics. Again, the application of these types of labels generally carries tradeoffs via public interest vs. scientific application. In both cases, they may be better applied when they can be justified properly by listener experimentation focused on emotional and favorable vs. unfavorable response to each individual voice type on a given set of scales or of one voice type as compared to another, on a given set of dichotomous criteria, such as general pleasantness, trustworthiness, authority, perceived intelligence and intellectual prowess, and so on.

Personality type labels in particular give dimension and emotional association to the types that can provide, in conjunction with other types, easy mental association and quick allusion to the general characteristics and sound of the voices in the given type.

Table 1-5. Emotional response based personality type labels

"The Whiner"
"The Golden Voice"
"The Kid"
"The Preacher"
"The Seducer"
"The Droner"
"The Mouse"
"The General"
"The Thinker"
"The Reluctant Participant"

The final labeling of voices could be done one of two ways, either by machine and mathematical, data reduction analysis, based on acoustic cue measurements of the voices, as was performed on pilot data, or by expert or naïve listeners working with either a fixed set or open array of vocal descriptors. Use of trained "categorizers" working with a limited set of pre-selected vocal descriptors would be a "top-down" approach. The advantage of the later approach is that it would be comprehensive in its use of available acoustic information, and not limited to measured variables which may or may not be the optimal measurements for determining the parameters for definition of each type. The disadvantage of this method is general human bias of the listeners (inter-rater and intra-rater reliability), which would be accentuated by a limited number of experts.

The acoustic analysis method would determine the parameters of the vocal type through a "bottom-up" approach. The machine analysis of acoustic cue method carries the advantage of being the superior method of accurately classing new, incoming voices into new categories, without the need of the bottom-up approach of listener similarity judgments in each case. Acoustic analysis would be repeatable across novel stimuli, which is highly desirable in this case. Acoustic analysis also carries the advantage over

expert listeners of being unbiased. The disadvantage of a machine method lies in uncertainty as to the total cue inventory, and producing an “over-fitting” problem with specific databases, making novel stimuli classification ineffective. These disadvantages are manageable, however, and it was for this reason that the acoustic analysis machine method of classification was ultimately used for the current study, along with data reduction analysis, to determine the parameters and number of types for voices in a pilot study, with expert listeners’ aural perceptual judgments being used through a supplemental experiment after the natural class types were established through data reduction from the untrained listeners judgment data, and mainly only for confirmation and verification of the acoustic analyses results.

There are three statistical methods that could be of greatest utility in establishing and modeling a vocal taxonomy. They are Hierarchical Clustering Scheme (HCS) analysis, Multidimensional Scaling (MDS), and Linear Discriminant Analysis (LDA). HCS is the most indispensible statistical method to be used in this study. It is a bottom-up approach to building a hierarchy of clusters, where each voice (in this case) starts in its own cluster, with voices and pairs of voices merging as you move up the hierarchy. This is an agglomerative approach to clustering members of the sample, which is fueled by the listener similarity judgments. HCS determines the taxonomy, by revealing the natural classes within the sample of voices. The procedure begins with a number of clusters equal to the number of cases that make up the matrix (i.e., 50 voices in the case of the pilot study per sex group). Next, the clusters are denoted 1 through N. Then, the most similar pair of clusters are determined, merged through a method (i.e., arithmetic averaging), and then the matrix is revised to reflect the similarities between

the newly merged cluster and all remaining clusters. This procedure is then repeated until all clusters are finally merged into a single cluster. Each new change in the number of clusters can be seen in a dendrogram hierarchically, which provide a useful tool for initially considering the best number of clusters, or type number candidates, at each level where the HCS can provide a new number of justified similarity clusters. Though this method alone is not sufficient for making the final determination of the number of category types.

Likewise, while this is a task of enormous importance to the current study, HCS is not an appropriate tool to determine the dimensions, number, definitions, or labels of the categories that are revealed in the hierarchy (please see Figure A-1, Figure B-1, Figure C-1, and Figure D-1 for visual representations of HCS hierarchy dendrograms). HCS, in addition, does not finally determine the number of relevant categories, or the perceptual dimensions that account for them, nor does it determine the capacity of N number of acoustic cues to account for the taxonomy.

For these tasks, MDS and LDA is needed. MDS refers to a class of techniques for exploring similarities and dissimilarities in data and developing a resulting matrix. In this case, MDS can help determine the number of dimensions and helps one determine the identity of the perceptual dimensions (i.e., for voice type purposes, a single acoustic cue or a combination of two or more cues). This is done by assigning voices to specific locations in an n-dimensional space (i.e., 1 – 5 dimensions) such that the distances between points in the space match the given dissimilarities as closely as possible. The number of dimensions can be estimated by picking the maximal number that provides a significant increase in the degree of fit between the original similarity judgments and the

transformed values used to plot the n-dimensional space. Of course, at some point, adding dimensions results in minimal gains in fit. On the down side, MDS does not determine the identity of the dimensions. For that, we can use discriminant analysis. Discriminant analysis can provide a predictive model for each dimension of the MDS analysis using linear combinations of the predictor variables, acoustic cues in this case, that provide the best discrimination between the dimensions.

As each technique has its utility and limitations, all three methods in conjunction are needed to develop a useful taxonomy of voice types. The process is to first code the voices by the categories generated by the HCS. Then, select N number of acoustic correlates. Each correlate can then be analyzed by combining them into a linear model, whereby one can learn the success of a given model in categorizing the voices on a 0 to 1 scale, 1 being perfect, through a process of regression. In this way, the relative importance of each cue can then be reported. However, this process is limited in that it cannot tell you the optimal number of perceptual dimensions used by actual listeners. This is a general limitation of machine analysis.

CHAPTER 2
PILOT STUDY

Methods

Overview

A pilot study has provided an excellent proving ground for the prudence and viability of these particular methods applied to vocal typology. An existing database, the same one that will be used for the eventual larger experiment was used for the pilot study. The original database consists of 150 voices total, 75 female and 75 male, each of which is evenly divided into “young” (18-35 years old), “middle-aged” (40-55), and “old” (over 55). In the pilot, “old” voices were discarded from the database, because it was assumed that the “old” voices were too aberrant from the other voices and would be clustered only be age, making the data not useful. The database consists of 16 “SPIN” (Speech Perception In Noise) sentences. These sentences compose a classic database, in which 14 are mirrored directly from an existing database. All are designed to hit the right phonetic balance as a representative set of phonemes in terms of frequency of occurrence in spoken American English. Only 10 of the 16 sentences from the existing database were selected at random from each voice for use in the pilot and full experiments.

With the culled database of 50 mixed young and middle-aged voices per sex, 500 trials were run per experiment, with mixed male and female untrained listeners participating in the experiments. 10 experiments were performed in total, with 10 individual listeners hearing 500 trials and judging similarity on a numerical scale of 1 to 7 (a Likert Scale), with 7 being the most similar and 1 being the least similar. All pairings were same to same sentence comparisons. These sentence pairings were randomly

generated, with each participant hearing a randomly generated set of 500 match-ups, that included same to same voice match-ups and opposite orderings of “Voice A” to “Voice B” and “Voice B” to “Voice A,” including same voices speaking different sentences, but only ever matched with same to same sentences by content. The decision to match only same to same sentences was made in consideration of ease or difficulty for the listener in making accurate perceptual judgments of similarity between voices. Having the least amount of differences in linguistic content, recording conditions and so on between the samples being compared allows for the greatest ease of comparison for the listener. Also, matching different sentences to different sentences would open the experiment to a greater degree of methodological criticism, as linguistic content incongruity might be seen as playing a difference in the rating of vocal similarity.

All experiments were performed at The Speech Perception Laboratory, Institute for Advanced Study of the Communication Processes (IASCP) at the University of Florida, with UF IRB approval (#2011-U-0828). The experiments were computer based and controlled for consistency across all 10 listener experiments. The sentences were played in brief trials with a less than 1 second pause between paired sentences. Each pairing could only be played once, but rating time was free to the listener. Each experiment took roughly 50 to 75 minutes per participant.

Stimulus Materials

A speech database, consisting of 100 voices total, 50 female and 50 male, each of which is evenly divided into “young” (18-35 years old) and “middle-aged” (40-55) native speakers of American English reading ten “SPIN” (Speech Perception In Noise) sentences was utilized. All speakers were recruited from the Gainesville, Florida, USA area. No attempt was made to control for dialect background, although all three groups

were predominately represented by individuals who had lived in many regions of the country over their lifespans. All participants were native speakers of American English with no known history of speech or hearing problems. Talkers were recorded in a quiet environment using a head worn microphone fixed at a constant distance from the corner of the mouth.

Participants

Ten native speakers of American English were recruited to participate in this study (21–35 years of age). Six were male, four were female, and none reported any history of speech or hearing problems.

Procedures

Similarity estimates were collected for the purpose of generating a complete matrix of comparisons for submission to a hierarchal clustering analysis in the initial step to form a voice type taxonomy. For this purpose, male and female voices were never compared to each other. Within each sex set, a given voice was paired with every other voice, in both orders (A-B, B-A) for a total of 2,500 trials. All pairings involved a common sentence, and all ten sentence types were randomly sampled from for trial generation. An interstimulus interval of 1s was used; the test was self-paced, and listeners were permitted only one opportunity to listen to each trial. Similarity was assessed using a seven point Likert Scale, with 7 being the most similar. Each listener did not rate all 2,500 trials. Rather, the trials were divided randomly into five sets of 500 trials each per sex (10 sets total). Each listener was assigned to one stimulus set. Each listener completed the similarity ratings task in 50 – 75 minutes.

Results

General

The 50*50 similarity matrix generated by the experiment was submitted to hierarchical clustering scheme (HCS) to permit the clustering of voices by perceived similarity. The procedure begins with a number of clusters equal to the number of cases that make up the matrix (i.e., 50 voices in the case of the pilot study). The clusters were denoted 1 through N. Next, the most similar pair of clusters were determined, merged through a method (i.e., arithmetic averaging), and then the matrix was revised to reflect the similarities between the newly merged cluster and all remaining clusters. This procedure was then repeated until all clusters were finally merged into a single cluster. The resulting dendrograms were then used to posit a male and female taxonomy of voice types based on two criteria: 1) Establishing a maximum number of descriptive categories and 2) Ensuring that each category had at least two members, fulfilling the goal of reducing large numbers of individual voices to similar, functional, descriptive units.

Male Voices

Using the criteria stated above, the fifty male voices were reduced to seven voice types of uneven membership. Some basic demographics of these types are provided in Table 2-1, including the percentage of voices in the database that were classified into each type, the mean actual age of the particular voice type, and the mean perceived age of the voice type (perceived age data was available for this database from Harnsberger, Shrivastav, & Brown, 2010). Several initial observations can be made from Table 2-1. First, voice types for males can be described in terms of a dominant group of modal voices with six other types of much smaller membership. Secondly, the

minority voice types tended to skew older relative to the modal voice type (labeled “m1” in Table 2-1), with the exception of type “m7.” Finally, the modest age effects were largely the same for perceived age as well as actual age. Vocal aging effects were not a major focus of this study, but were explored only to ascertain that vocal similarity judgments by listeners were not based purely on the age of the talkers.

Table 2-1. Demographics of male voice types

Voice Type	Size	Actual Age	Perceived Age
m1	66%	33	32
m2	8%	35	43
m3	6%	49	42
m4	6%	46	45
m5	4%	53	53
m6	4%	48	43
m7	6%	23	28

The HCS levels function to group all of the voices by similarity, but also provide the necessary framework for the eventual defining of parameters and correlates for each group. The HCS analysis yields dendrogram figures (Figures A-1 and B-1 in the Appendices), which display the divisions of similarity. The first division, starting from right to left, divides the voices into the initial two groups based on perceived similarity. Moving back from right to left, each division reveals another useful number of types at each level, with a potential number of levels from 2 to 50 per gender. Each number of groups can then be analyzed against independent variable acoustic correlates to make determinations about the parameters of each group, and to determine the right number of groups, using discriminant analysis. The dendrograms for both the female and male data are shown in Appendices A and B respectively. Some useful information can be gleaned by eyeballing the dendrograms. For example, in the female data dendrogram, one can observe that the first strong division yields an initial splinter group of 5 female

voices (mf11, mf26, mf46, mf44, and mf12). It might at this point be assumed that this particular division has age as one of it's significant independent variables, as all members of the group begin with "m" for "middle-aged." In Figure B-1, there is also a strong first division observed in the male samples. In this case, 6 male voices are shown to compose the first distinguished group by perceived similarity (mm51, mm52, mm13, ym41, mm50, and ym55). No discernable division by age can be observed here based on the speaker labels alone, as the group composes both young and middle-aged voices. Age as a more salient distinguishing characteristic for female voices than for male is born out by further study in the full experiment.

The seven male voice types were further explored in a auditory analysis by two phonetically-trained expert listeners. Age was not found to be a distinguishing characteristic in males in the same way as the females tended to show apparent saliency to age. Rather, the male voices tended to be associated with a hierarchy of acoustic and speech characteristic correlates. Five characteristics were identified as crucial to differentiating the seven voice types, and are represented in Table 2-2 in terms of a coding system developed to map acoustic-auditory characteristics to each type:

- Speaking Rate: S(low) vs. F(ast) – Relative to the average rate observed in the database
- Voice Quality: C(lear) vs. R(ough)
- Mean Pitch: H(igh) vs. L(ow)
- Nasality/Orality: O(ral) vs. N(asal) – Degree of atypical nasality in speech relative to normal utterances
- Pitch Variability: M(onotone) vs. D(ynamic)

- Other: A = Average values for database; X = Variable within the voice type for the characteristic in question

Table 2-2. Male voice types coded with five speech characteristics

Voice Type	SR	VQ	MP	PV	N/O
m1	A	C	A	D	O
m2	S	C	L	M	O
m3	A	R	H	A	O
m4	F	C	H	M	N
m5	F	C	L	M	O
m6	A	R	X	D	O
m7	F	C	X	D	O

This coding system presents another, more evolved VT labeling system from those proposed in the Introduction, for resulting voice types in the full experiment, e.g., ACADO for type “m1,” SCLMO for type “m2,” and so on. From the expert judgments, the voice type with the largest membership, “m1,” consisted of very “typical” voices, unremarkable in all characteristics with the exception of pitch variability, in which speaker members were judged as using a relatively broader pitch range as compared with “m2,” “m3,” “m4,” and “m5.” The acoustic-auditory characteristics that proved most useful were mean pitch, pitch variability, and speaking rate. Voice quality and nasality/orality differentiated only “m3” / “m6” and “m4,” respectively.

Female Voices

The same analysis applied to the female voices resulted in a smaller number of vocal categories, six rather than seven, voice types. Their demographics are given in Table 2-3. Unlike the male voice types, the fifty females voices cohered into two major categories (f1 and f6). Actual age appeared to play a role in some of these groupings: “f6” skewed young, while “f1,” “f2,” and “f4” skewed older. Considering perceived age,

the voice types can be ordered in increasing grades of age, with “f2” and “f6” as real outlier types in the taxonomy.

Table 2-3. Demographics of female voice types

Voice Type	Size	Actual Age	Perceived
f1	28%	46	39
f2	10%	51	54
f3	6%	36	35
f4	6%	52	43
f5	8%	33	30
f6	42%	20	22

The expert judgments of the speech characteristics that distinguish each group appear in Table 2-4, using the same coding system employed with the male voice types.

Table 2-4. Female voice types coded with five speech characteristics

Voice Type	SR	VQ	MP	PV	N/O
f1	F	C	X	M	O
f2	S	R	L	A	O
f3	A	C	A	D	O
f4	F	C	H	D	O
f5	F	C	A	M	O
f6	A	C	H	A	N

In this more balanced set of voice types by membership size, the largest group (f6) interestingly was not modal in all of its characteristics. It was the youngest group by far, and was judged to have the higher mean pitch voice, as well as being the only type with a greater degree of nasality. The second largest type (f1) spoke more rapidly and with a more restricted pitch range than the others. As with the male voice types, voice quality and nasality/orality were more limited in their utility as differentiating characteristics, determining the traits of one type from another.

Discussion and Conclusions

Two taxonomies of groups of similar voices, or “voice types” were separately developed for male and female voices using a database of 100 American English speakers. The resulting similarity matrices were submitted to a hierarchical clustering analysis, resulting in the identification of six female and seven male voice types. From the judgment of expert listeners, five speech and voice characteristics were required to uniquely identify these types and their characteristic dimensions, and the same set proved viable for both sexes: voice quality, nasality/orality, mean pitch, pitch variability, and speaking rate. These characteristic dimensions of each type were based on limited data in the pilot experiment and proved to be somewhat similar, but also significantly different in important ways to the dimensions that a full acoustic analysis and expert listener aural perceptual confirmation and verification analysis revealed to be the important distinguishing speech and voice characteristics that formed the dimensions of the final voice types that resulted from the full experiment of 100 listeners. Likewise, the number of types diverged from the pilot in the main experiment, where more data yielded appropriately similar, yet significantly different results. Some of these differences in pilot versus full experiment results were also a result of minor, but significant changes in the experimental design, in addition to the methodology of the analyses. The pilot experiment was highly useful to the main experiment both in providing validity to the thesis of this dissertation, i.e., that natural class voice types exist and human voices are not uniformly similar to one another and to revealing experimental design improvements toward yielding better and more comprehensive results in the full experiment.

A preliminary attempt was made to validate the judgments of the expert listeners with automated acoustic measures in the pilot experiment. Mean pitch and pitch variability were calculated using the Praat software (which employs an autocorrelation technique (Boersma, 1993)). Speaking rate was calculated in terms of syllables/second using custom software for syllable detection. Voice quality was measured using cepstral peak prominence (Hillenbrand & Houde, 1996). No attempt was made to quantify degree of overall atypical nasality in running speech. The voice quality measure proved to be significant in differentiating “f2,” “m3,” and “m6” from the counterparts, and both pitch measures and speaking rate showed the expected trends relative to the expert judgments, although not all pairwise comparisons proved to be significant. A much more comprehensive analysis is required in order to model these voice types and test them with larger numbers of speakers representing a greater range of dialects. This gap is significantly reduced through a full experiment with 10 times the amount of speakers and significant improvements in experimental design, analytical methodology, and the instillation of result confirmation and verification methods in the experimental process for the full experiment.

CHAPTER 3
EXPERIMENT 1: MATRIX AND POSITING CATEGORIES

General

Experiment 1 is the full experiment using untrained listener judgments on vocal similarity, with 100 listeners as opposed to the 10 used in the pilot study, and was performed with lessons learned from the pilot study through appropriate modification to the design and subsequent analysis. After input from the supervising committee and agreed revisions, changes were made to the experimental design to improve the validity of the results and method. Changes included the addition of calibration pairs that were uniform across all of the experiments. There were a total of 10 trial sets, 5 male and 5 female. In this design, unlike the pilot, each listener heard 540 trials—the same 500 as in the pilot, minus 10 from each of the five sex-specific sets of comparison pairs, to form the 50 calibration pairs. These uniform 50 calibration pairs was then added to each of the sets of 490, to make the new number of 540 trials per listener similarity judge. This resulted in an average listening time of 60 to 75 minutes per participant, for which the participants were paid at an hourly rate, as per the IRB approved experimental design.

The purpose of this experiment is to create the data matrix, based on data from 100 acceptable untrained listener judgment experiments (many more were actually performed, with invalid results based on set, standard criteria, being excluded pre-analysis), consisting of 540 trials each, 50 dealing with the male and 50 dealing with the female voice sets, and to use this data to posit the number of types for each sex, using common criteria and based on the data matrix.

Methods

Stimulus Materials and Procedures

An existing database, the same one that was used for the pilot study, was employed for the larger study. The database consists of 150 voices total, 75 female and 75 male, each of which is evenly divided into “young” (18-35 years old), “middle-aged” (40-55), and “old” (over 55). As in the pilot, “old” voices were discarded from the database, because it was assumed that the “old” voices were too aberrant from the other voices and would be clustered only be age, making the data not useful. Also, the pilot confirmed that saliency of age could be a problem if included in the full study, as listeners were probably affected by perceived age on a limited basis in the female pilot study (where age appeared to be more salient to the listener than with male voices—a proposition strengthened by the results of Experiment 1). This would be exacerbated in the full study and interfere with attempts at isolating the identity indexical property of the voice, which is a goal of this study generally. The database consists of 10 SPIN sentences. These sentences compose a classic database, where the 10 SPIN sentences used in this study’s vocal database are listed in Appendix E, along with the 6 unused sentences from the existing database. All are designed to strike the right phonetic balance as a representative set of phonemes in terms of frequency of occurrence in spoken American English, with the 10 used being selected from the total 16 at random, but using the same 10 uniformly across the total voices for this experiment’s database.

With the culled database of 50 mixed young and middle-aged voices per sex, 540 trials were run per experiment, with mixed male and female untrained listeners participating in the experiments. Over 125 experiments were performed in total, with all

individual listeners hearing 540 trials and judging similarity on a numerical scale of 1 to 7 (Likert Scale), with 7 being the most similar and 1 being the least similar. All pairings were same to same sentence comparisons by content. These sentence pairings were randomly generated, with each participant hearing a randomly generated set of 500 match-ups (minus the 10 pairs from each set for the 50 calibration pairs per sex, added back to the 490 pairs per set, for a total of 540 trials per listener), that included same to same voice match-ups and opposite orderings of “Voice A” to “Voice B” and “Voice B” to “Voice A,” including same voices speaking different sentences, but only ever matched with same to same sentences. The decision to match only same to same sentences was made in consideration of ease or difficulty for the listener in making accurate perceptual judgments of similarity between voices. Having the least amount of differences in linguistic content, recording conditions and so on between the samples being compared allows for the greatest ease of comparison for the listener. Also, matching different sentences to different sentences could expose the experiment to methodological issues, as linguistic content incongruity might be seen as playing a difference in the rating of vocal similarity.

All experiments were performed at The Speech Perception Laboratory, Institute for Advanced Study of the Communication Processes (IASCP) at the University of Florida, with UF IRB approval. The experiments were computer based and controlled for consistency across all 100+ listener experiments. The sentences were played in brief trials with a less than 1s pause between paired sentences. Each pairing could only be played once, but rating time was free to the listener. Each experiment took roughly 50 to 75 minutes per participant. Participants were recruited from the Gainesville, Florida,

USA area through advertisements in local media (print, online, social, and announcement), and all participants were compensated for their participation. All hundred participants used in the final data analysis were native speakers of American English, 18–34 years of age. 20 were male, 80 were female, and none reported any history of speech or hearing problems.

Matrix and Data Reduction

As with the pilot study, the data was subjected to HCS data reduction analysis to yield the dendrograms for female and male results separately (shown in Appendices C and D respectively). The resulting similarity matrix from the experimentation data was submitted to Hierarchical Clustering Scheme (HCS) analysis to group the voice types into "N" number of voice types, hierarchically (from as few as two to as many as half the voice sample size, 25 per gender). The resulting HCS levels function to group all of the voices by similarity, but also provide the necessary framework for the eventual defining of parameters and correlates for each group. The HCS analysis yields dendrogram figures, which display the divisions of similarity. The first division, starting right to left, divides the voices into the initial two groups based on perceived similarity. Moving back from right to left, each division reveals another useful number of types at each level, with a potential number of levels from 2 to 50 per gender. Each number of groups can then be analyzed against independent variable acoustic correlates to make determinations about the parameters of each group, and to determine the right number of groups, using discriminant analysis. The dendrograms for both the female and male data are shown as Figures C-1 and D-1. Some useful information can be gleaned by an initial visual inspection of the dendrograms. For example, in the first female data dendrogram, one can observe a strong, large group composed entirely of young, “y,”

voices. It might be assumed that this particular division has age as one of it's significant independent variables, as all members of the group begin with "y" for "young-aged."

In the case of the females, Figure C-1 (Appendix C), initial group membership in several cases was heavily reliant on and determined by age. At the 4 group (2^{nd}) level in the dendrogram, it can be observed that "group 1" is composed exclusively of "yf" or "young female" members. This is not desirable, as age would appear to be the sole or most important characteristic determining group membership in females at the initial level. This was confirmed by expert aural perceptual auditory survey of group membership at that level.

Figure D-1, there is also a strong first division observed in the male samples. In this case, a large group of 41-43 male voices is persistent at the 2, 3, and 4 group levels in the dendrogram. This is significantly larger and more persistent of a bimodal distribution than was seen with the female voices. This distribution is disfavorable in creating a useful, working taxonomy. However, it is not dominated by perceived age saliency as is the case with early groupings in the female voice taxonomy. Though bimodal distribution is a challenge to the insurance of significant data reduction in both the male and female HCS generated taxonomy, age saliency as a dominant interference at the early levels of discrimination for this sample is not a problem for the judgment of male voices in the same way that it was for the female voices. No discernable division by age can be observed here based on the speaker labels alone, as the group composes both young and middle-aged voices.

Structure of Taxonomy

For determining the number of types from the matrix, it is necessary to adhere to a set methodology for both the male and female data. Potential group number candidates

revealed from the HCS analysis present even distribution of membership number at several levels, as well as approximate consistency between the male and female group number that is highly desirable from an optics standpoint for the overall conclusions. However, the methodology used incorporates not only the utility and desire for consistency and evenly distributed membership, while also using a set and scientifically ground criteria for the final vocal typology structure.

For the pilot, deciding the amount of categories was a tentative decision, due to the smaller amount of perceptual data (5 listeners per gender). In the case of Experiment 1, this was not an issue, and stricter criteria was employed to determine the structure of the taxonomy. For the male or female participants respectively, only each level of the dendrogram can be a potential candidate for the number of types. That is to say, separation of categories in the statistical model does not actually happen at every potential level, 2 through 50. Each potential candidate "type set" level number was analyzed via its membership and potential utility for the optimal amount of candidates. For the males, 10 types were posited and for the females 9 groups were posited. These groups were based on three primary criteria:

1. Establishing a maximum number of descriptive categories
2. Ensuring significant data reduction (no bimodal distribution), and
3. Breaking up groups dominated by age saliency alone, e.g., young females.

A fourth criteria of ensuring that each category had at least two members, fulfilling the goal of reducing large numbers of individual voices to similar, functional, descriptive units was considered, but found to be without cause or sufficient merit in this experiment, due to the limited size of the database. One member groups may in fact not be "outliers," but rather groups with under-represented, smaller membership, due to the

size of the database (50 per sex, after separation of male and female and the discarding of “old” voices from this experiment). It could not be determined if single-member groups were outlier or unique voices, as opposed to simply being a group of less members (under-represented membership in this database). Even “outlier” voices, such as those of unique celebrities along the lines of Gilbert Gottfried, may actually represent a small portion of the general population “type,” (an underrepresented type), as opposed to being strictly “unique” as a characteristic. Some of the voices that composed one membership groups may simply be under-represented in this particular sample of 100 voices, and, therefore, cannot be discounted as outliers, rather than groups in their own right from the present study only.

Figures 3-1 and 3-2, show the number of group candidates (as seen at each level of the dendrograms) plotted against the maximum group membership that occurs in the largest group at each potential candidate number of groups. In Figure 3-1, the female voice type group number candidates can be seen to exhibit significant “elbows” at 4 groups, 7 groups, and 9 groups. This information can be seen in more detail in Table 3-1, where each group number candidate level is listed along with the maximum group membership for each candidate, with the number of 1 member groups (under-represented or possible outlier/unique voices) indicated at each level where applicable. Though the 4 group candidate shows the largest elbow and is therefore an appealing candidate, it does not satisfy criteria #3 (as can be seen on the female taxonomy dendrogram), as the first group would be composed entirely of “yf,” voices, or young females, meaning that group is likely dominated by age saliency alone. The 7 group candidate level provides a smaller elbow in maximum membership, which helps to

satisfy criteria #2, in that it reduces bimodal distribution and ensures significant data reduction, while slightly breaking the young female group up. However, the 9 group member best satisfies all three criteria by establishing a maximal number of descriptive categories, ensuring no bi-modal distribution and further breaking up the age salient first group of young females and dividing a later cluster of “mf” middle-aged females significantly. Additional candidates at 11 and 15 groups do not exhibit another elbow in ensuring significant data reduction, nor do they contribute significantly to the other 2 criteria.

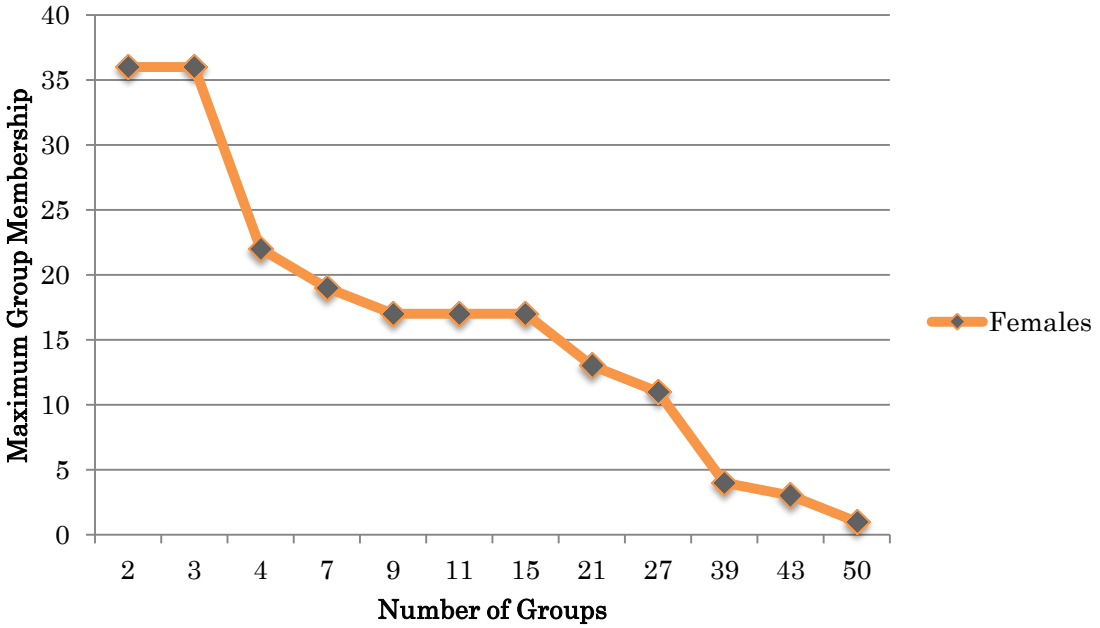

Figure 3-1. Female group membership at each dendrogram level by group number

In addition to the maximum group membership at each group number candidate level, Table 3-1 also shows the number of groups excluding 1 member groups. As stated earlier, this second number of groups represents a possible modification of the existing and final typology. Outlier voices were deemed, in the absence of further

evidence to the contrary or additional experimentation to be under-represented groups in this database, rather than “outlier” or “unique voices in the larger taxonomy of American English voice types. In the presence of research on a larger database or some other additional evidence, these under-represented groups could change in status, and have for this reason, been specifically identified here.

Table 3-1. Female group membership candidates at each dendrogram level with maximum membership and number of groups excluding outliers; with optimal candidate indicated

Group Candidate Number by Dendrogram Level with 1 Group Membership Noted	Maximum Group Membership	Number of Groups Excluding 1 Member Groups (Outliers)
2 Groups	36	2
3 Groups	36	3
4 Groups	22	4
7 (1 Outlier*)	19	6
9 (1 Outlier*) [Optimal]	17	8
11 (2 Outliers*)	17	9
15 (5 Outliers*)	17	10
21 (9 Outliers*)	13	12
27 (15 Outliers*)	11	12
39 (32 Outliers*)	4	7
43 (38 Outliers*)	3	5
50 (All Single Membership)	1	50

*”Outliers” are regarded in the current methodology for determining the structure of the taxonomy to be under-represented groups, absent further evidence to the contrary from further experimentation

In the case of the male voice taxonomy, maximum group membership is shown to create “elbows” via the number of group candidate levels most significantly at 6 groups and 12 groups. The 6 group candidate does little to satisfy the criteria, as there is both strong bimodal distribution and a lack of significant and useful data reduction. There is, however, a significant data reduction “elbow” at 12 groups, which provides excellent data reduction, while establishing a maximum number of descriptive categories. However, as will be outlined in the acoustic analysis, 2 of the 1 member

groups were determined through expert listener aural perceptual analysis to be classified exclusively or near exclusively by dialect interference for those 2 voices in the male voice database. Therefore, those 2 groups were excluded as useful descriptive units of identity as an indexical property in the typology. The validity of using to the aural perceptual expert listening experiment as a criteria to augment slightly the results of the HCS and acoustical analysis was confirmed by the otherwise total lack of conflict between the acoustic and aural perceptual test results on the discernable speech and speaker characteristic correlates of each type.

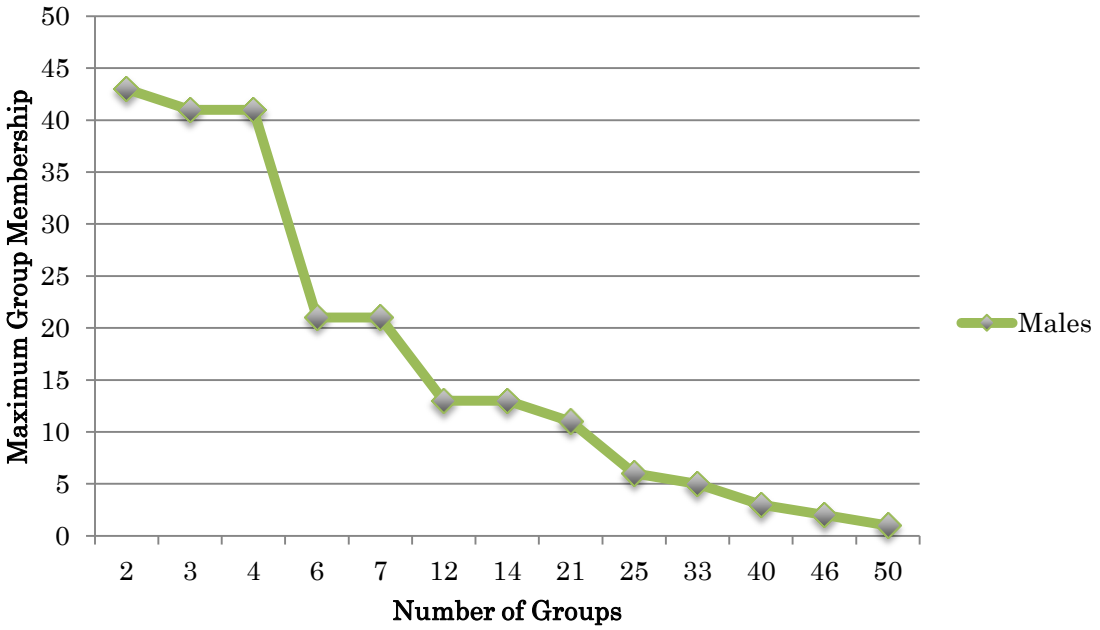

Figure 3-2. Male group membership at each dendrogram level by group number

Table 3-2 again displays not only the maximum group membership of each group number candidate level, but also denotes how many under-represented, 1 member groups or “outlier” voices were found to exist in each case. This information is extraneous to the current typology, but could change if either the methodology were

altered or in the face of additional evidence from further research on a larger database of American English voices.

Table 3-2. Male group membership candidates at each dendrogram level with maximum membership and number of groups excluding outliers; with optimal candidate indicated

Group Candidate Number by Dendrogram Level with 1 Group Membership Noted	Maximum Group Membership	Number of Groups Excluding 1 Member Groups (Outliers)
2 Groups	43	2
3 Groups	41	3
4 Groups	41	4
6 (1 Outlier*)	21	5
7 (1 Outlier*)	21	6
**12 (3 Outliers*) [Optimal]	13	9
14 (6 Outliers*)	13	8
21 (11 Outliers*)	11	10
25 (14 Outliers*)	6	11
33 (31 Outliers*)	5	2
40 (37 Outliers*)	3	3
46 (42 Outliers*)	2	4
50 (All Single Membership)	1	50

*"Outliers" are regarded in the current methodology for determining the structure of the taxonomy to be under-represented groups, absent further evidence to the contrary from further experimentation

**12 groups is reduced to 10 groups after Experiment 3 (expert listener aural perceptual evaluation), where 2 groups are judged to be determined by dialect saliency alone)

Results

Tables 3-3 and 3-4 show the final structure of the taxonomy by group, including new labeling appropriate to the acoustic modeling results and with full membership listing. For the females, 9 groups were determined to exist after the above stated criteria for group number candidate selection was applied. This number additionally accomplishes the goal of providing a "useful" set of working voice types, e.g., one that can be cognitively grasped by a practitioner or researcher attempting to apply the

groups to a forensic analysis or to additional study on the emotional response and other possible vocal pleasantness based correlates to each type.

Table 3-3. 9 Female voice types with size and membership

Voice Type	Size (Count)	Members
F-VT-01	17	yf01, yf03, yf04, yf19, yf20, yf21, yf24, yf30, yf31, yf33, yf38, yf39, yf44, yf49, yf57, yf58, yf59
F-VT-02	14	mf10, mf14, mf22, mf28, mf29, mf33, mf34, mf36, mf41, mf42, mf49, yf28, yf35, yf60
F-VT-03	6	mf17, mf23, mf25, mf32, mf38, mf47
F-VT-04	3	mf19, mf30, mf37
F-VT-05	3	mf11, mf26, mf44
F-VT-06	2	mf12, mf46
F-VT-07	2	yf23, yf37
F-VT-08	2	yf27, yf34
F-VT-09	1	yf65

Similarly, the male types also accomplish this goal, with a working set of 10 groups, according to the same criteria used to establish the female types. The membership of the largest group (maximum group membership) was 13, well satisfying the criteria of reducing the data to a useful number of groups with no bimodal distribution or age saliency domination in dictating of the membership. As stated earlier, Table 3-4 makes note of the elimination of 2 of the groups during the expert aural perceptual phase of the acoustic analysis. These 2 groups were determined through two independent aural perceptual tests by two different expert listeners to have been grouped by dialect interference in these two cases.

Table 3-4. 12* Male voice types with size and membership

Voice Type	Size (Count)	Members
M-VT-01	13	mm40, mm48, mm59, mm60, ym02, ym08, ym26, ym43, ym50, ym51, ym52, ym61, ym63
M-VT-02	12	mm21, mm24, mm45, ym17, ym32, ym36, ym40, ym45, ym46, ym47, ym53, ym64
M-VT-03	5	mm39, mm43, mm55, mm56, ym42
M-VT-04	4	mm03, mm27, mm57, mm58
M-VT-05	4	mm04, mm31, mm53, mm54
M-VT-06	3	mm18, ym48, ym62
*M-VT-07	2	mm13, ym41
M-VT-08	2	ym05, ym54
M-VT-09	2	mm50, ym55
*M-VT-10	1	mm51
M-VT-11	1	mm35
M-VT-12	1	mm52

*Groups 7 and 10 for the Males were discounted after expert aural perceptual listener experimentation determined them to be a result of dialect interference

The structures of the taxonomies for both male and female voices was determined according to common criteria and yielded a useful, workable set of voice types for each sex independently. The HCS analyses of the male and female voices relied upon averaged raw similarity scores, with each mean representing ten observations (one each from ten different listeners). To determine the stability of the derived taxonomies relative to potential individual listener differences in bias in similarity ratings, each raw score from an individual listener was converted to z-scores, and the similarity matrix was recomputed using averaged z-scores and submitted to HCS analyses (again, divided by gender). The resulting dendrograms of both the male and female voices were extremely similar to those observed with the raw values. Broad trends common to both analyses included a larger modal male voice type, strong age effects among the female

voices, and a greater plurality in group size of female voice types relative to male voice types. Agreement in voice type membership was high. Among male voice types, 90% remained grouped with the same voices as in the HCS of the raw similarity values. The corresponding female rate was 94%. Of the five male voices that changed groups, three involved one-member groups clustering into larger groups. Of the three female voices that changed groups, all of these involved exchanges between F-VT-01, F-VT-02, and F-VT-03.

Next, acoustic analysis, with aural perceptual expert confirmation was performed to determine the dimensions and speech and speaker correlates for each type within each sex. This analysis will also yield more meaningful, non-random labeling for the types within the posited categories and newly established vocal typology.

CHAPTER 4
EXPERIMENT 2: MODELING

General

Modeling was performed to determine the dimensions and acoustic correlates of each of the newly established nine female and twelve male voice types. The eventual goal of the acoustic model would be the ability to classify new voices into their appropriate voice type. This task would be facilitated by a reliable acoustic “profile” for each type, which could then be coded for semi-automatic or even fully automatic classification of new incoming voices to determine voice type. Incoming voices might initially or in cases of degraded evidence materials or less than perfect speech samples be typed by type range ID, rather than an exact match. Incoming voices could be identified as belonging to one of a range of limited types within the typology, accomplishing the goal of reducing the incoming voice to belonging to a subset of the general population, or possibly to exclude the evidence voice from belonging to the same range of possible types as the suspect’s voice, in the case of forensic SPID.

Prior work in the correlates of SPID by human listeners have identified a broad range of potential voice and speech cues, such as speaking fundamental frequency characteristics (Atal, 1972; Iles, 1972; Jassem, et al., 1973; LaRiviere, 1975; Mead, 1974), nasality (Glenn & Kleiner, 1976; Su, et al., 1974), temporal speech features (Abberton & Fourcin, 1978; Johnson, et al., 1984), voice quality (Hollien & Majewski, 1977; Johnson, et al., 1977; Zalewski, et al., 1975), and articulation, including fricative production (Ingemann, 1968; Schuartz, 1986). As speaker vocal types have never been formally studied, a broader aural-perceptual analysis was first undertaken to identify the full range of possible cues to consider. Accordingly, it was an open set task, where the

cues listed above were used to create a standard inventory of seven cue categories, plus a general classification of nonstandard “Dialect”:

1. Dialect
2. Nasal
3. Voice Quality
4. Articulatory Effort
5. Speaking Rate
6. Mean Pitch
7. Pitch Variability
8. Fricatives

Dialect was included to determine if any of the voice types generated in Experiment 1 reflected groupings of speakers by common socio- or regional dialect rather than the intended focus on voice type. In addition, an option was available to provide alternative or more detailed characteristics to guide the acoustic modeling effort. All 500 voice samples in Experiment 1 were classified by two phonetically-trained expert listeners. Given the open set nature of the task, a true reliability comparison was not possible. However, the modal choices of each listener for each talker could be compared directly. Taking this approach yielded an estimated agreement rate of 74% across all talkers classified. No other unique characteristics received more than one occurrence, with the exception of a characteristic labeled “End Rise,” referring to a terminal increase in pitch. This term was used in 0.5% of samples. A collation of the percentage of speech samples that were classified with the predicted categories are shown in Figure 4-1 for male voices, and Figure 4-2 for female voices.

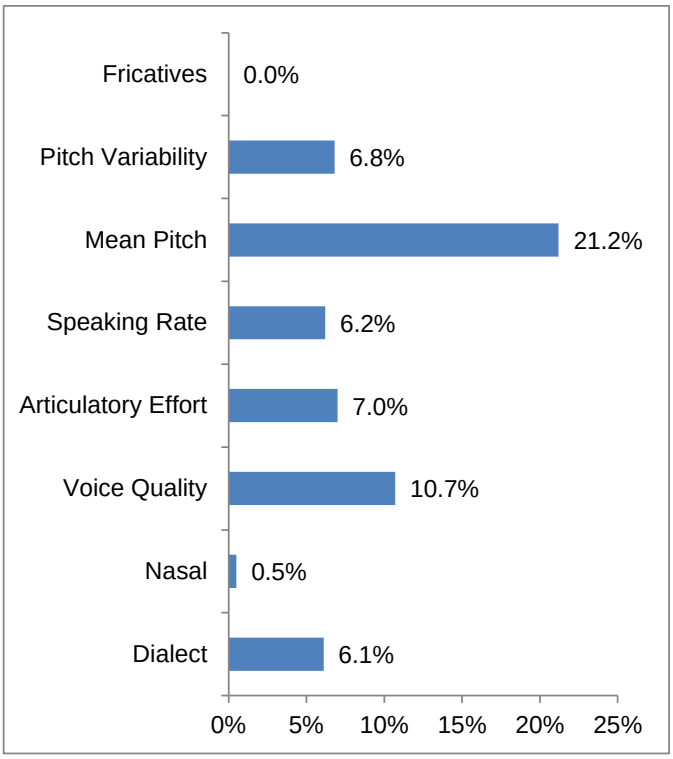

Figure 4-1. Percentage of male speech samples classified using a predetermined cue

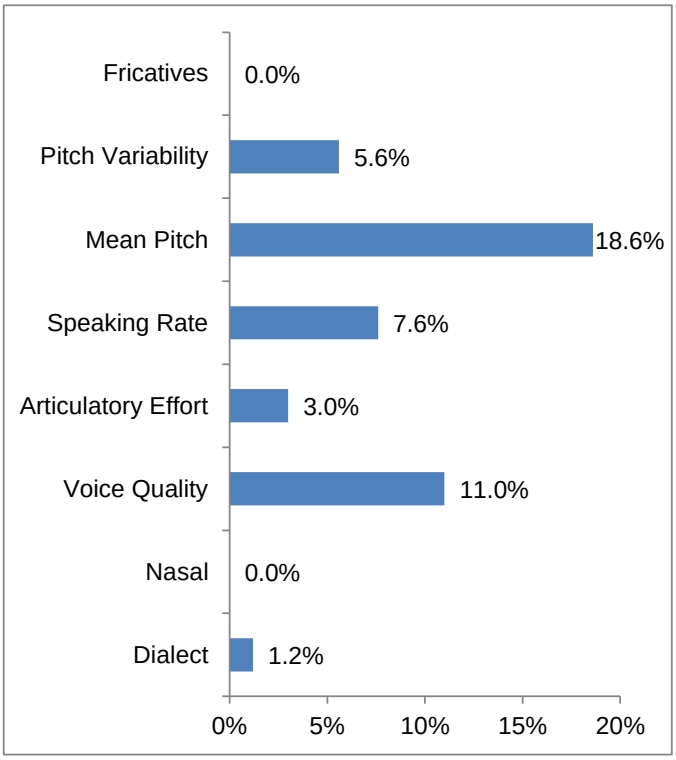

Figure 4-2. Percentage of female speech samples classified using a predetermined cue

As expected, the main categories of cues were utilized by both raters in the aural-perceptual task, with two possible exceptions: use of fricatives and the nasality/orality distinction. In addition, Dialect was also applied infrequently, although it represented over 6% of the responses to the male voices. Finally, while cross-gender trends were not of direct interest to the study, it can also be observed that individual cue types elicited common percentages among both male and female voices, with the exception of articulatory effort, which was applied more frequently among male voices to reflect the subset of materials that were perceived as hypoarticulated.

The voices classified as exhibiting some type of nonstandard dialect were examined further to determine if 1) any of the voice types generated in Experiment 1 consisted of a majority of voices that were classified with a nonstandard dialect and, if so, 2) did those voices share a common perceived nonstandard dialect. Two voice types derived from the HCS analysis met both of these criteria: M-VT-07 and M-VT-10, which constitute just 6% of the male voices in the speech database. No female voices were consistently labeled as speakers of nonstandard dialects, and thus no resulting voice types can be said to be a byproduct of a dialect-based similarity judgment rather than a voice type judgment. Given the results of the aural-perceptual analysis, the acoustic analysis focused on the five classes of cues that were both originally identified in the literature on individual SPID and were utilized in 5% or more of the perceptual judgments.

Methods of Acoustic Analysis

Acoustic measures of mean fundamental frequency, fundamental frequency variability, voice quality, articulation, and speaking rate cues were taken either manually or automatically from the 100 male and female voice samples of the speech database,

minus the three male voices that were classified as nonstandard dialect speakers. Effectively, this revision left ten male voice types to go with the nine female voice types. The male voice types were recoded with the elimination of two types, and are listed in Table 4-1. Henceforth, voice type codes are defined in the text by Table 4-1 rather than Table 3-4. The recoding simply omitted types M-VT-07 and M-VT-10 and recoded MT-V-08, MT-V-09, MT-V-11, and MT-V-12 into MT-V-07, MT-V-08, MT-V-09, and MT-V-10, respectively.

Table 4-1. 10 Male voice types with size and membership

Voice Type	Size (Count)	Members
M-VT-01	13	mm40, mm48, mm59, mm60, ym02, ym08, ym26, ym43, ym50, ym51, ym52, ym61, ym63
M-VT-02	12	mm21, mm24, mm45, ym17, ym32, ym36, ym40, ym45, ym46, ym47, ym53, ym64
M-VT-03	5	mm39, mm43, mm55, mm56, ym42
M-VT-04	4	mm03, mm27, mm57, mm58
M-VT-05	4	mm04, mm31, mm53, mm54
M-VT-06	3	mm18, ym48, ym62
M-VT-07	2	ym05, ym54
M-VT-08	2	mm50, ym55
M-VT-09	1	mm35
M-VT-10	1	mm52

Table 4-2 provides a list of all measures taken, the software tool utilized in data collection, and references (when possible) for the particular algorithm used for the measure. The mean pitch, pitch variability, and speaking rate cues are ones commonly utilized to examine many linguistic and indexical properties of speech. The latter two require some elaboration. Voice quality refers to the perception to the acoustic qualities of the motion of the vocal folds. Nonmodal voice quality can vary in degree and in

manner, the latter incorporating such terms as “breathy,” “rough,” and “strained.” Numerous measures have been proposed to quantify nonmodal voice in general or more specific terms. For this application, a current, widely-used measure was employed, cepstral peak prominence (with smoothing). Cepstral peak prominence is defined as “a measure of the amplitude of the cepstral peak corresponding to the fundamental period, normalized for overall signal amplitude” (Hillenbrand and Houde, 1996). A cepstral peak, in turn, is the prominence in a log power spectrum of a log power spectrum of an acoustic signal.

Table 4-2. All cues measured in the acoustic analysis, the software tool utilized and references (when applicable) for the particular algorithm used for the measure

Cue Type	Measure	Reference/Method
Mean Pitch	F0 Mean	Boersma (1993)
Pitch Variability	F0 σ / Mean	Boersma (1993)
Speaking Rate	Total Duration	Milenkovic (2001)
Voice Quality	CPP-S	Hillenbrand and Houde (1996)
Articulatory Effort	Vowel Space (Mean)	(See Text)

Articulatory effort in individual voices was assessed by estimating the average vowel space volume of the speaker measuring the first two formant frequencies of two tokens of each of three target vowels ([a], [i], [u]) from stressed syllables across six randomly selected sentences (for examples, see Johnson, Flemming, & Wright, 1993 and Harnsberger, Shrivastav, & Brown, 2010). Formant measures were taken from a combined FFT and LPC (14-coefficient autoregressive analysis) display, with a 25ms analysis window at the temporal midpoint of the vowel. These frequencies in Hertz were converted to Barks and then used to calculate distances in an F1-F2 Bark space for all three pairings of vowels ([a]-[i], [i]-[u], [a]-[u]). These three distances were then averaged for a final score for mean vowel space size. This measure was explored to

determine if individual voice types displayed any greater propensity to produce less effortful, more hypoarticulated speech.

The goal of the acoustic analysis was to both describe the common characteristic of each voice type for each gender and, in turn, examine the relative importance of different categories of cues for voice types in general. To this end, descriptive statistics were generated for each voice type, and all measures were submitted to linear discriminant analysis to examine the rank order of weights applied to each type of acoustic cue.

Results

The means of each of the five cue categories are shown in Figures 4–3 through 4-7 respectively for mean pitch, pitch variability, speaking rate, voice quality, and articulatory effort (based on vowel space size) for the ten male voice types. In total, 50 values are reported, and thus, these results are simplified in Table 4-3, which use the acoustic measures to code the types in terms of extreme measures along each acoustic dimension. Voice types whose mean value was greater than one standard deviation from the group mean received a code, while all means within one standard deviation remained unspecified. The corresponding results for the nine female voice types are shown in Figures 4–8 through 4–12 and Table 4–4. The codes in Tables 4–3 and 4–4 are defined as follows:

1. Speaking Rate: (S)low vs. (F)ast
2. Voice Quality: (C)lear vs. (N)onmodal
3. Mean Pitch: (L)ow vs. (H)igh
4. Pitch Variability: (M)onotone vs. (D)ynamic
5. Articulatory Effort: (R)educed vs. (E)nunciated

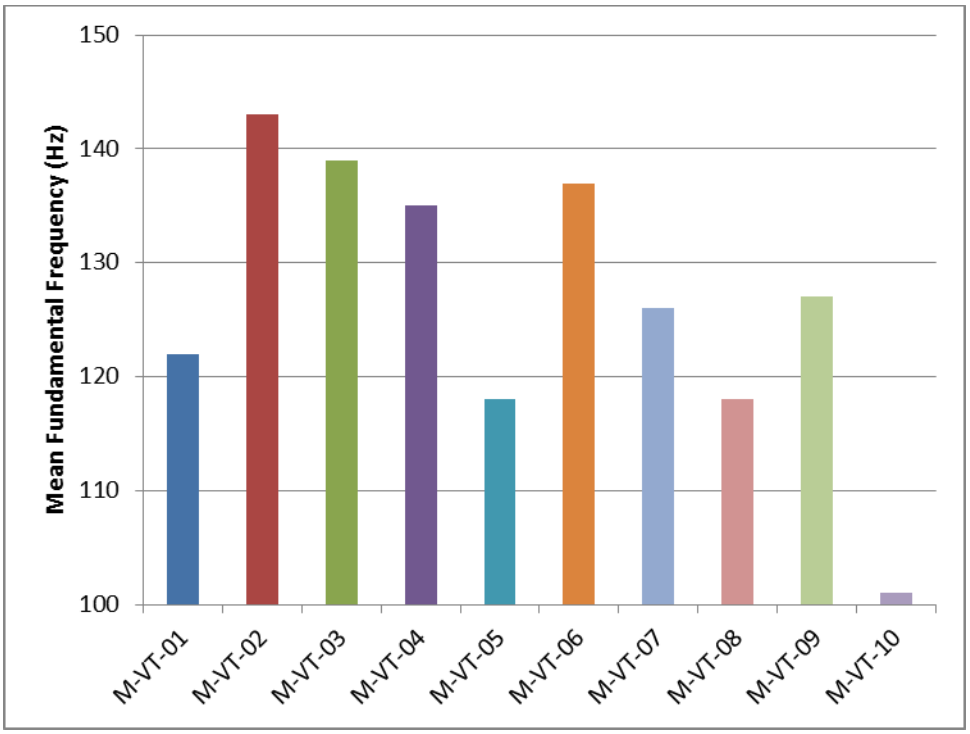

Figure 4-3. Mean fundamental frequency (mean pitch) for ten male voice types

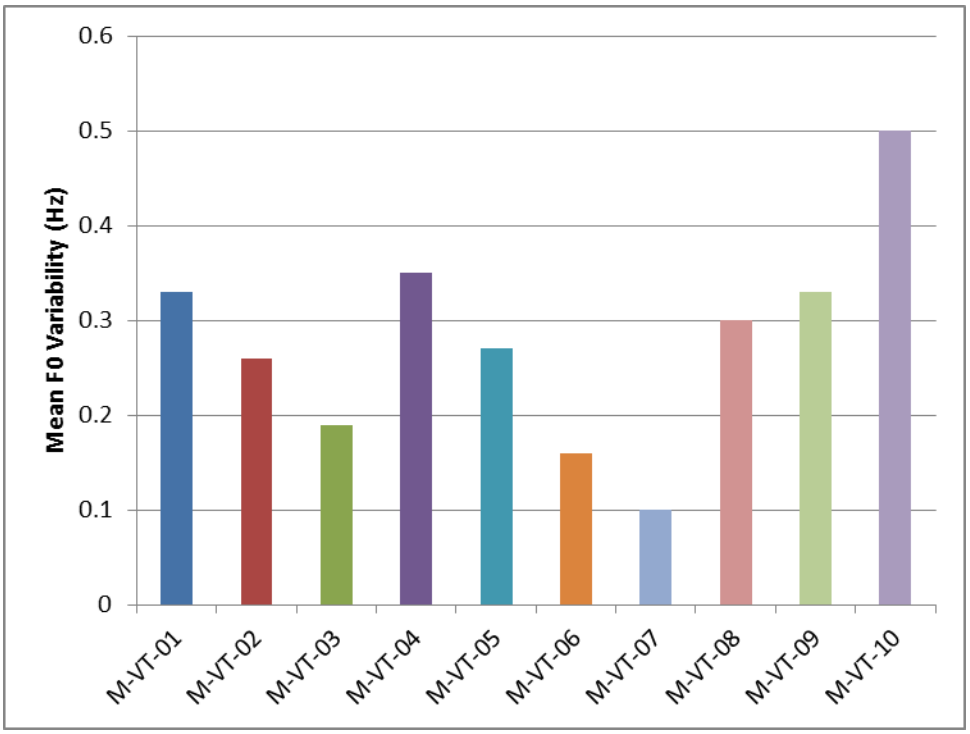

Figure 4-4. Mean fundamental frequency variability for ten male voice types

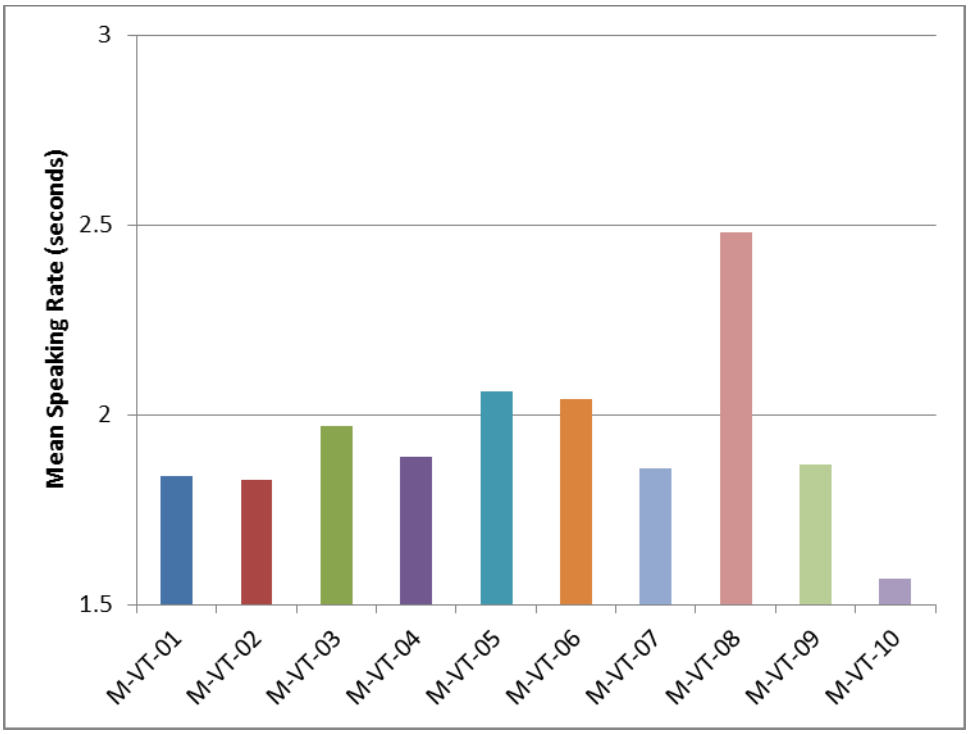

Figure 4-5. Mean speaking rate (overall duration) for ten male voice types

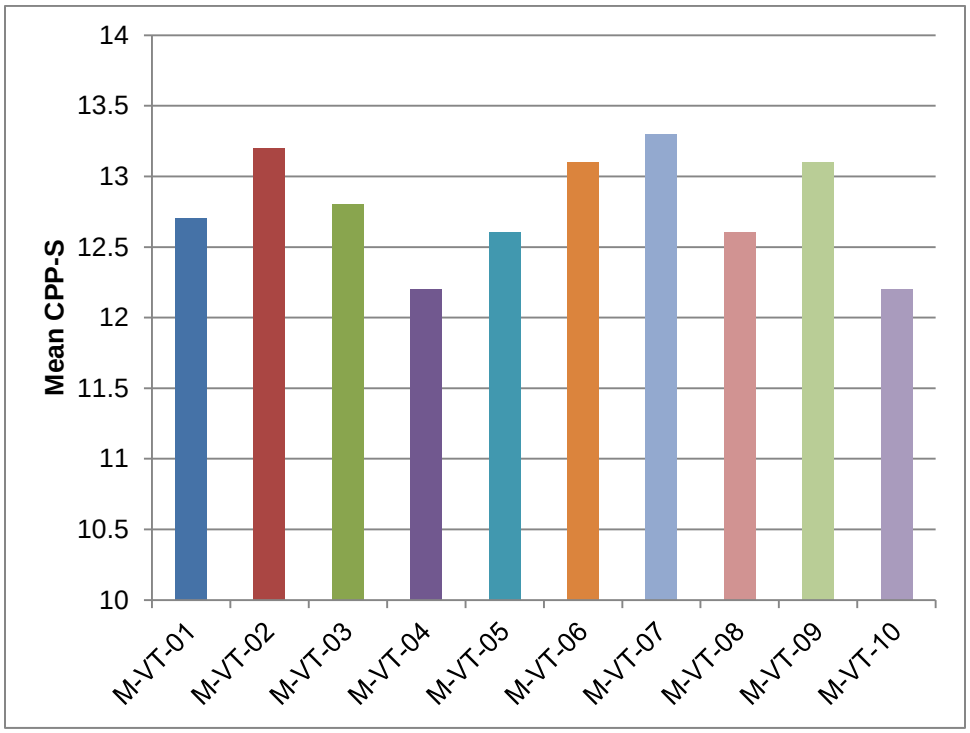

Figure 4-6. Mean CPP-S (for nonmodal voice quality) for ten male voice types

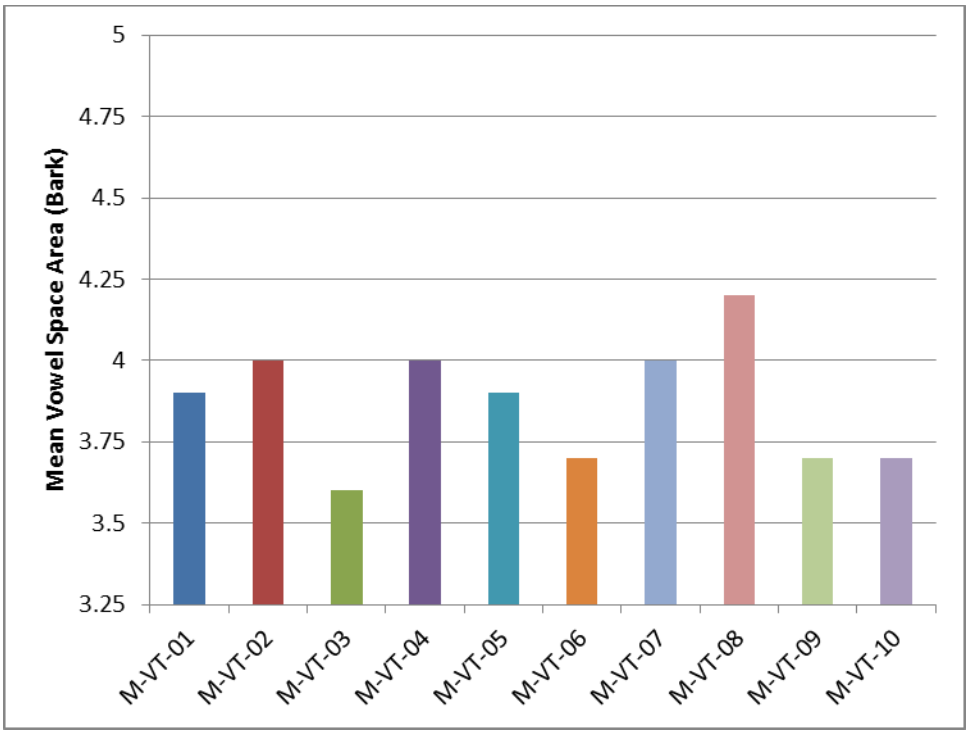

Figure 4-7. Mean articulatory effort (measured in vowel space area) for ten male voice types

Table 4–3 reports the degree of success in using five categories of acoustic cues to categorize male voices into ten voice types. It must first be noted that the modal voice type, M-VT-01, was also modal in all five characteristics. Secondly, seven of the voice types, M-VT-01 through M-VT-07, could be differentiated via just three characteristics, mean pitch, pitch variability, and voice quality. In total they constitute 92% of the voices in the database. While M-VT-03 also showed a more hypoarticulated vowel space, this feature could have been excluded and the type characterized by just its relatively high pitch and reduced pitch variability. Of the top three voice types in terms of size, they can be most parsimoniously characterized as Male Modal, Male High and Clear, and Male Monotone. The remaining voice types represent, individually, 2%-8% of the male voice speech database. M-VT-10 is populated by a single voice, which is highly unique

compared to the others, raising the issue of whether or not it stands in for a larger group in the population of American English males or whether it is an idiosyncratic voice.

Table 4-3. Recording of the acoustic analysis of the male voice types into extreme values along the five acoustic dimensions explored

Voice Type	Size (Count)	Speaking Rate	Voice Quality	Mean Pitch	Pitch Variability	Articulatory Effort
M-VT-01	13					
M-VT-02	12		C	H		
M-VT-03	5			H	M	R
M-VT-04	4		N		D	
M-VT-05	4			L		
M-VT-06	3				M	R
M-VT-07	2		C		M	
M-VT-08	2	S		L		E
M-VT-09	1					R
M-VT-10	1	F	N	L	D	R

*See text code for definitions

As stated earlier, in the group number decision criteria for the structure of the taxonomy of the vocal typology in this study, in cases such as M-VT-10, i.e., single voice membership groups, the groups were identified as types. This decision was made both for lack of firm evidence that these single membership groups represented idiosyncratic voices and also in recognition of general realities in dealing with a sex-specific database of only 50 voices. While it is impossible to determine whether or not a group of this type represents an idiosyncratic voice, it is more likely that the group is simply an under-represented type in the database. Likewise, expert listener aural-perceptual testing confirmed, to the degree possible, that the voice was not likely a strictly idiosyncratic voice among the general population of speakers of American

English. The issue of “uniqueness” of a human voice is more likely an issue of degree of rarity, than it is an issue of strict idiosyncrasy among the general population of speakers of American English or any other large population of speakers. A primary motivation of this study was he assumption that human voices are not uniformly similar to one another. Degree of similarity or dissimilarity is a primary motivator for both the distance, in terms of similarity of one type to another in the taxonomy, and also for the size of group types within the similarity. As the largest groups are shown to be the most under-specified by acoustic correlate code, or “average” in nature compared to the larger population, the most acoustically specific coded groups are likewise the smallest or most under-represented types.

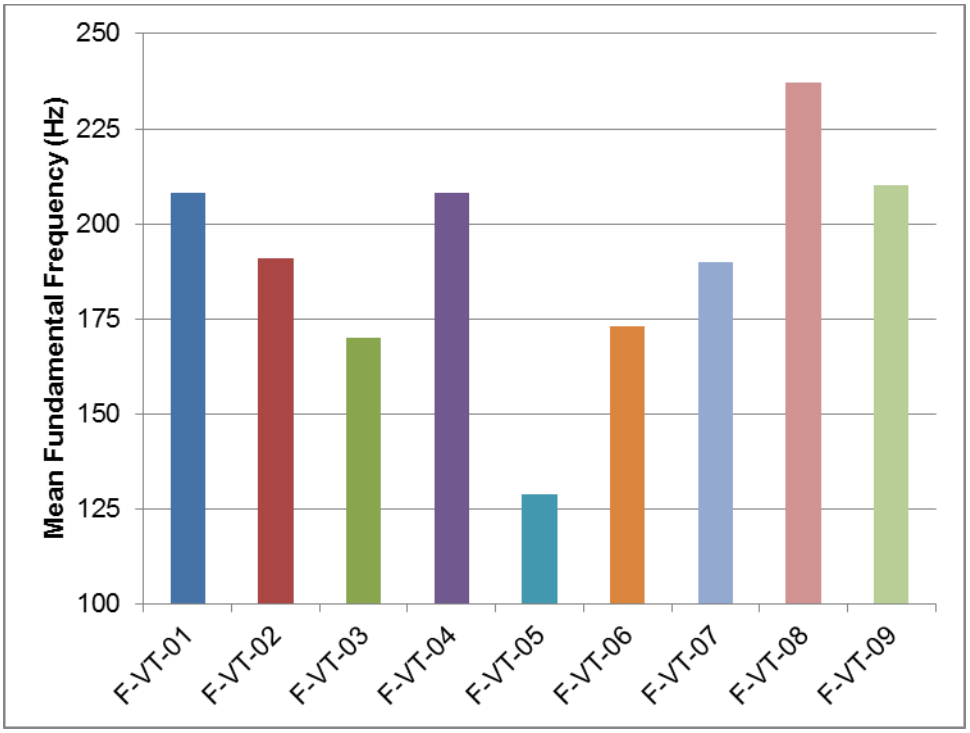

Figure 4-8. Mean fundamental frequency (mean pitch) for nine female voice types

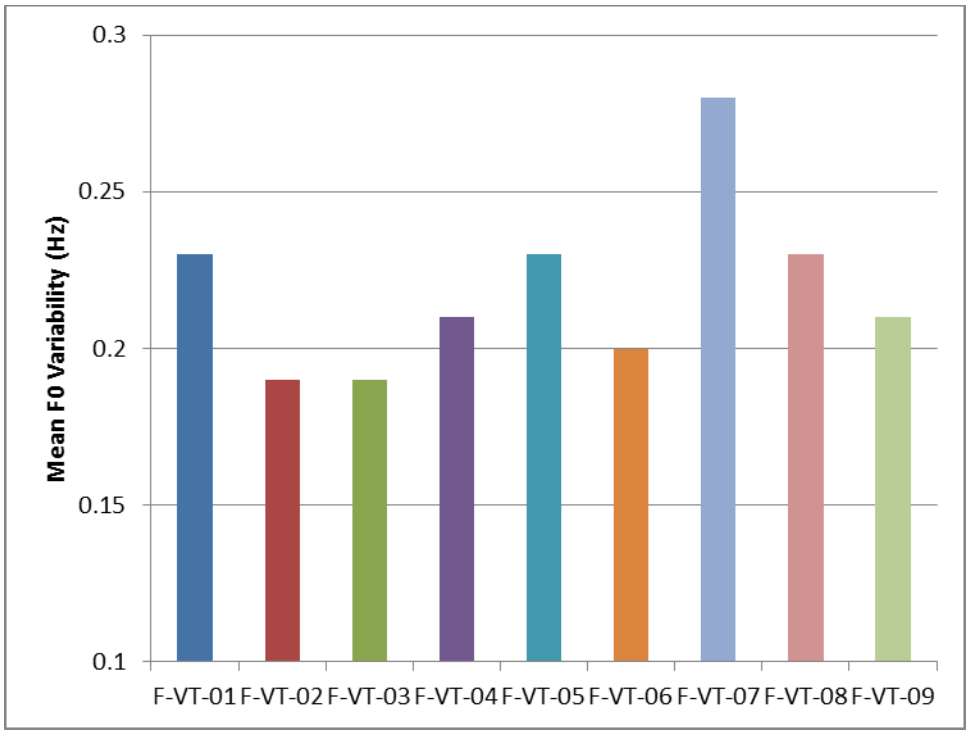

Figure 4-9. Mean fundamental frequency variability for nine female voice types

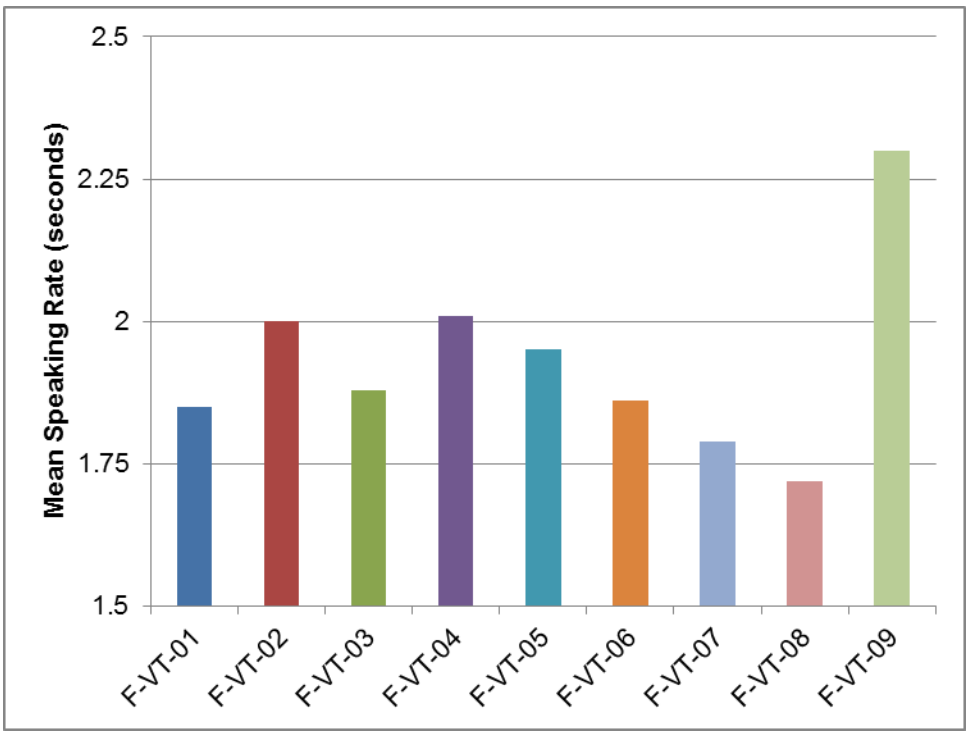

Figure 4-10. Mean speaking rate (overall duration) for nine female voice types

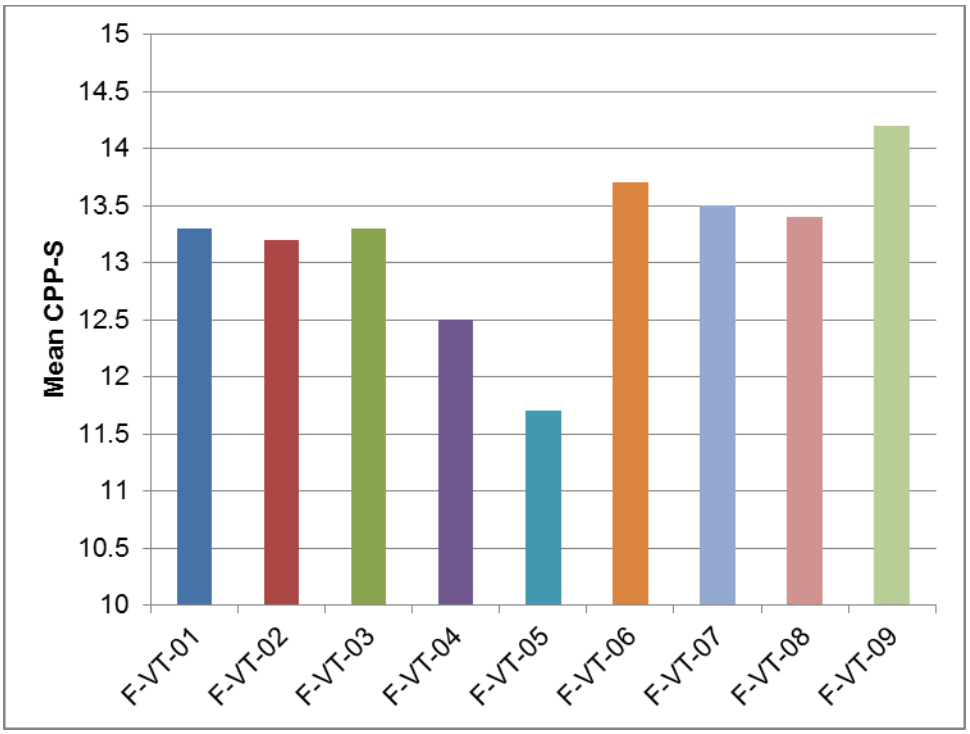

Figure 4-11. Mean CPP-S (for nonmodal voice quality) for nine female voice types

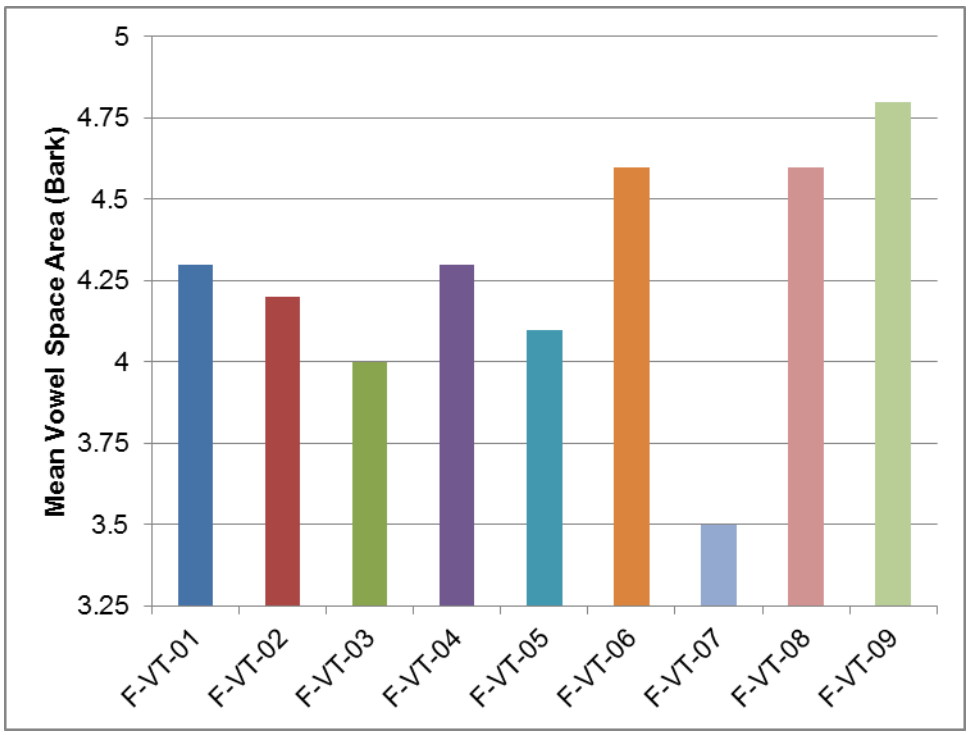

Figure 4-12. Mean articulatory effort (measured in vowel space area) for nine female voice types

Table 4-4. Recoding of the acoustic analysis of the female voice types into extreme values

Voice Type	Size (Count)	Speaking Rate	Voice Quality	Mean Pitch	Pitch Variability	Articulatory Effort
F-VT-01	17					
F-VT-02	14				M	
F-VT-03	6			L	M	
F-VT-04	3		N			
F-VT-05	3		N	L		
F-VT-06	2		C	L		
F-VT-07	2	F			D	R
F-VT-08	2	F		H		
F-VT-09	1	S	C			E

*See text code for definitions

Table 4–4 summarizes the five categories of acoustic cues to categorize female voices into nine voice types. Recall that the female voice types were more starkly divided than the male voice types in terms of perceived and actual vocal age. Four voice types were exclusively composed of young voices: F-VT-01 (the largest), F-VT-07, F-VT-08, and F-VT-09, while F-VT-03 through F-VT-06 were exclusively middle-aged. Only F-VT-02 was mixed by age category, although it was the second largest of the nine and was predominately middle-aged. Effectively then, four young female voice types and four to five middle-aged female voice types were observed. Comparison between voice types, therefore, were made primarily within age category.

Beginning with the young female voice types, the largest in size was also entirely typical in the acoustic cues measured in the study and can be henceforth termed “Young Female Modal.” Four other voice types were partially or wholly consisting of young female voices, and four of the five speech cues were needed to acoustically

differentiate them. The young voices of F-VT-02 were the only ones speaking in a more monotone voice; dynamic did appear in F-VT-07 (Young Female Excited), but it was not necessarily to match voices to that type. Likewise, voice quality as measured by CPP-S played a very limited role in young female voices, and was not needed to characterize the one female voice populating F-VT-09 (Young Female Slow and Clear). Articulatory Effort and Speaking Rate appeared to trade-off and distinguish F-VT-07 and F-VT-09. Finally, mean pitch was necessary to group voices in the Young Female High voice type shown as F-VT-08.

In contrast, the middle-aged female voice types relied more heavily on voice quality, and did not require either speaking rate or articulatory effort. Middle-aged voices constituted 11 of the 14 voices in F-VT-02 (Middle Female Modal), which makes it the largest middle-aged voice type, followed by F-VT-03 (Middle Female Droner). For these two, and for F-VT-04 (Middle Female Rough) vs. F-VT-05, mean pitch was the critical cue, with F-VT-03 and F-VT-05 displaying lower pitch voices on average. Voice Quality was critical for separating F-VT-05 (Middle Female Low and Rough) from F-VT-06 (Middle Female Low and Clear).

A summary of the prosaic labels for each voice type across genders is provided in Table 4–5. Across all nineteen voice types, the five acoustic cues were successful in differentiating voice types, although clearly mean pitch and pitch variability were relied on heavily by listeners. Nevertheless, they were sensitive to other systematic differences between the voices, and all five classes of acoustic cues were carried over in a Linear Discriminant Analysis (LDA) of the largest three voice types within each gender. LDA was not deemed an appropriate technique for estimating the relative

importance of the five acoustic cues with the smaller voice types, given that the factors submitted outnumbered the membership sizes of M-VT-04 through M-VT-10 and F-VT-04 through F-VT-09.

Table 4-5. The membership size and labels for all nineteen voice types, organized also by age group when appropriate

Voice Type	Size (Count)	Age Group	Label
M-VT-01	13		Male Modal
M-VT-02	12		Male High and Clear
M-VT-03	5		Male Monotone
M-VT-04	4		Male “Preacher”
M-VT-06	3	Young	Male “Mumbler”
M-VT-05	4	Middle	Male Low
M-VT-07	2	Middle	Male “Droner”
M-VT-08	2	Middle	Male “Explainer”
M-VT-09	1	Middle	
M-VT-10	1	Middle	
F-VT-01	17	Young	Young Female Modal
F-VT-07	2	Young	Young Female “Excited”
F-VT-08	2	Young	Young Female High
F-VT-09	1	Young	Young Female Slow and Clear
F-VT-02	14	Middle	Middle Female Modal
F-VT-03	6	Middle	Middle Female “Droner”
F-VT-04	3	Middle	Middle Female Rough
F-VT-05	3	Middle	Middle Female Low and Rough
F-VT-06	2	Middle	Middle Female Low and Clear

Two sets of results are provided from the LDAs of the male and female voice types: 1) tables of the successful classification of the relevant voices into the three voice types, to examine the necessity to expand the acoustic analysis in future studies and 2)

rank orders of the five acoustic cues, to denote their relative importance in modeling. Tables 4–6 and 4–7 list the classification of the male and female voice types, respectively

Table 4-6. Percent correct classification of 30 male voices into three common male voice types

	Classified		
Actual	M-VT-01	M-VT-02	M-VT-03
M-VT-01	76.9%	7.7%	15.4%
M-VT-02	33.3%	58.3%	8.3%
M-VT-03	0.0%	20.0%	80.0%

Table 4-7. Percent correct classification of 37 female voices into three common female voice types

	Classified		
Actual	F-VT-01	F-VT-02	F-VT-03
F-VT-01	94.1%	5.9%	0.0%
F-VT-02	21.4%	64.3%	14.3%
F-VT-03	0.0%	0.0%	100.0%

In both analyses, two discriminant functions were obtained, and both were significant in both analyses ($p<0.05$). Across both male and female voice types, classification performance was well above chance for all three voice types, although F-VT-02 and all three male voice types were classified below a high threshold of success (e.g., 90% correct). While the five acoustic cues examined accounted for much of the variance between the major voice types, a richer analysis will be required in future studies in order to successfully classify most voices.

Of greater interest was the relative importance of the five cues in classifying voices into types. A rank order was obtained via an examination of the structure matrix of the LDAs. For male voices, the five cues could be ranked in descending importance as follows:

mean pitch > pitch variability> voice quality > speaking rate > articulatory effort

The rank order obtained agrees with the observations made that inspired the classification system in Table 4–3, with the exception of the higher ranking of speaking rate over articulatory effort, although both cues were only important for distinguishing the smallest voice types. For female voices, the same rank order was obtained.

To further explore and define the optimal set of acoustic-perceptual dimensions that define the voice type inventories derived from the HCS analysis, the same similarity matrix was also submitted to ALSCAL Multidimensional Scaling, with two to four dimensions explored for, separately, male and female voices and young and middle aged voices. The decision to divide the entire database into four groups was made on the basis of the bifurcation of the female voices by age group as well as the age effects observed in the male voices.

The fits obtained between the original similarity values and the derived distances of the MDS analyses varied by group, but were all significant and strong. These proportions of variances are shown in Figure 4-13 for all four groups.

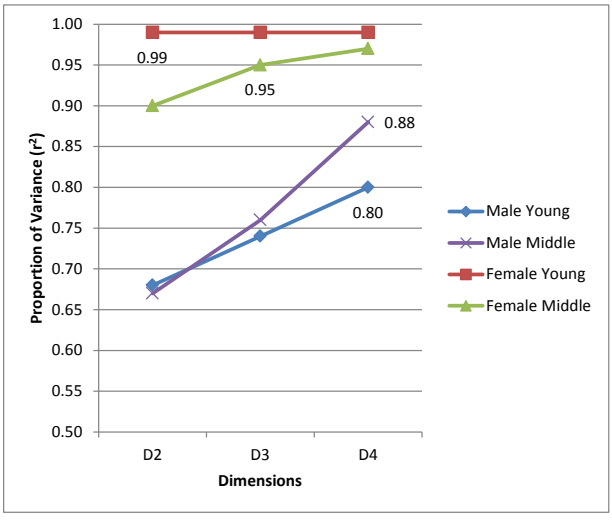

Figure 4-13. Fits between MDS spaces and similarity matrices for Males (Young vs. Middle-Aged) and Females (Young vs. Middle-Aged)

In Figure 4-13, r^2 values are shown for the optimal number of dimensions picked for each group; this number corresponded to the largest number of dimensions that still permitted a desired gain in fit. In the cases of Male voices, no “elbow” was observed in the dataset, but a greater number of dimensions was not feasible given the size of the stimulus set. Female similarity spaces were adequately represented by two and three dimensional solutions.

To characterize each of these dimensions, the acoustic measures used in the discriminant analysis (see Tables 4-6 and 4-7) were correlated with the derived distances of the MDS. Significant correlations (Pearson’s r) were observed between seven of the total of thirteen dimensions examined (across all four groups). To identify the remaining dimensions, an expanded acoustic analysis was performed in which measures of intensity variability, spectrum (F1 – F3 of selected target vowels), voice quality (NVB – number of voice breaks), and fundamental frequency (range) were incorporated based on earlier acoustic analyses of the same database (Harnsberger, Shrivastav, and Brown, 2010). With this expanded range of cues, preliminary definitions can be given to each definition; these are provided in Table 4-8. The strength of the correlations varied widely, between a 0.31 – 0.70, and while were significant, these lower values indicate that the correlates of voice types, beyond speaker pitch and some measure of voice quality, likely combine some of these measures and/or incorporate aspects of the signal not represented in the current analysis.

Table 4-8. Acoustic definitions for 2 – 4 dimensions of the MDS solutions for young and middle-aged male and female voice types.

Voice Type Group	Dimension	Cue	r
Middle -Aged Female	d2	Voice Quality	0.70*
	d1	Mean Pitch	0.61*
	d3	Voice Quality - NVB	0.51*
Young Female	d1	Speaking Rate	0.48*
	d2	Pitch Range	0.46*
Middle-Aged Male	d1	Pitch Variability	0.48*
	d3	Mean Pitch	0.47*
	d4	Pitch Range	0.37*
	d2	Vocal Tract Length (F2)	0.33*
Young Male	d4	Mean Pitch	0.42*
	d2	Pitch Variability	0.39*
	d1	Intensity Variability	0.36*
	d3	Pitch Range	0.31*

CHAPTER 5
DISCUSSION

General

The typology resulting from the present research was developed using untrained listener similarity judgments of voice pairs and several modeling techniques. As predicted, both male and female voices were not judged as uniformly dissimilar to one another, but formed natural groups whose acoustic-auditory properties were further investigated in systematic auditory judgments which guided an acoustic analysis. While the existence of voice types was strongly suggested by earlier literature on speaker identification (and its error patterns), there was no way of knowing *a priori* how many voice types would be discovered, whether these varied in number and kind by gender, and whether or not other indexical properties of speech might influence listener judgments of vocal similarity. This study is the first attempt to address these questions. The answers were many, and complicated.

First, for both genders, 2 – 3 voice types predominated, with all other types each representing small pluralities of the database. Second, male and female voices did prove to differ in the number of salient voice types, their membership size, and the influence of other indexical properties on these groups. Male voice types outnumbered female voice types, and chronological age strongly conditioned the grouping of female voices into types to a greater extent than male voices. For female voices, chronological age influenced distribution of voices into types, with younger voices grouped separately from middle aged voices, while male voice types were more heterogeneous with respect to age. Vocal age appears more salient in judging female voice type, and this observation is congruent with higher accuracy rates in estimating female vocal age in

prior studies (Harnsberger, et al., 2010; Schotz, 2005). Age is, however, not irrelevant to the male voice typing in this study. In the two largest groups of male speakers, M-VT-01 and M-VT-02, 70% and 75% respectively are skewed towards one of the two age groups. M-VT-01 is heavily populated with “young” speakers, while M-VT-02 is heavily populated by “middle-aged” speakers. The population is even more dominated by age in subsequent groups. Of course, membership population is progressively smaller in the typology for males, from M-VT-01 to M-VT-10, with relatively small memberships throughout, due to limitations in the sample size, so it is somewhat difficult to gauge how heavily age is playing a factor in listener judgment of similarity.

In the case of the females, there was little doubt that the typology was heavily influenced by age. It played a factor even in the selection of the final number of groups. In the final typology for the females, 100% of the first group was “young,” nearly 80% of the second group was “middle-aged,” and all subsequent groups in the typology were dominated completely by one age group or another. Again, because of the size of the membership of most of the groups, it is difficult to say to what degree exactly age was playing a factor in similarity judgment, but it is clear that it played a strong role for females (stronger than with the males).

This study eliminated “old” voices at the onset, as they were believed to be too aberrant, thus resulting in types that would be clustered by age alone. Not only did this assumption prove to be correct, the results of the study and the apparent influence of age on listener judgments indicate that in a further study, it might well be better to posit groups by age category. Separating the sample at the onset further by age would likely yield voice types that are more heavily composed of true identity properties of the voice,

with less age influence on the typology. A superior modeling approach in the future might be to capture all possible voice types among exclusively young, healthy voices and derive a vocal typing model for middle-aged and old voices by combining a standard vocal aging acoustic model to young types, rather than attempting to directly assess vocal typology in older generations. The best place to begin further research toward positing voice types by age category would be with young voices, because understanding the young, healthy model first is more important than seeing how voice types interact with age variability, though that would also be fertile ground for future study. The older a voice, the more likely it is to be affected by the aging process. This might be most apparent in “voice quality” as an acoustic measure, but pitch can also be affected by the aging process by changes in the shape and elasticity of the vocal folds.

It may be that all older voices are derivable from an inventory of voice types based on young voices, although rapid sound change within languages over successive generations could undermine that approach. The oldest voices, for instance, might be better typed by applying aging effects to a vocal taxonomy based on noncontemporary young speech. Essentially, the older and middle aged voice types were perceive today maybe derivable from past young voices that do not display articulatory, vocal, or prosodic characteristics utilized by younger speakers today.

Voice types are shown in this study to be real and to exist, with concrete acoustical correlation for utility in classifying new voices with an established vocal typology for American English and beyond, and are therefore independent of age. However, age effects so many aspects of the human voice, filtering the sample by age may be a necessary step in future study for establishing a workable typology that

accounts for age. Starting with young voices would be a good first step towards accounting for the effects of aging on older, even middle-aged voices.

While this dissertation has demonstrated that listeners are capable of grouping voices by similarity, the psychological reality of the types would be best supported by repeated experiments of this type, and more interestingly, studies of the any effects of voice types on the perception and learning of linguistic and/or indexical properties of speech. Expert listeners have identified a number of readily audible and discernable characteristics of the natural class types. These characteristics were later determined to coincide without conflict to those correlate measures found in the acoustic analysis for each type, making it clear that the established types do have an audible reality beyond machine measurable acoustic factors. The bottom-up experimental techniques used to build the taxonomy and yield the vocal stereotypes, i.e., untrained listeners' auditory judgments on perceived similarity and dissimilarity of voices in the database, have also demonstrated this. However, to what degree listeners perceive individual speakers as members of a group of similar speakers is unclear and requires further study, reflection, experimentation, and research. Whether such a reality exists or not in the mind of the listener might also go a long way towards furthering our current understanding of vocal pleasantness and sociolinguistic profiling. How we perceive a person based on their voice type might well classify them as a person with certain social, intellectual, and personality traits, creating a subtle profile in our minds that may affect how one individual treats and regards another in social and professional conversational situations throughout life.

What remains an open question is to what degree vocal similarity is used in the listener's mind to form rough voice types, which are in turn judged by social and vocal pleasantness criteria by many listeners to form stereotypes related to social status, personality traits, and so on. That social stereotyping at this level takes place is known and has been shown in several studies, though most of them involve dialect as a determining factor in assigning social stereotype by voice (Baugh, 2000; Smalls, 2004). In a notable study on linguistic profiling, Stanford researcher, Dr. John Baugh, demonstrates how speakers of African American English and Hispanic English are discriminated against when they apply for housing (Baugh, 2003). This study was performed by phone, with real-estate agents routinely denying or reducing the selection of available apartments when confronted with dialects associated with particular racial identities.

How vocal similarity vis a vis the indexical property of speaker identity in voice may play a role in this process and in SPID by untrained listeners is a question for the psychological reality of voice types. The utility of a working vocal typology for explicit rather than implicit speaker SPID by expert practitioners and academics is apparent as an application of this research, but the degree to which this typology may be already being applied by listeners as part of their subconscious SPID procedures remains unclear. Future studies should directly assess the efficacy of voice typing in forensic SPID, specifically in studies with a large sample of voices that systematically match or mismatch in type. The stimulus materials in question would need to emulate common characteristics of evidence recordings, particularly with respect to background noise the

frequency range constraints imposed by typical covert recording equipment and conditions.

The role of the listener in speech often receives less attention in modern linguistic departments, in favor of theoretical and psychological modeling of speakers' production processes, the study of which dominates much of the core fields of present day American linguistics departments and related associations of linguistic inquiry. Crucial to this thesis' results is the role that the listener plays in forming the natural class categories that drive a practical and useful taxonomy of an American voice database. (Graumann & Herrmann, 1989). The reality of voice types and speaker identity in the mind of the listener and in the speech signal as an indexical property on par with gender, age, and other apparent properties of the speech signal may be provable through evidence of its effects on word recognition and other linguistics processing. Talker effects in word recognition have been demonstrated in other studies (Johnson & Mullennix, 1997). Competing theories on normalization or episodic coherence in listener processing of the speech sample to account for indexical property "filtering" of the speech signal would in either case also have to be in play for accounting for talker identity and voice type in the signal, if in fact the voice type is a real property in the signal.

Speech Characteristic Coding of Voice Types

Coding the speech cues measured in this study was completed to provide a simple rubric to characterize each male and female voice type. Coding involved matching of acoustic measurements of the voices in the database with the population of each voice type for each sex. Characteristics of a given type were assigned a speech code when the mean value for that type was one standard deviation above or below the

mean across all voice types, i.e., “Fast” or “Slow” for Speaking Rate, “High” or “Low” for Mean Pitch, “Monotone” or “Dynamic” for Pitch Variability, and so on. This process revealed a possible outcome of 3 correlates for each speech characteristic scale. For example, in the case of Speaking Rate, the potential correlates could be “Fast” for 1 standard deviation above the mean; “Slow” for one standard deviation below the mean speaking rate, or [blank] for average (no correlation). This resulted in under-specified groups having the least correlate labeling and, predictably, being the most populous groups. This under-specification and heavier member population is reflected in the group ordering and labeling for the types within each sex, with type “01” for each sex being the group with the largest membership and also under-specified. As the types move higher in label number they become smaller in membership number and more specified by characteristic coding. This was true of both sexes. In cases where the membership number might be the same for two types, the less specified group by coding is giving the lower number label (higher slot) in the typology, within its given sex.

Characteristic Dominance in the Typology

Speech characteristics dominate the typology established in this study by both sex and size of the group. The types are separate by sex and also ordered by group size, which corresponds generally to under-specificity of acoustic correlate and speech characteristic from largest to smallest group by membership. In the case of the males, 60% of the voices were able to be classified in their coding by only 3 of the speech characteristics—voice quality, mean pitch, and pitch variability. Females similarly can be classified at the higher levels by the same three dominate characteristics, with the most populous types being highly under-specified. While such a limited number of codes were sufficient to characterize the largest voice types, the subsequent use of all five

acoustic correlates in linear discriminant analysis suggested that the full inventory of possible cues has not been exhausted. The acoustic analysis of this study was guided, first by prior literature on the acoustic correlates of speaker identity and then by a systematic auditory analysis for the particular stimulus materials constituting the database for this experiment. The goal was not evaluate any and all acoustic cues that have been posited for speaker identity or other indexical properties, but only those map back easily on well-understood attributes of phonation and articulation. However, the limited success in LDA modeling, additional cues should be incorporated, such as mid to high frequency spectral information, with the goal of linking such cues to vocal tract characteristics of individuals (e.g., F4 and lip protrusion; Pennington, 2011).

Applications

Application and Value of Research

Are you a Clooney or a Gottfried? A Diva or a Droner? While this dissertation is a reporting on the experimentation into speaking voice types and an analysis of those results with their implications for our current understanding of natural vocal typology, it is desirable that this new typology of American speaking voices might find opportunity for future study, but most especially, for practical application outside of academia. For this reason, these voice types and the data itself should provide an initial model for vocal typology that will find potential revision through future experimentation and eventual utility in the public, academic, and private sectors. What follows is a partial listing of possible applications for a vocal typology across an array of fields.

Academic and Scientific Study

Upon completion and publication of this study, other phoneticians, acoustic experts, and linguists might seek to perfect or recast these types, leading to a more

universal typing system, to be of academic and scientific utility to a broad range of different fields. Geneticists might seek to explore whether genetically related people share common voice types separate of environmental factors, such as dialect. Government bodies might seek to use this system to determine which voice types are most pleasant or well-received for different types of public service announcements, warning systems, or automated verbally-administered testing procedures. Cross-linguistic as well as unified voice types might be used to determine regional typology trends in acoustic signals of voices, to complement existing linguistic typology research on the world's languages.

Forensic

The science of SPID and its different methods have long been plagued by issues of degree of accuracy. Both in and out of the courtroom, forensic acousticians and phoneticians have had their work fall under criticism for the degree to which they are able to make a positive match or to rule out a voice. One of the issues with acoustic as well as perceptual-aural analyses have been their estimated degree of accuracy by percentage. While the usefulness of traditional SPID is apparent, it’s admissibility in a court of law, where lives often hang in the balance, has been under scrutiny for years. In light of the high degree of accuracy in identification methods such as DNA matching in recent years, it is not surprising that judges and the judicial system might look unfavorably on scientific identification methods that yield a shaky 60% match or often even less. Adding to the skepticism of the judicial system, is the “CSI Effect” exhibited by juries in recent years.

Though blood typing in the absence of DNA evidence may produce only evidence that the defendant and the perpetrator of a crime share a blood type belonging to

hundreds of millions of others in the world, it is admitted freely into the courts, as in the case of blood typing, a match is a match. So, it would seem that for the courts, and often the public at large, a 100% match to a type belonging to the defendant and millions of others would seem to be preferable to a lesser certainty, though far more personalized match, as in the case of a SPID match made by a forensic linguistics professional. Additionally, despite ample experimental and academic evidence suggesting the fallibility of voice line-up and other SPID matches by witnesses, both the judicial system and juries have traditionally looked favorably upon such evidentiary presentations in the courts. What seems to be missing is an alternative identification method that might pass muster with judges, juries, and professionals—one that would be neither too stringent to raise accuracy issues, yet not broad enough to be lost in futility. If a reliable vocal typing system were available to the courts and forensic linguists, it could be of great utility.

Government and Public Interest

Of course, voice typing would yield not only a possible greater application in the forensic/judicial realm, but would also have the advantage of opening up the field to a new and wide range of civilian, corporate, and government applications—where its application would be less controversial. The narrow criminal application outlined above is only one of thousands of possible applications, including everything from marketing and advertising to warning systems and public service, safety, and welfare announcements. In addition, vocal types, which naturally include celebrity voices and are accessible to the public would have commercial viability and could bring much needed resources and attention to the field of forensic phonetics and linguistics generally. Perception of ones voice by others, and the implications involved therein,

would be valuable information to individuals, corporations, and political entities alike. Online and elsewhere, services that offer the visitor the opportunity of having their voice typed would be of great public interest. Individuals could learn what the characteristics of their voice type are, as well as what famous voices they resemble typologically. In addition, individuals could experiment to see if their voice belongs to the same type as their loved ones and others, opening the door to interesting speculation on how couples and friends might be drawn together in part by voice, and whether or not voice is passed down genetically and/or by proximity and nurturing of ones children.

Business and Marketing

Business and Marketing applications abound for a usable voice type system. Determining which voice type would be most effective for marketing of specific products and services (trustworthy for insurance and financial products, exciting or relaxing for travel services, etc.) through marketing research, would in the end be of shared interest to academics looking to determine the personality qualities associated with certain indexical properties of voice co-assigned to each type. This would also be useful to voice casting agents to provide their clients with short lists of available voice actors filtered by requested type or quality. Voice type could become an indispensable element of an actor's resume. How this research might be of value to singing voices and jingles is also an open and intriguing question that would require further research. Marketing researchers interested in creating personality-reflective voice types for narrative marketing services to specific industries could provide valuable research partners for vocal typologists interested in vocal pleasantness and voice perception in sociologically-geared experiments on telephone discrimination and other areas of inquiry.

Sociolinguistics and Discourse

How ones voice is perceived in discourse and daily human interaction has many implications for how a person is treated in society. The study of how certain voice types are perceived via their identity indexical property in the speech signal and how this perception affects their overall treatment in discourse would be of interest to sociolinguists and discourse analysts. Issues of attraction, repulsion and other emotional responses based on the physical properties of the speech signal would be an important overlap to the study of vocal pleasantness and discourse analysis within and between vocal types.

Future Study

Upon completion and publication of this study, other phoneticians, acoustic experts, and linguists might seek to perfect or recast these types, leading to a more universal typing system, to be of academic and scientific utility to a broad range of different fields. Geneticists might seek to explore whether genetically related people share common voice types separate of environmental factors, such as dialect. In the current study, one set of identical twins was observed superficially by two expert listeners for voice similarity. This observation was not acoustically measured, but consisted instead of a short interview only. Generally, the twins were observed to have a remarkable similar voice type, with their voices often being mistaken over the telephone and in other environments. These interviews were not scientific enough to yield data, but it did open the door to a wider study of people who share similar body types or genetically identical vocal tract constructions to be studied for vocal similarity.

Government bodies might seek to use the current typology system to determine which voice types are most pleasant or well-received and complied with for different

types of public service announcements, warning systems, or automated verbally-administered testing procedures. Cross-linguistic as well as unified voice types might be used to determine regional typology trends in acoustic signals of voices, to complement existing linguistic typology research on the world's languages. It is the intention of the current author to complete a possible follow-up study on native speakers of Japanese. A study of this type would help determine 1) to what degree voice types are universal and cross-linguistic in nature and 2) how valid the results of this study are when replicated with a different population of speakers and/or listeners generally.

CHAPTER 6
CONCLUSIONS

Results Summary

That individual voices are not uniformly similar to one another can be seen as intuitive. The untrained listener understands on a basic level that some voices are more or less similar to others. The current research has scientifically confirmed what laypeople, human listeners at large, have unconsciously recognized and known to be the case—that a given listener perceives one voice to belong to a similar type of voice to another, much as the human mind can classify other elements useful to identifying a known or unknown individual from previous encounters and experience. Physical appearance, facial recognition, even syntactic patterns in letter writing, emails, texts, or instant messaging can be used by the mind of a listener/reader/viewer to recognize and identify an unknown individual as being the same or similar to a previously encountered or known personality. In some cases, the perceiver of a given mode of communicative stream might misidentify a communication as being from an individual from whom it did not in fact originate. This can happen when the communication is an intentional imitation of a individual known to the perceiver or when the communication is unintentionally and coincidentally similar to that of an individual known to the perceiver. That voice types follow a similar pattern is by no means surprising. We have probably all had the experience of answering the telephone or turning our heads to face someone who is verbally addressing us, sure in an instant of whose voice it is that we just heard, only to find that we have misidentified an incoming voice incorrectly. This is because some voices are more similar, in some cases very similar, to some voices than others. While dialect, age, and voice quality can be factors that influence our decisions as listeners to

identify some similar as more similar to each other than others, it is clear that anatomical, fixed factors of the human voice also affect these similarity and identification judgments.

The present research and experimentation has yielded a working set of voice types for American English, based on the above described type of similarity judgments by untrained listeners. This research has tapped directly into the human's intuitive understanding of voices (in this case) inherent non-uniform similarity to one another. Data reduction and acoustic analysis were then used to cull, sculpt, confirm, and label these natural class voice types into a working typology of American voice types separately for males and females. Use of a larger database or larger sample of the listener population might result in nominal changes to this typology, without changing the fundamental aims and results of the project—which were to build a natural class themed taxonomy of vocal stereotypes that is based on the intuitive judgments of the untrained human ear. The untrained listener has been shown by Hollien (2002) and others to remain as the most reliable mechanism for human voice identification. Likewise, fully automated systems of SPID remain a wholly unrealized dream of SPID researchers and practitioners. Semi-automatic methods of forensic SPID have remained the most effective method, as they tap into the continuing supremacy of the human ear and the technological aides that modern advancements have brought to the field. This dissertation has adhered to this view on the current state of SPID by using untrained listener judgments (the power of the human ear) as a basis for coherence of voices from the database into voice types, while also using acoustic, technological measures, data reduction techniques, and expert listener judgments to confirm and further clarify the

resulting typology. This approach, flowing from the bottom-up approach to data collection of Experiment 1, with acoustic confirmation from machine analysis in Experiment 2, was further validated by the expert listener data from the ‘expert listener aural perceptual judgments’ phase—where nothing was found in the final analysis to have been contradicted from the acoustic analysis by the expert listener aural-perceptual data. That is to say, nothing was yielded from the acoustic analysis of the speaker stereotype correlates that was not confirmed by the expert listener judgments. Some additional information was provided by the expert listener data that was not found in the acoustic analysis at a significantly correlated level, but there were no contradictions between the perceptual and acoustic data. Further study in the same spirit of untrained listener base, layered with trained listener and acoustic measurement enhancement on larger and also crosslinguistic databases would further the present study and bring additional clarity to the typology presented for consideration here.

Furthermore, interdisciplinary studies and correlations with sociolinguists, psychologists, geneticists, voice acting and identification professionals, software engineers, and vocal pleasantness researchers would yield further insight, as well as a wealth of applications for the current and future studies in vocal typology.

A Note on Online Accessibility

In the case of graduate-level theses research and in academic research generally, there can be little excuse these days not to offer one’s data, methodology, literature review, and evolving conclusions online for general or limited consumption and feedback.

The idea of a "3D Dissertation," where fellow academics can easily access and offer advice on a student's or colleague’s work, and where the visitor can learn more

about specific areas by clicking deeper or not into the content, is highly appealing to the contemporary graduate student. The ability to learn tangentially by checking definitions of key terms, authoring notes and details, resources, and data links is of high value to both the author and the reader in a work designed for academic consumption, where it is always difficult for the author(s) and their advisers to determine to what degree assumptions should be explained or justified within the text of a traditional linear paper thesis. Tangential reading, where the reader can delve deeper into certain topics by clicking deeper into certain areas, while skimming or ignoring areas where the reader is already familiar with the material or disinterested in the details of the sub-heading or topic, for better or worse, has become the modern reading style of the internet age. This method also allows the author to effectively share and elicit feedback on their research from colleagues in a potential wide-variety of disconnected fields. Depth of knowledge is neither assumed in a particular specialty related to the research top, nor belabored, thus facilitating more effective interdisciplinary collaboration. Furthermore, conference presentations are enhanced beyond subsequent proceedings and journal publications, by supplemental, online access to the presentation material, which can be accessed by those who did not attend the conference, but learned of the topic, those who attended, but needed further, subsequent time to review the material, and distance colleagues, who are separated from potentially significant input and collaboration by geography and language communication barriers.

While electronic accessibility to traditionally written articles and theses has been a great leap forward for academic research, versions of theses written and designed specifically for online consumption can take online accessibility a good deal further in its

evolution. Of course, this author is under no illusions—a thesis available entirely electronically, without a traditional paper complement (which meets all the current graduate school requirements) would not be institutionally acceptable quite yet (in fact, such an online version would probably be ignored from a certification perspective).. But can the day when this will be the norm, with the linear, two dimensional thesis all but ignored, be far off? For these reasons, the rough strokes of this research, as well as its results, aims, and applications have been kept online throughout the drafting and research process to elicit feedback from academics and field practitioners where possible. Into the future, the information will be maintained, with forum and author contact access online at the www.voicetypes.com website. Parts of this research project have already been accepted for presentation and publication in at least two highly visible venues and further publications of this research and continuing research offshoots are in the planning and submission phase. As the research project and its publishing and collaborative branches grow, contributing to a more comprehensive typology of human voices, these publications, collaborations, and further research will be chronicled and outlined online at the existing website. This website will come to include comments from collaborating scholars as well as references and introductions to others doing work that enhances, predicates, or is resultant from this research project. This reflects the reality that vocal typology is both novel and at the same time, a natural evolution from the work of previous researchers; it fits nicely into work in other fields and practical application and also goes a long way towards establishing a new field of inquiry into the classification of human voices generally.

Eventually, after considerable further research and development, the website aspires towards an online, interactive voice typing tool, where visitors can type their own voices, and learn about the aural-perceptual and acoustic correlates, sociolinguistic biases, and perceived social traits that correspond to and are associated with their personal voice type. Other, non-academic correlates might also be of great popular interest to the public, such as career implications by voice type, as revealed by vocal pleasantness studies, as well as a listing of celebrities who share a visitor's determined voice type. In association with software engineers, another goal would be to make available online typing tools and guidelines to forensic practitioners, in an effort to assist with voice identification and VTID training.

APPENDIX A
FEMALE PILOT HCS DENDROGRAM

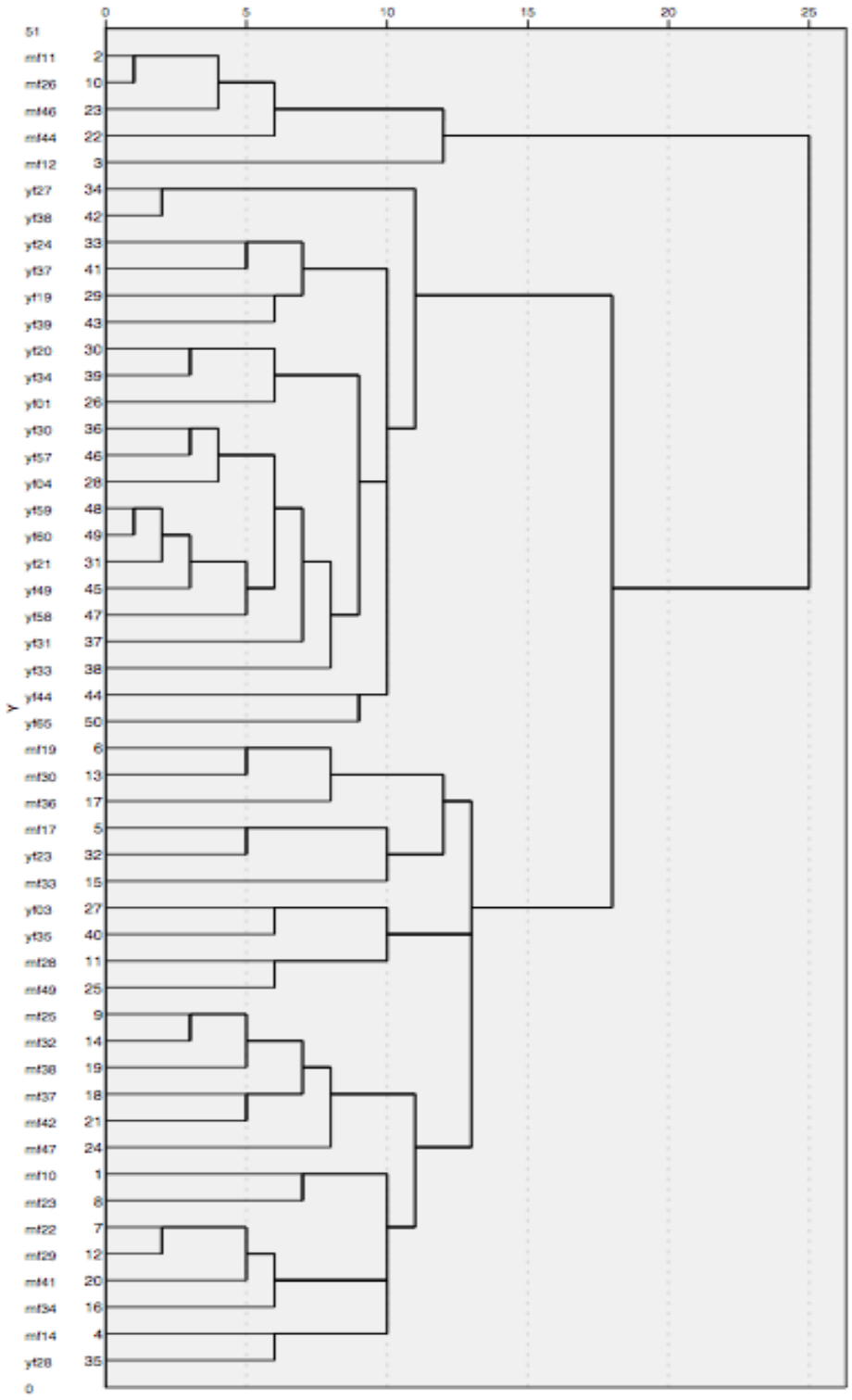

Figure A-1. Female pilot HCS dendrogram with individual young male (ym) and middle-aged male (mm) voices shown on the vertical axis.

APPENDIX B
MALE PILOT HCS DENDROGRAM

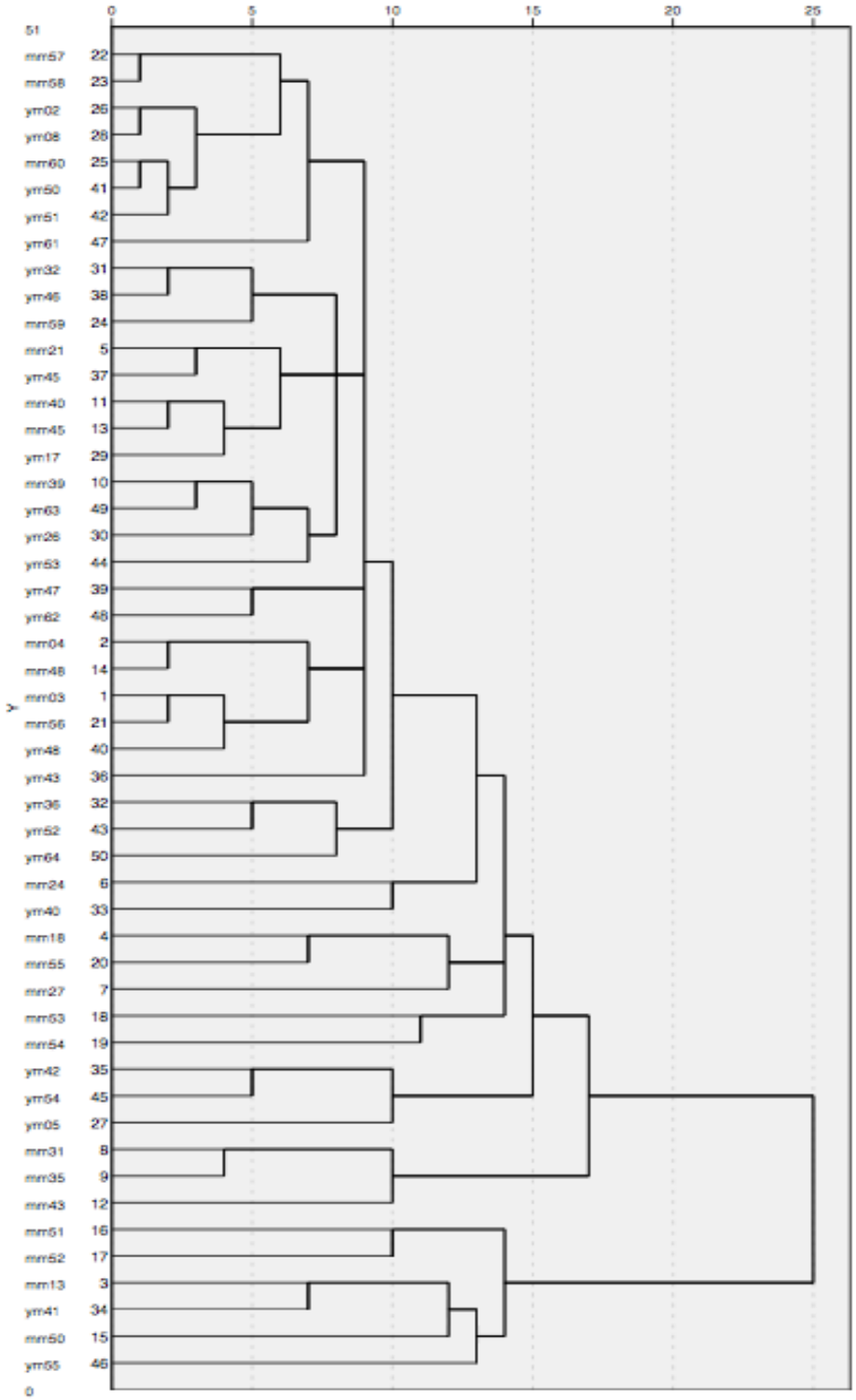

Figure B-1. Male pilot HCS dendrogram, with individual young female (yf) and middle-aged female (mf) voices shown on the vertical axis.

APPENDIX C
FEMALE HCS DENDROGRAM

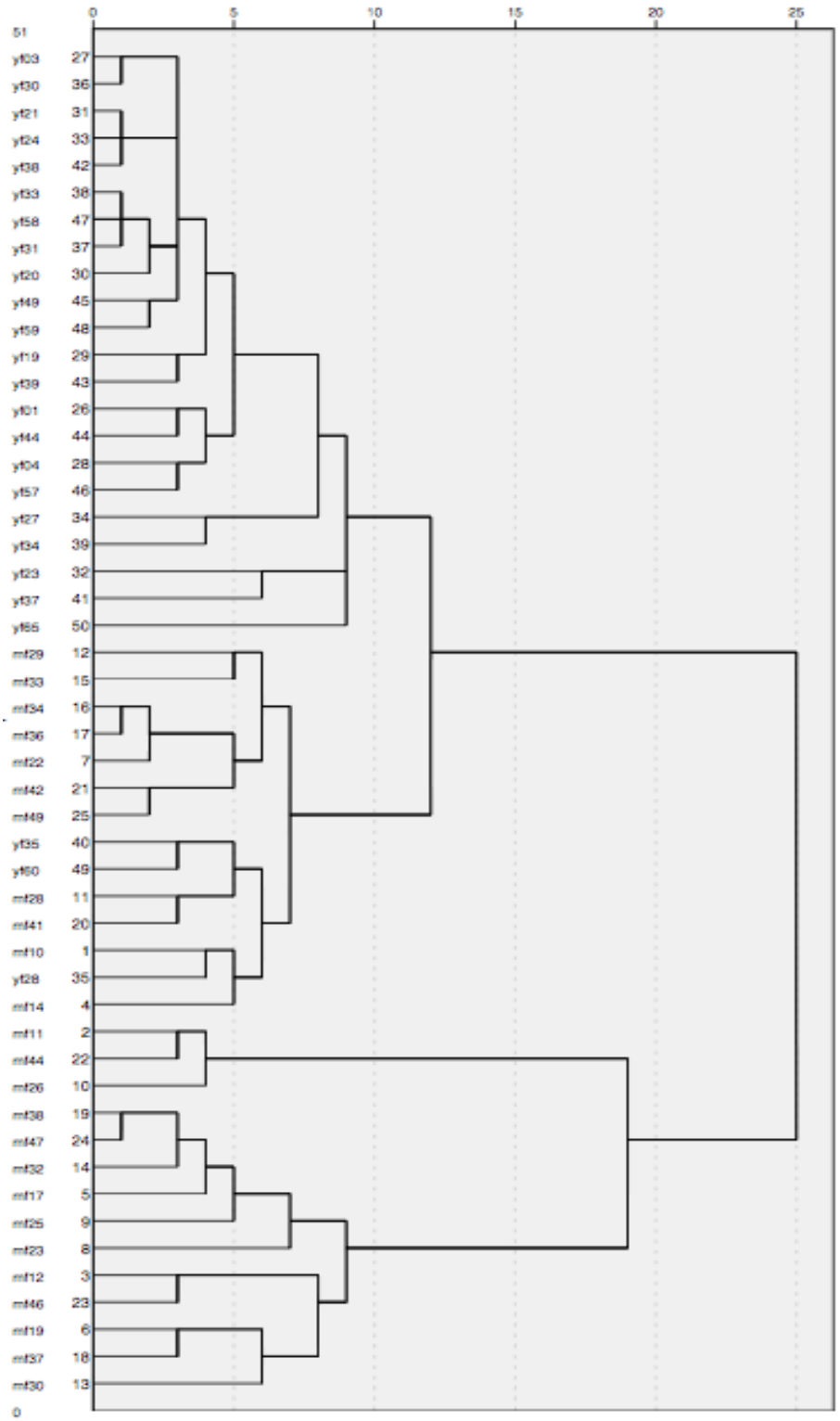

Figure C-1. Female HCS dendrogram, with individual young female (yf) and middle-aged female (mf) voices shown on the vertical axis.

APPENDIX D
MALE HCS DENDROGRAM

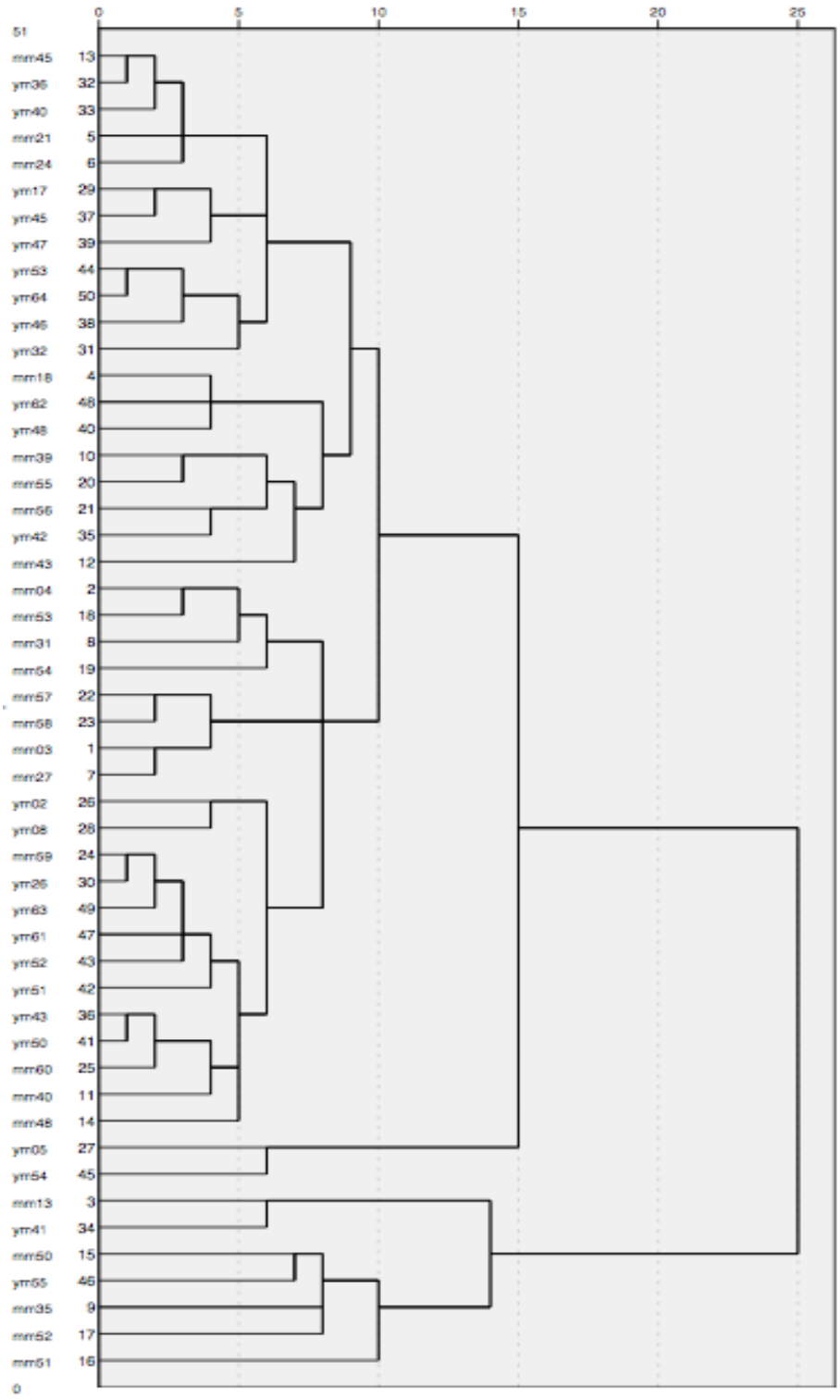

Figure D-1. Male HCS dendrogram, with individual young male (ym) and middle-aged male (mm) voices shown on the vertical axis

APPENDIX E
TEXT OF 10 SPIN SENTENCES INCLUDED IN THE DATABASE

1. “His boss made him work like a slave.”
2. “He caught the fish in his net.”
3. “The beer drinkers raised their mugs.”
4. “I made the phone call from a booth.”
5. “The cut on his knee formed a scab.”
6. “I gave her a kiss and a hug.”
7. “The soup was served in a bowl.”
8. “The cookies were kept in a jar.”
9. “The baby slept in his crib.”
10. “The cop were a bulletproof vest.”

*11. “How long can you hold your breath?”

*12. “At breakfast he drank some juice.”

*13. “I ate a piece of chocolate fudge.”

*14. “The judge is sitting on the bench.”

*15. “The boat sailed along the coast.”

*16. “The pirates buried the treasure.”

*unused sentences from database

APPENDIX F
EXPERIMENT QUESTIONNAIRE

Subject Code: _____

SUBJECT INFORMATION FORM

Age (in years)_____Place of Birth_____

Sex: M F

1) First language spoken if not English (or in addition to English):_____

2) List any other languages you speak, along with your proficiency:_____
_____
_____

3) Have you ever had a hearing or speech disorder? Yes No
If "Yes," please explain:_____

4) Please list all of the places you have lived and the years when you lived there in chronological order (include only places in which you spent at least a year in residence):

City, State, Country Years

REFERENCES

Abberton, E. & Fourcin, A. J. (1978). Intonation and speaker identification. *Language and Speech, 21*, 305-315.

Alderman, T. G. (2005). Forensic speaker identification: A likelihood ratio-based approach using vowel formants. *München: LINCOM Gmbh (LINCOM Studies in Phonetics, 1)*. 2nd printing, 2007.

Alexander & A. Drygajlo (2004). Scoring and direct methods for the interpretation of evidence in forensic speaker recognition. *Interspeech, Jeju, Korea, October*. 2397-2400.

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time, *Journal of the Acoustical Society of America, 113*. 544-552.

Amino, K. & Arai, T. (2009). Effects of linguistic contents on perceptual speaker identification: comparison of familiar and unknown speaker identifications. *Acoustical Science and Technology, 30(2)*, 89-99.

Atal, B. S. (1972). Automatic speaker recognition based on pitch contours. *The Journal of the Acoustical Society of America, 52*, 1678-7697.

Baugh, J. (2000). Racial identification by speech. *American Speech*, 75(4), 362-364.

Baugh, J. (2003). Linguistic Profiling. In Makoni, Sinfree (Ed.). *Black linguistics: Language, society, and politics in Africa and the Americas* (155-168). New York: Routledge.

Beebe-Center, J. Gilbert. (1965). *The psychology of pleasantness and unpleasantness.* New York: Russell & Russell.

Bimbot, I., Magrin-Chagnolleau, & Mathan L. (1995). Second-order statistical measures for text-independent speaker identification. *Speech Communication, 17*, 1-2.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of The Institute of Phonetic Sciences. 17*, 97–110.

Bolt, R. H., Cooper, F. S., Green, D., Hamlet, S. L., Hogan, D. L., Mc Knight, J. G., Pikett, J. M., Tosi, O., & Underwood, B. D. (1979). On the theory and practice of voice identification. Technical report, National Academy of Sciences.

Botti, F., Alexander, A., & Drygajlo, A. (2004). An interpretation framework for the evaluation of evidence in forensic automatic speaker recognition with limited suspect data. *Proceedings of 2004: A Speaker Odyssey, Toledo, Spain*, 63-68.

Bricker, P. D. & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *The Journal of the Acoustical Society of America, 40(6)*, 1441-1449.

Broeders, A. P. A. (2001). Forensic Speech and Audio Analysis. *Forensic Linguistics, 1998 to 2001, A Review, in 13th INTERPOL Forensic Sciences Symposium*, 16-19.

Bugental, B. B. & Lin E. K. (1997). Attention-grabbing vocal signals: Impact on information processing and expectations. *Personality and Social Psychology Bulletin*, 23(9), 965-973.

Callaghan, J. (2000). Singing and Voice Science. San Diego, California: Singular Publishing Group.

Campbell, J. P. (1997). Speaker recognition: A tutorial. *Proceedings of the IEEE, 85(9)*, 1437-1462.

Champod, C. & Meuwly, D. (2000). The interference of identity in forensic speaker recognition. *Speech Communication, 31*, 193-203.

Cleveland, T. F. (1993). Voice Pedagogy for the Twenty-First Century: Toward a Theory of Voice Classification (Part 1). *The NATS Journal, 49(3)*, 30-31.

Clifford, B. R., Rathborn, H., & Bull, R. (1981). The effects of delay on voice recognition accuracy*. Law and Human Behavior, 5*, 201–208.

Clifford, B. R. (1980). Voice identification by human listeners: On earwitness reliability. *Law and Human Behavior, 4(4)*, 373-394.

Compton, A. J. (1963). Effects of filtering and vocal duration upon the identification of speakers, aurally. *The Journal of the Acoustical Society of America, 35(11)*, 1748-1752.

Doddington. (1985). Speaker recognition - Identifying people by their voices. *Proc. IEEE, 73(11)*, 1651-1664.

Doherty & Hollien, H. (1978). Multiple-factor speaker identification of normal and distorted speech. *Journal of Phonetics, 6*, 1-8.

Dommelen, W. A. van. (1987). The contribution of speech rhythm and pitch to speaker identification. *Language and Speech, 30(4)*, 325-338.

Dommelen, W. A. van. (1990). Acoustic parameters in human speaker recognition, *Language and Speech, 33(3)*, 259-272.

Dommelen, W. A. van. (2001). Identification of twins by spoken syllables. *Perceptual and Motor Skills, 92*, 8-1.

Doty, N. D. (1998). The influence of nationality on the accuracy of face and voice recognition. *American Journal of Psychology, 111*, 191–214.

Drygajlo, A., Meuwly, D., & Alexander, A. (2004). Statistical methods and Bayesian interpretation of evidence in forensic automatic speaker recognition. *Eurospeech*.

Doherty, E. T. (1976). An evaluation of selected acoustic parameters for use in speaker identification. *Journal of Phonetics, 4*, 321-326.

Endres, W., Bambach, W., & Flosser, G. (1971). Voice spectrograms as a function of age, voice disguise, and voice imitation. *The Journal of the Acoustical Society of America, 6(2)*, 1842-1848.

Erickson, M., Perry, S., & Handel, S. (2001). Discrimination Functions: Can They Be Used to Classify Singing Voices? *Journal of Voice, 5(4)*, 492-502.

Farrús, M. (2009). Fusing prosodic and acoustic information for speaker recognition [Thesis abstract]. *International Journal of Speech, Language and the Law, 16(1)*, 169-171.

Fatokakis, N., Tsopanoglou, A., & Kokkinakis, G. (1993). A text-independent speaker recognition system based on vowel spotting. *Speech Communication, 12(1)*, 57-68.

Furui, S. (1981). Comparison of speaker recognition methods using statistical features and dynamic features. *IEEE Trans. Acoustics, Speech, Signal Processing, 29(3)*, 342-350.

Furui, S. (1986). Research on individuality features of the speech waves and automatic speaker recognition techniques. *Speech Communication, 5(2)*, 183-197.

Furui, S. (1997). Recent advances in speaker recognition, *Pattern Recognition Letters, 18*, 859-872.

Gfroerer. S. (2003). Auditory Instrumental Forensic Speaker Recognition. In Proc. Eurospeech 2003, pages 705-708, Geneva, Switzerland.

Glenn, J. W. & Kleiner, N. (1976). Speaker identification based on nasal phonation. *The Journal of the Acoustical Society of America, 43*, 368-372.

Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory and Cognition, 19(5)*, 448-458.

Goldstein, A. G., Knight, P., et al. (1981). Recognition memory for accented and unaccented voices. *Bulletin of the Psychonomic Society, 17*, 217–220.

Goldstein, U. G. (1976). Speaker-identifying features based on formant tracks. *The Journal of the Acoustical Society of America 59*, 176-182.

Graumann, C. F. & Herrman, T. (Eds.). (1989). *Speakers: The role of the listener*. Clevedon, UK: Multilingual Matters Ltd.

Harnsberger, J. D., Shrivastav, R., & Brown, W. (2010). Modeling perceived vocal age in American English. *Proceedings of Interspeech*, 466-469.

Hillenbrand, J., & Houde, R. A. (1996). Acoustic correlates of breathy vocal Quality: dysphonic voices and continuous speech. *Journal of Speech, Language, and Hearing Research, 39*, 311–321.

Hirson A. & Duckworth, M. (1993). Glottal fry and voice disguise: A case study in forensic phonetics. *J. Biomed. Eng., 15(3)*, 193-200.

Hirson, A. & Duckworth, M. (1995) Forensic implication of vocal creak as voice disguise. BEIPHOL, *Studies in Forensic Phonetics, 64*, 67-76.

Hollien, H. (1990). *The acoustics of crime: The new science of forensic phonetics*. New York, NY: Plenum Press.

Hollien. H. (2002). *Forensic voice identification*. London, England: Academic Press.

Hollien, H. & Harnsberger, J. (2011). Speaker identification: The case for speech vector analysis. *Upcoming*, 1-47.

Hollien, H. & Harnsberger, J. (2012). “Speaker identification using the tricks of the voice-science trade.” Presented at the 41st Annual Symposium: Care of the professional voice, The Voice Foundation.

Hollien, H. & Majewski, W. (1977). Speaker identification by long-term spectra under normal and distorted speech conditions. *The Journal of the Acoustical Society of America, 62*, 975-980.

Hollien, H., Majewski W., & Doherty, E. T. (1982). Perceptual identification of voices under normal, stress and disguise speaking conditions. *Journal of Phonetics, 10*, 139–148.

Hollien, H. & Schwartz, R. (2000). Aural-perceptual speaker identification: Problems with noncontemporary samples. *Forensic Linguistics, 7*, 199–211.

Hollien, H. & Schwartz, R. (2001). Speaker identification utilizing noncontemporary speech. *Journal of Forensic Sciences, 46*, 63–67.

Howard, Hirson, A., French, J. P., & Szymanski, J. E. (1993). A survey of fundamental frequency estimation techniques used in forensic phonetics. *Proceedings of the Institute of Acoustics, 15 (Part 7)*, 207-215.

Iles, M. (1972). Speaker identification as a function of fundamental frequency and resonant frequencies. PhD dissertation, University of Florida.

Ingemann, F. (1968). Identification of the speaker's sex from voiceless fricatives. *The Journal of the Acoustical Society of America, 44*, 1142-1144.

Jassem, W. (1968). Format frequencies as cues to speaker discrimination. *Speech Analysis and Synthesis, 1*, 9-41.

Jassem, W., Steffen-Batog, M., & Czajka, S. (1973). Statistical characteristics of short-term average Fo distribution as personal voice features. *Speech Analysis and Synthesis, 3*, 209-228.

Jessen, M. (2009). Forensic phonetics and the influence of speaking style on global measures of fundamental frequency. In G. Grewendorf & M. Rathert (Eds.), Formal linguistics and law. (pp. 115-40). Berlin - New York: Mouton de Gruyer.

Johnson, K., Flemming, E., & Wright, R. (1993). The Hyperspace effect: Phonetic targets are hyperarticulated. *Language, 69(3)*, 505-528.

Johnson, C. C., Hollien, H., & Hicks, J. W. (1984). Speaker identification utilizing selected temporal speech features. *Journal of Phonetics, 12*, 319-326.

Johnson, K. & Mullennix, J. W. (1997). *Talker variability in speech processing*. Waltham, MA: Academic.

Kinoshita, Y., Ishihara, S., & Rose, P. (2009). Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition. *International Journal of Speech, Language and the Law, 16(1)*, 91-111.

Koopman, J. (1999). A Brief History of Opera. Retrieved from http://www.lawrence.edu/fast/koopmajo/standard.html.

Köster, O. & Köster, J. P. (2004). The Auditory-perceptualeEvaluation of voice quality in forensic speaker recognition. *The Phonetician, 89*, 9-37.

Köster, O. & Schiller, N. O. (1997). Different influences of the native language of a listener on speaker recognition. *Forensic Linguistics, 4*, 18–28.

Kreiman, J., Gerratt B. R., Kempster, G. B., Erman A., & Berke, G. S. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research, 36(1)*, 21-40.

Künzel, H. (2000). Effects of voice disguise on speaking fundamental frequency. *Forensic Linguistics, 7*, 149–179.

Künzel, H. J. (2001). Beware of the 'telephone effect': The influence of telephone transmission on the measurement of formant frequencies. *Forensic Linguistics, 8*, 80–99.

Künzel, H. J. (2004). Tasks in forensic speech and audio analysis: A tutorial. *The Phonetician, 90*, 9-22.

Kuwabara, H. & Sagisaka, Y. (1995). Acoustic characteristics of speaker individuality: Control and conversion. Speech Communication, 16(2), 165-173. Retrieved October 16, 2009, from http://dx.doi.org/10.1016/0167-6393(94)00053-D

Ladefoged, P. & Ladefoged, J. (1980). The ability of listeners to identify voices. *UCLA Working Papers in Phonetics, 49*, 43–51.

LaRiviere. C. L. (1975). Contributions of fundamental frequency and formant frequencies to speaker identification. *Phonetica, 31*, 185-197.

Lass, N., Hughes, K., Bower, M., Waters, L., & Bourne, V. (1976). Speaker sex identification from voiced, whispered and filtered isolated vowels. *The Journal of the Acoustical Society of America, 59*, 675-678.

Lavner, Y., Gath, I., & Rosenhouse, J. (2000). The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Communication, 30(1)*, 9-26.

Lundy, D., Roy, S., Casiano, R., Xue, J., & Evans, J. (1999). Acoustic Analysis of the Singing and Speaking Voice in Singing Students. *Journal of Voice, 14(4)*. 490-493.

Marrero, V., Battaner, E., Gil, J., Llisterri, J., Machuca, M. J., Marquina, M., et al. (2008). Identifying speaker-dependent acoustic parameters in Spanish vowels. In *Proceedings of Acoustics '08*. (pp. 5673-7). Paris, France, June 29 - July 5, 2008. Acoustical Society of America - European Acoustics Association - Société Française d'Acoustique. Retrieved January 20, 2010, from http://liceu.uab.cat/~joaquim/phonetics/VILE/VILE_Acoustics08.pdf

Marrero, V., Gil, J., & Battaner, E. (2003). Inter-Speaker variation in Spanish: An experimental and acoustic preliminary approach. In ICPhS 2003. *Proceedings of the 15th International Congress of Phonetic Sciences*. (pp. 703-6). Barcelona, Spain, 3-9 August 2003. Retrieved August 2, 2009, from http://liceu.uab.cat/~joaquim/phonetics/VILE/VILE_ICPhS03.pdf

Masthoff, H. (1996). A report on a voice disguise experiment. *Forensic Linguistics, 3*, 160–167.

McDougall, K. (2006). Dynamic features of speech and the characterization of speakers. Toward a new approach using format frequencies, *The International Journal of Speech, Language and the Law, 13(1)*, 89-126.

McGehee, F. (1937). The reliability of the identification of the human voice. *Journal of General Psychology, 17*, 249-271.

Mead, K. (1974). Identification of speakers from fundamental frequency contours in conversational speech. Joint Research Unit Report 1002, 1-22.

Meuwly, D. (2000). Voice Analysis, in SIEGEL, J.- SAUKKO, P.- KNUPFER, G. (Eds.) Encyclopedia of Forensic Science. London: Academic Press. pp. 1413-1420.

Milenkovic, P. H. (2001). *TF32 user's manual*. Madison: University of Wisconsin.

Narayan, C. & Yuan, J. (2008). Acoustic correlates of speaker discrimination in English. *The Journal of the Acoustical Society of America, 124(4)*, 2457-2457.

Niessen, M. (2004). Speaker specific features in vowels. Thesis. Artificial Intelligence, University of Groningen, Retrieved July 13, 2009, from http://scripties.fwn.eldoc.ub.rug.nl/FILES/scripties/Kunstmatigeintellige/doctoraal/2004/Niessen.M.E./AI-DKI-2004-M.E.Niessen.pdf

Nolan, F. (1983). The Phonetic Bases of Speaker Recognition. Cambridge: Cambridge University Press (Cambridge Studies in Speech Science and Communication).

Nolan, F. (2001). Speaker identification evidence: its forms, limitations and roles, in Proceedings of the conference "Law and Language: Prospect and Retrospect". 12-15 December 2001. Levi, Finnish Lapland.

Nolan, F. (2002). Intonation in speaker identification: An experiment on pitch alignment features. Forensic Linguistics: *The International Journal of Speech, Language and the Law, 9(1)*, 1-21.

Nolan, F. (2002). The 'telephone effect' on formants: *A response. Forensic Linguistics, 9*, 74–82.

Nolan, F., Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *The International Journal of Speech, Language and the Law, 12(2)*, 144-173.

Orchard, T. L. & Yarmey A. D. (1995). The effects of whispers, voice-sample duration, and voice distinctiveness on criminal speaker identification. *Applied Cognitive Psychology 9(3)*, 249–260.

Oyer, H. J. & Trudeau, M. D. (1984). Temporal speech characteristics and vocal pleasantness in the aging voice. *The Journal of the Acoustical Society of America, 75(S1)*, S41-S41.

Papcun, G., Kreiman, J, & Davis, A. (1989). Long-term memory for unfamiliar voices. *The Journal of the Acoustical Society of America, 85(2)*, 913-924.

Pennington, M. (2011). Toward phonetically grounded distinctive features. Part I: Acoustic-articulatory correlations in a four-region model of the vocal tract *IULC Working Papers, 11(1)*, 1-27.

Perrot, P., Morel, M., Razik, J., & Chollet, G. (2009). Vocal forgery in forensic sciences. France, 1-7.

Pollack, I., Pickett, J. M., & Sumby W. H. (1954). On the identification of speakers by voice. *The Journal of the Acoustical Society of America, 26(3)*, 403-406.

Quatieri, T. F., Jankowski, C. R. Jr., & Reynolds, D. A. (1994). Energy onset times for speaker identification. *IEEE Signal Processing Letters, 1(11)*, November 1994.

Reich, A. R. (1981). Detecting the presence of vocal disguise in the male voice. *The Journal of the Acoustical Society of America, 69(5)*, 1458-1461.

Reich, A. R., & Duke, J. E. (1979). Effects of selected vocal disguises upon speaker identification by listening. *The Journal of the Acoustical Society of America, 66(4)*, 1023-1028.

Rodman, R., McAllister, D., Bitzer, D., Cepeda, L., & Abbitt, P. (2002). Forensic speaker identification based on spectral moments. *Forensic Linguistics, 9(1)*, 22-43.

Rose, P. & Duncan S. (1995). Naive auditory identification and discrimination of similar voices by familiar listeners. *Forensic Linguistics, 2*, 1–17.

Saslove, H. & Yarmey A. D. (1980). Long-term auditory memory: Speaker identification. *Journal of Applied Psychology, 65(1)*, 111-116.

Scherer, K. R. (2002). Vocal communication of emotion: A review of research paradigms. *Speech Communication, 40*, 227-256.

Schiller, N. O. & Köster O. (1996). Evaluation of a foreign language speaker in forensic phonetics: A report. *Forensic Linguistics, 3*, 176–185.

Schmidt-Nelson, A. & Stern K. R. (1985). Identification of known voices as a function of familiarity and narrow-band coding. *The Journal of the Acoustical Society of America, 77(2)*, 658-663.

Schotz, S. (2005). Stimulus duration and type in perception of female and male speaker age. *Proceedings of Interspeech 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology*. Lisbon, Portugal, September 4-8.

Shrivastav, R. (2006). Perceptual attributes and assessment of the singer's voice. *The Performer's Voice*, ed. by Michael S. Benninger and Thomas Murry, 103-115. Abingdon, UK: Plural Publishing.

Shrivastav R., Sapienza C., & Nandur V. (2005). Application of psychometric theory to the measurement of voice quality using rating scales. *Journal of Speech, Language, and Hearing Research, 48(2)*, 323-335.

Schuartz, M. F. (1986). Identification of speaker sex from isolated voice fricatives. *The Journal of the Acoustical Society of America, 43*, 1178-1179.

Skelton, R. B. (1970). Individuality in the vowel triangle. *Phonetica, 21*, 129-137.

Smalls, Dawn L. (2004). Linguistic Profiling and the Law. *Stanford Law and Policy Review*, 15, 579.

Su, L., Li, K., & Fu, K. (1974). Identification of speakers by use of nasal coarticulation. *The Journal of the Acoustical Society of America, 56*, 1876-1882.

Sundberg, J (1977). Acoustics of the Singing Voice. *Scientific American, 236(3)*, 82-91.

Titze, I. R. (1994). Principles of Voice Production. Englewood Cliffs, N.J.: Prentice Hall.

Thompson, T. P. (1987). A language effect in voice identification. *Applied Psychology, 1*, 121-131.

Voiers, W. D. (1964). Perceptual bases of speaker identity, *Journal of the Acoustical Society of America 36*, 1065-1067.

Washi, Rumi. (2004). Japanese Female Speech. In S. Okamoto & J. S. Shibamoto Smith (Eds.), *Japanese Language, Gender, and Ideology: Cultural Models and Real People* (76). New York: Oxford University Press.

Wolf, J. J. (1972). Efficient acoustic parameters for speaker recognition. *The Journal of the Acoustical Society of America, 51*, 2044-2055.

Yarmey, A. D. (1991). Voice identification over the telephone. *Journal of Applied Social Psychology 21*, 1868–1876.

Yarmey, A. D. & Matthys E. (1992). Voice identification of an abductor. *Applied Cognitive Psychology 6*, 367–377.

Young, M. A. & Campbell, R. A. (1967). Effects of context on talker identification, *The Journal of the Acoustical Society of America, 42(6)*, 1250-1254.

Zalewski, J., Majewski, W., & Hollien, H. (1975). Cross-correlation of long-term speech spectra as a speaker identification technique. *Acustica, 34*, 20-24.

BIOGRAPHICAL SKETCH

Tyler McPeek has been keenly interested in creative approaches to language structures since he began writing poetry and fiction in his early teens. This interest was was the key to his decision to study formal linguistics at the graduate level, after a background in literature, earth sciences, and international business. He has been published in numerous academic journals, magazines, literary journals, online, and through individual publications. In his second year of college at St. Andrews University, he published his first collection of poetry, entitled “Questions of a Nightingale” with St. Andrews Press. This collection won the 1997 Bunn-McClelland Chapbook Award. In the fall of 1997, Tyler studied for a semester under the tutelage of the internationally renowned daughter/scholar and family of Ezra Pound at Brunnenburg Castle in the South Tirol of northern Italy. He has also studied for credit under various other scholars as an undergraduate student, including Ronald H. Bayes, Donald Keene, Carl Walters, Barbara Millhouse, and others, at an array of different institutions, including: Reynolda House Museum of American Art, Wake Forest University, the University of Hawaii at Manoa, and field study in India, South America, Europe, and throughout Asia.

In 1999, Tyler moved to Japan where he was a high school teacher in the Japanese government sponsored JET Program, as well as a small business owner and consultant to several companies in the Hokuriku region of western Japan. In 2005, he completed an MBA in International Business and Marketing from Ritsumeikan Asia Pacific University (立命館アジア太平洋大学). Following the completion of his MBA, he lived, worked, and studied in Seoul, South Korea for 1 year, where he collaborated with professors at Yonsei University and oversaw the implementation of business plans that

he wrote for a Korean health product company as part of his MBA thesis. During this time, he also traveled to San Francisco as a guest lecturer at UC Berkeley, as well as trips to Taipei, Bangkok (coordinating with professors and graduate students at Chulalongkorn University), and Tokyo for marketing and technical presentations and distribution contract negotiations. Tyler spent 2007 studying and doing research at Sophia University (上智大学) in Tokyo. Returning to his native Florida in 2008, he began his doctoral studies and a teaching assistantship in the Linguistics Department at the University of Florida. He received his MA in Linguistics and Graduate TESL and SLA Certificates from UF in 2010. Tyler is the President of the Florida Linguistics Association (FLA), www.floridalinguistics.com, proprietor of tylermcpeek.com and tencolors.com and has his academic research posted in numerous places online, including his doctoral dissertation results, audio samples, and related research on Vocal Typology, at voicetypes.com. Tyler's academic contributions in the fields of Linguistics and Business have been published and presented at various academic institutions, including Harvard University, Université de Paris-Sorbonne, The University of California, Berkeley, Chulalongkorn University, and Yonsei University.

Tyler's newest works of fiction and poetry, dealing with his time in Japan and his travels to various countries, including the continental US and Hawai'i, Canada, Mexico, Ecuador, Cuba, The Netherlands, Germany, Austria, Switzerland, Italy, France, Egypt, India, Indonesia, Malaysia, Thailand, Singapore, Australia, South Korea, Hong Kong, Mainland China, Taiwan, and various other countries and locales around the globe, are scheduled for publication in 2013 and 2014. This includes mixed language, experimental literary works, such as those in his most recent book of poetry—an

English/Japanese bilingual and language meshing collaboration with prominent Japanese visual artist Akira Hasegawa, entitled “Three Reflections of Cherry Blossoms.”

Receiving his Ph.D. from the University of Florida in 2013, Tyler plans on becoming a professor of language and language science at a university in the East Asian region, where he will continue both his creative literary work and linguistics research in Japanese and East Asian theoretical linguistics, Second Language Acquisition, Orthographic Influence on Borrowing Phonology Models, and Forensic Phonetics.