Naval Research Laboratory

Washington, DC 20375-5320

NRL/FR/5550--02-10,044

Voice Biometrics for Information Assurance Applications

George S. Kang
Yvette Lee

Transmission Technology Branch
Information Technology Division

December 5, 2002

Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)*
December 5, 2002

2. REPORT TYPE

3. DATES COVERED *(From - To)*
October 1, 2001-September 1, 2002

4. TITLE AND SUBTITLE

Voice Biometrics for Information Assurance Applications

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
33904N, 61553N

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
3415

6. AUTHOR(S)

George S. Kang and Yvette Lee

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Research Laboratory, Code 5550
4555 Overlook Avenue
Washington, DC 20375-5320

8. PERFORMING ORGANIZATION REPORT NUMBER

NRL/FR/5550--02-10,044

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Commander
Space and Naval Warfare Systems Center
4301 Pacific Highway
San Diego, California 92110-3127

10. SPONSOR / MONITOR'S ACRONYM(S)

11. SPONSOR / MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

In 2000, the President of the United States established an organization within the DoD to develop and promulgate biometrics technologies to achieve security in information, information systems, weapons, and facilities. NRL has been tasked to study *voice biometrics* for applications in which other biometrics techniques are difficult to apply. The ultimate goal of voice biometrics is to enable the use of voice as a password.

Voice biometrics are "man-in-the-loop" systems in which system performance is significantly dependent on human performance. This aspect has not been properly emphasized by previous researchers in this field. Accordingly, we let each speaker choose his (or her) own test phrase that can be uttered consistently. The speech waveform is then pre-processed (i.e., equalized and normalized) to reduce the effect of inconsistent speaking. Subsequently, we extract five different voice features from the speech waveform. Some of them have never been used for voice biometrics. Finally, individual feature errors are combined to indicate a confidence level of speaker verification.

Initial laboratory testing under various conditions shows encouraging results. We will be prepared to fleet-test our voice biometrics system in FY03.

15. SUBJECT TERMS

Speaker recognition; Speech pre-processing, Speech feature extraction

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					George S. Kang
a. REPORT	b. ABSTRACT	c. THIS PAGE	UL	43	19b. TELEPHONE NUMBER *(include area code)*
Unclassified	Unclassified	Unclassified			(202) 767-2157

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

CONTENTS

INTRODUCTION 1

BACKGROUND 2

DoD Biometrics Thrusts 2
Navy Biometrics Programs 3
NRL Voice Biometrics Tasks 3
Two Objectives of Voice Biometrics 4
Differences Between Voice Biometrics and Other Biometrics 5
Historical Perspective of Voice Biometrics R&D 6

SPEAKER ENROLLMENT 7

Test Phrase Selection 7
Template Information 10
Template Depository 10

SPEECH PRE-PROCESSING (SPEECH NORMALIZATION) 10

Peak Amplitude Normalization 11
Adaptive High-Frequency Boost 12
Speech Waveform Cropping 13
Speech Bandwidth Expansion 13
Gas Mask Speech Improvement 15

VOICE FEATURE EXTRACTION 21

Speech Spectral Histogram 22
Vocal Tract Area Histogram 23
Speech Loudness Histrogram 25
Average Pitch Frequency 25
Speech Energy Distribution 27

SPEAKER VERIFICATION 29

Time Alignment (Time Warping) 29
Individual Voice Feature Errors 31
Total Voice Feature Error and Weighting Factors 32
Self-Test Score 33

Inter-Speaker Scores 33
Template Update 33
Test Scores with Tape-Recorded Input Speech 34
Test Scores of Impostors 34
Test Scores with Gas Mask 36

CONCLUSIONS 36

ACKNOWLEDGMENTS 37

REFERENCES 37

VOICE BIOMETRICS FOR INFORMATION ASSURANCE APPLICATIONS

INTRODUCTION

The Department of Defense (DoD) has a critical need to ensure security and integrity in all its information and information systems, weapons, and facilities. To meet this need, DoD is exploiting biometrics technology that uses measurable physical characteristics of an individual for identification or verification. LTG Peter Cuviello, Director of the DoD Biometrics Management Office (BMO), envisions that, "on the battlefield of the future, a soldier's voice will be his password."

In support of DoD biometrics efforts, we at NRL are focusing on voice biometrics for applications where other biometrics techniques are difficult to apply. An example of voice biometrics is remote speaker verification over computer or communication links in which the user and the information source are not collocated. Another example is when the speaker wears a gas mask and gloves in a nuclear-biological-chemical (NBC) environment. We are also pursuing research and development aimed at improving voice biometrics accuracy and reducing the time required to perform speaker verification. Reduction of the time required to perform speaker verification is an important issue because people are used to operating a cipher lock, which takes only two seconds. Biometrics access control devices might take longer than two seconds, and this is generally not acceptable.

Presently, voice biometrics is technologically behind other biometrics in terms of verification accuracy, user-friendliness, and even availability. Voice biometrics devices are seldom demonstrated at vendor exhibits because they take too much time to demonstrate. This indicates a lack of user-friendliness in current voice biometrics products. Furthermore, few voice biometrics devices are available in the marketplace. Recently, BMO tested 56 biometrics devices, of which only two were voice biometrics devices [1].

In an attempt to fill the technology shortfall in voice biometrics, we have developed a reliable and user-friendly voice biometrics system for the DoD and the Navy. In the beginning, we carefully studied why voice biometrics performance trailed other biometrics. We finally realized that voice biometrics devices are *man-in-the-loop systems* in which overall system performance is dependent on human performance or, as we state, the speaking manner of the person to be verified.

A fingerprint pattern does not change. In contrast, voice features change constantly, depending on the emotional and physical condition of the speaker. Therefore, critical elements in the voice biometrics are:

- *Selection of test phrase* — that can be uttered consistently;
- *Speech pre-processing* — to normalize voice biometrics features prior to speaker verification;
- *Speech processing* — to extract several speaker-dependent voice features for speaker verification; and
- *Methods of comparing voice features* — that are not usually aligned in time.

Manuscript approved September 17, 2002.

We made a fresh look at these areas. For every approach presented, we discuss several tradeoffs and provide recommendations. We also developed our own biometrics device and are getting satisfactory speaker verification performance.

The sponsor of our efforts (Space and Naval Warfare Systems Command (SPAWAR) PMW-161) is currently testing various biometrics devices in Navy offices and on Navy platforms. To take advantage of this ongoing program, we plan to conduct tests on our voice biometrics device in similar environments in FY03.

BACKGROUND

DoD Biometrics Thrusts

On July 13, 2000, the President of the United States established an organization within DoD to develop and promulgate biometrics technologies to achieve information system security (Section 112 of PL 106-246). Accordingly, DoD established the Biometrics Management Office (BMO) with the Army as its executive agent. The mission of BMO is to promote the use of and to ensure the availability of biometrics technologies within DoD. In addition, the Biometrics Fusion Center (BFC) has been established to test and evaluate commercial off-the-shelf (COTS) products and to serve as a repository of biometrics information. All three Services are committed to the pursuing biometrics efforts, as noted by the visionary statements of the nation's top-ranking officers in the field of information warfare and information assurance (Fig. 1).

VADM Richard Mayo, US Navy
Director of Space, IW, Command and Control (CC)
OPNAV, N6

"Ensuring rapid, secure and authorized access to DoD, military and intelligence community systems is a challenge of the highest national importance."

LTG Peter Cuviello, US Army
Director of Information Systems for Command, Communications and Computers (DISC4)

"On the battlefield of the future, a soldier's voice will be his password."

MG Charles Croom, US Air Force
Joint Staff, Director of C4 Systems

"Biometrics will play a critical role in securing the information systems of the future."

Fig. 1 — Three Service leaders in DoD biometrics efforts

Navy Biometrics Programs

SPAWAR PMW-161 leads the Navy biometrics programs. In SPAWAR, biometrics is regarded as "emerging technologies." Jim Davies is the program director and Dave Guerrino is the Navy point-of-contact (POC) to BMO. Currently, SPAWAR is pursuing five different biometrics efforts:

- Introduce biometrics technologies to various Navy operational sites and evaluate user acceptance.
- Test and evaluate biometrics products for Navy applications.
- Identify candidates of Navy weapons or C3I systems and determine the benefits and deployment costs.
- Investigate biometrics assurance issues, including protection profiles, security targets, and vulnerability assurance.
- Focus on voice biometrics: develop a reliable and user-friendly speaker authentication for use in unique Navy situations and where other biometrics methods are difficult to apply. (This is our biometrics project at NRL, and it is further described in this report.)

NRL Voice Biometrics Tasks

We are in the Voice Systems Section (Code 5555), which is part of the Transmission Technology Branch of the Information Technology Division at NRL. We support the Navy's biometrics effort that was begun on November 17, 2000 at the Navy-BMO conference. The Navy representatives in this meeting were Earle Kirkley and Jim Davies, both of the Technical Direction Division of SPAWAR PMW-161, Dave Guerrino, the Navy Biometrics point of contact at BMO, and our NRL section members. At this meeting, NRL Code 5555 was designated as the Navy Voice Biometrics R&D Group. Currently, we have been performing the following three projects:

1. *Test and evaluate remote biometrics over Navy secure voice links* — In this application, the person who wishes to access the controlled information is not collocated with the information system (Fig. 2). It is convenient to perform biometrics by voice over the voice link. NRL is currently prototyping a three-way voice conferencing system with speaker verification. NRL is also evaluating voice biometrics performance using vocoded speech and noisy speech. This phase of investigation will be documented separately when completed.

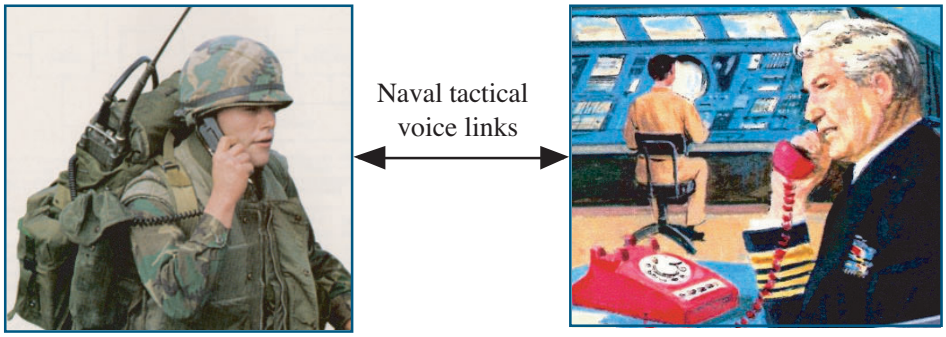


Fig. 2 — Remote biometrics. When a speaker tries to access classified information via voice communication links, remote speaker verification over the existing communication system is convenient.

2. *Develop speech equalization techniques for voice biometrics when wearing gas mask and gloves —* In future warfare, it is expected that soldiers will use gas masks, referred to as the Mission Oriented Protective Posture, Level 4 (MOPP-4) (Fig. 3). When a person wears a gas mask, an opaque window in front of the eyes make iris scans, retina scans, or facial mapping techniques difficult to apply. Likewise, if the soldier is wearing gloves, the fingerprint technique cannot be used. The use of voice biometrics is more practical.

Fig. 3 — MOPP-4 attire. There are two acoustic paths from inside the mask to outside: one is via the internal built-in microphone; the other is a direct path through a plastic membrane, called the "voicemitter." Speech from either path sounds muffled, fuzzy, and tinny. Voice biometrics need speech equalization techniques to remove speech distortion.

3. *Develop a reliable and user-friendly voice biometrics system for access control —* A major application of voice biometrics is for controlling access to classified facilities (Fig. 4), information systems, and weapon systems.

Fig. 4 — Voice biometrics for access control. The term "user-friendly" means "quick to verify" voice biometrics for access control applications. The ultimate goal of voice biometrics is to use our voice as our password.

Two Objectives of Voice Biometrics

Two major applications of voice biometrics are speaker verification for access control and speaker recognition for speaker monitoring over communication links:

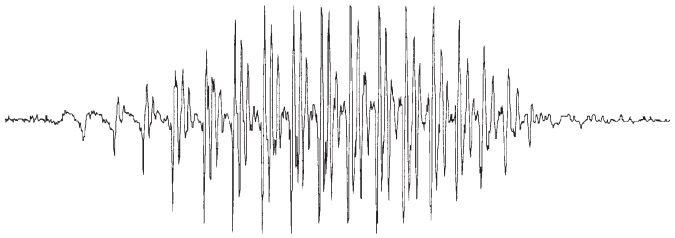
- *Speaker verification for access control —* The purpose is to verify or authenticate the speaker under test by means of his (or her) own spoken voice. Since the biometrics operation is performed in a cooperative environment, the spoken text can be specified beforehand. The biometrics device provides a binary decision—yes or no. This report is written for this application.

- *Speaker recognition for speaker spotting or speaker monitoring over communication link* — The purpose is to spot a specific person from the file containing voice features or from live speech intercepted over communication links. This report does not cover this voice biometrics application.

Differences Between Voice Biometrics and Other Biometrics

Voice biometrics is different from other biometrics in many respects. Table 1 lists three major differences using the fingerprint-comparison biometrics as an example of nonvoice biometrics. These differences make the implementation of a reliable and user-friendly voice biometrics device difficult. Although we show the case of fingerprint comparison biometrics as an example of nonvoice biometrics, other nonvoice biometrics have similar differences (Table 1). Stated briefly, nonvoice biometrics features are stable, and they are primarily related to the person (unlike the speech waveform, which is related to both the words uttered and speaker's voice characteristics).

Table 1 — Differences Between Voice and Nonvoice Biometrics

(a) Fingerprint	(b) Speech Waveform
	/help/ 
What we see (i.e., the fingerprint) is the information directly used for personnel verification.	What we see (i.e., the speech waveform) is complex and repetitive, that is, in a form not directly useful for voice biometrics. Therefore, we must process the speech waveform to extract voice biometrics features.
The fingerprint is unique to each person.	The speech waveform is related to both speech and speaker. Hence, we have to extract voice features that are more sensitive to voice characterics than to speech characterictics.
Fingerprints do not change over a long period of time. Hence, the biometrics performance is stable (i.e., repeatable).	Voice characteristics change from time to time. Hence, we must do the following to minimize adverse effects of voice variabilities on the voice biometrics performance: • Select a phrase that each speaker can utter consistently • Normalize the speech loudness • Normalize utterance speed • Provide adaptive high-frequency boost • Use several voice features jointly • Update voice template, as necessary

Historical Perspective of Voice Biometrics R&D

The advent of high-speed digital signal processors in the early 1970s made it possible to use complex signal processing algorithms in speech processing applications. These applications include speech encoding, speech recognition, and speaker verification. In those days, voice biometrics was a very popular subject in speech research [2,3,4]. By the mid-1970s, voice biometrics technology had advanced enough to produce technology survey papers [5,6].

A major topic in voice biometrics has always been the voice feature and its effect on speaker verification performance. In the early 1970s, linear prediction analysis of the speech waveform was developed as a means to represent the speech spectral envelope by a limited number of parameters [7]. As a result, linear predictive coefficients were widely used as a voice feature [8-10]. Speech intensity [11] and pitch [5,12] were other, often-used voice features. Likewise, studies were made on the effects of processing raw voice features; for example, long-term averaging [13] or clustering [14,15]. Other voice biometrics issues, such as feature weighting [16], feature updating [17], and time warping [18,19], were also studied.

With all these voice biometrics R&D efforts behind us, it could be claimed that everything that can be done for voice biometrics has been done. The fact remains, however, that no existing voice biometrics device exists that can utilize one's voice as a password. We feel that previous voice biometrics works have not emphasized the man-in-the-loop aspect of the problem. In other words, voice biometrics performance is significantly dependent on human performance.

Furthermore, previous research has neglected the effect that stress on the speaker (physical, mental, and/or emotional stress—all of which are common in military situations) can have a detrimental effect on the performance of any voice biometrics system. Measurements have shown that people's speech can change markedly under stress conditions: they often talk faster, louder, and raise the their pitch. Sometimes there are also changes to the speech spectra, particularly vowels [20]. These stress-induced voice reactions can severely impact accuracy, potentially leading to real problems in real-world conditions. In the design of our voice biometrics system, we have considered the impact of human performance, and have developed techniques to overcome inconsistency in human speaking. We place a special emphasis on test word selection as part of enrollment. It must be noted that pre-processing to equalize speaking inconsistency has not been emphasized in previous work. As indicated by Fig. 5, our voice biometrics system has four major components. We introduce new ideas and improve old ideas in the implementation of each component:

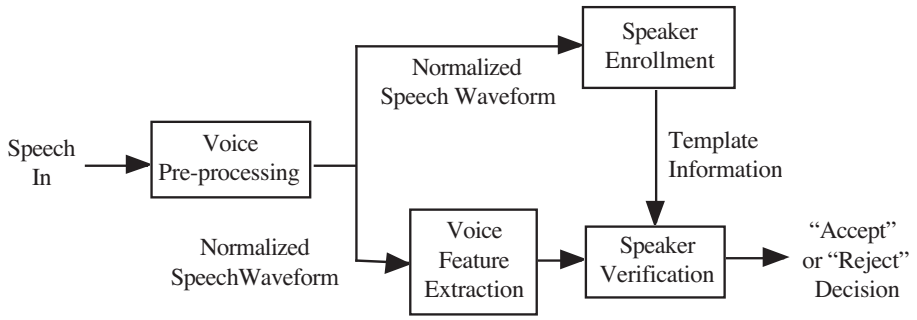


Fig. 5 — Voice biometrics system

1. *Enrollment* — We let the user select his (or her) own test phrase to minimize intra-speaker distance while maximizing inter-speaker distance.

2. *Speech pre-processor* —The speech waveform is normalized and equalized to minimize the effect of speaking inconsistencies that are reflected in the raw speech waveform.

3. *Voice feature extraction* — We use as many as five voice features jointly.

4. *Comparison of voice features* — We take a fresh look at time-alignment, weighting factors for the individual error, and updating the voice template to cope with slowly changing voice characteristics.

One area that we do not discuss is voice biometrics in noisy environments because we already have a solution to this problem. In 1999, we developed what-is-called the second-order gradient microphone that can reduce ambient noise as much as 20 to 30 dB [21]. The use of this microphone also solves the problem of reverberation because it effectively attenuates farfield sounds (i.e., late arriving multipath signals). The effect of vocoded speech on voice biometrics will be investigated and documented by the principal investigator of another task at NRL.

SPEAKER ENROLLMENT

The enrollment process characterizes each person's voice biometrics features and stores the information into the biometrics device to be used for speaker verification (Fig. 6). We discuss several important issues related to enrollment for voice biometrics.

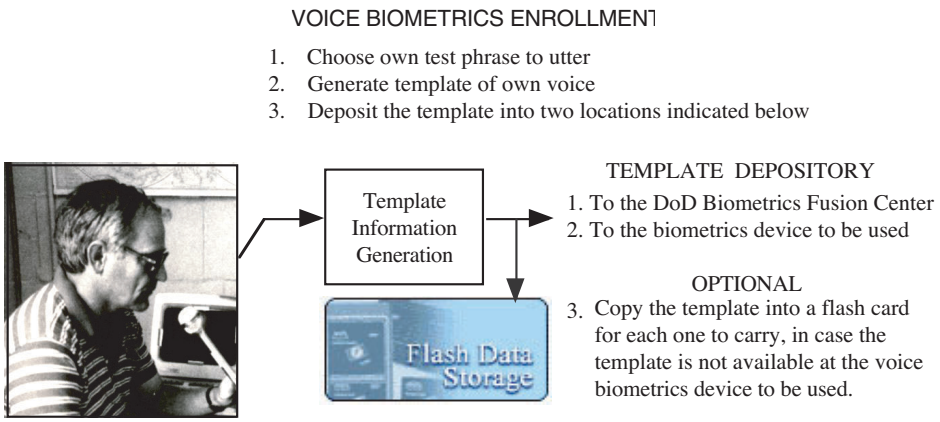


Fig. 6 — Enrollment process for the NRL voice biometrics

Test Phrase Selection

Voice biometrics needs a person's voice to verify identity; however, the voice need not be just spoken words. A singing voice is an alternative to spoken words because a singing voice contains enough information to recognize a particular person. Likewise, other vocalizations (such as the howling sound of a coyote) would acceptable as the speaker verification test signal. For most people, however, speech is a convenient form of test signal. Therefore, a test phrase must be selected. The test phrase should be selected to achieve the following general goals:

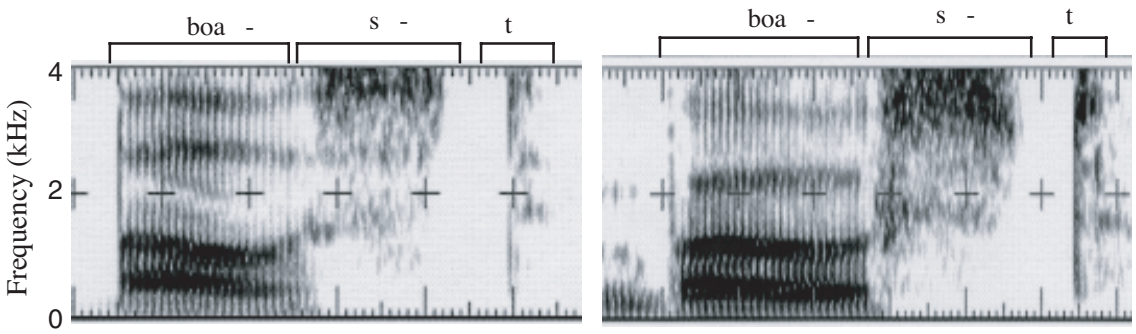
- *The inter-speaker difference is increased* — A test phrase is selected so that resultant voice features are easily distinguished between different people.

- *The intra-speaker distance is decreased* — The selected phrase must be such that the person can utter it consistently (i.e., in a similar rhythm, loudness, pitch, and speed). Then, the voice features will be tightly clustered together so that the speaker will always be accepted by the voice biometrics system.

There are two ways by which a test phrase can be selected: (1) the voice biometrics device specifies a unique phrase for each speaker verification session, or (2) the speaker chooses a preferred phrase beforehand and uses it at all speaker verification sesions. We discuss the pros and cons of each approach.

Approach #1: *The biometrics device selects the test phrase for each speaker verification session* — An advantage of this approach is that it discourages an imposter from impersonating the authorized speaker by using tape-recorded speech. However, specifying a phrase by the voice biometrics device is a poor idea because:

- Since the phrase varies at each speaker verification session, the user will not have an opportunity to practice the chosen phrase so as to speak it consistently.
- A test phrase will be selected without consideration of an individual's speaking idiosyncrasies to maximize the voice feature distance to other speakers.
- If the selected phrase is too simple (as in Fig. 7), everyone's voice features will look alike (i.e., it has no speaker verification capability).



(a) Speaker A (a) Speaker B

Fig. 7 — Speech spectrograms of the same word (/boast/) spoken by two different persons. This figure illustrates that if the test phrase is too simple, voice features (the voice print in this example) generated by different speakers are very similar. Therefore, appropriate test phrases must be chosen to highlight individual voices.

Approach #2: *The individual speaker selects a test phrase (our approach)* — In the NRL voice biometrics system, the individual speaker chooses his (or her) own test phrase. The selected phrase must be such that the speaker can utter it consistently (with the same pitch, stress, rhythm, and speed), yet differently from how other people would say it. An advantage of this approach is that each speaker verification session provides an opportunity for the speaker to practice the chosen phrase to utter more consistently. As a result, the speaker verification score will improve as speaker verification is performed often.

In selecting a test phrase, we suggest the following:

- Choose a phrase that has been used by the speaker many times in the past (for example, children's names, high school name, etc.). The speaker will have a definite way of saying these familiar words.
- Choose a phrase that other people cannot imitate easily (for example, a phrase in another language such as "bis mächste mal," "como se yama," etc.).
- Choose more complex phrases because this provides more diversified voice biometrics features.

- Avoid co-articulated words (e.g., This is an end) because transitional speech features are harder to compare.

- Choose three or four isolated nonsense words (e.g., man sh man shi buna ghun ga mog — this phrase was actually used by a person at NRL). Since there is no meaning, we tend to say nonsense phrases without varying inflections or intonations. [According to Stephanie Everett of NRL, however, the word "Chaubaunagungamaug" is the traditional Indian (Algonquian) name for Webster Lake in south central Massachusetts. Therefore, the cited test phrase may not be entirely "nonsense" words for some of the speakers.]

- Avoid a complete sentence because there are many different ways of placing stress, pause, and pitch inflection—depending on the emphasis. People tend to forget how they previously said a complete sentence. Inconsistent inflections are detrimental to voice biometrics (Fig 8).

In conclusion, we recommend the choice of three or four, multisyllable, nonsense words. Live tests by numerous persons walking though our laboratory confirmed that such test phrases worked well.

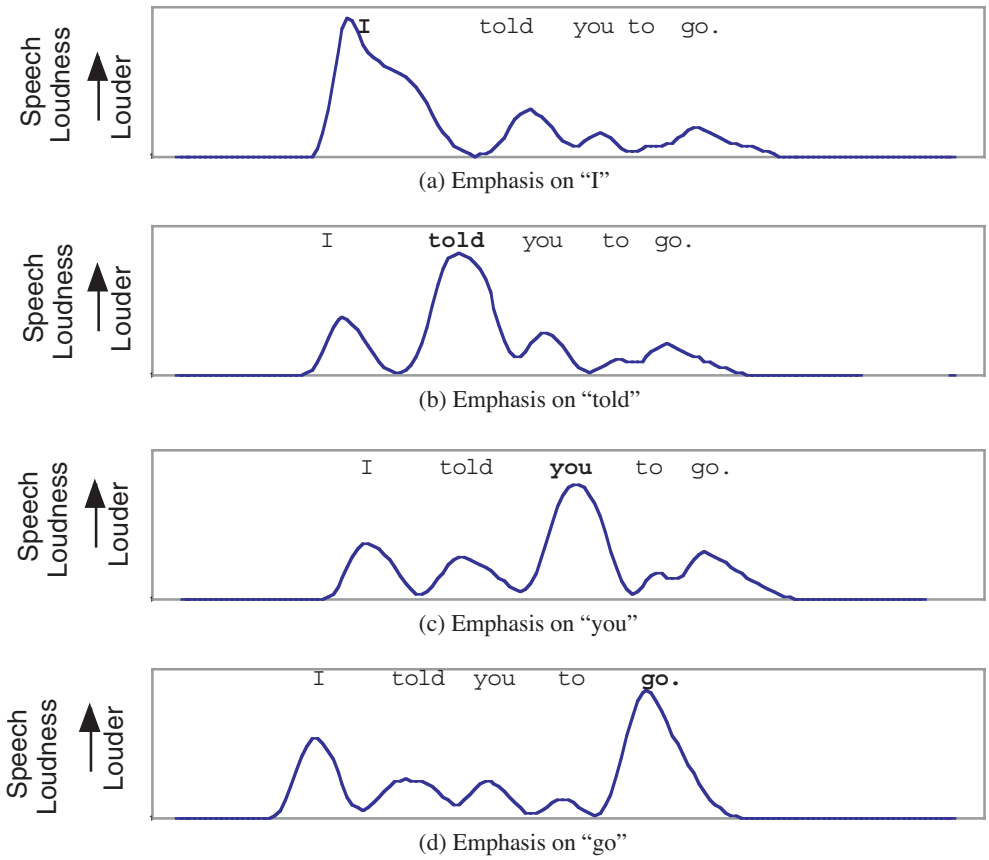


Fig. 8 — Four different prosody patterns. The sentence, "I told you to go," could have four different rhythms, stresses, and intonations, depending on the emphasis in the speaker's mind. These differences are shown in the trajectories of speech loudness (speech energy) here. A complete sentence often does not consistently generate voice features. Alignment of voice features is a critical issue for voice biometrics. Therefore, we do not recommend selection of a complete sentence as a test phrase for voice biometrics.

Template Information

Two general classes of voice information can be recorded in the template:

Approach #1: *The voice parameters used by the specific biometrics device* — The advantage of this approach is that the total amount of the information to be stored in the template is small (about 1 to 5% of data representing the raw speech waveform). But this approach has serious disadvantages that were eventually proven to be both costly and flawed*:*

- The template will be locked in one biometrics device that is currently in use.
- If a new or improved voice biometrics device is available, new templates must be issued to all users.

Approach #2: *Unprocessed speech waveforms (our approach)* — If the fingerprint-matching biometrics method stores each fingerprint as it is, we do not see why we cannot store the speech waveform as it is for the voice biometrics. The disadvantage of this approach is that the amount of data to be stored is larger compared to the previous approach. The minimum amount of information we need to perform voice biometrics is two sets of spoken voice. Since the test phrase is usually no more then two seconds, and the speech-sampling rate is 16 kHz, the amount of data needed to be stored is approximately 64 kbits or 8 kbytes, which is still rather small for present-day memory card capacity. A 1.25×1.5-inch flash card can hold 128,000 kbytes of data, 2000 times more than we need. The advantages of this approach are:

- Once a template is made, it is universally usable with all other current or future voice biometrics devices.
- It is easy to update the template information.

Template Depository

BMO mandated that all biometrics data be deposited at the Biometrics Fusion Center (BFC). In addition, the template must be resident in the biometrics device to be used. If the voice biometrics device does not have the speaker's voice template, the voice template must be downloaded from the BFC. This will require time. If there is no communication link, the biometrics processor cannot be used.

To circumvent this difficulty, the NRL voice biometrics system is equipped with a flash card reader. Each speaker can feed in his (or her) own voice biometrics features. The information in the template will be encrypted to prevent misuse if the template is lost.

Another advantage of the individual carrying his (or her) own template is that the template information can be updated at the speaker verification site, if the speaker verification decision is "acceptable" but the score is near the rejection threshold. In this way, the voice biometrics system adapts to any change of the speaker's voice.

SPEECH PRE-PROCESSING (SPEECH NORMALIZATION)

For all biometrics other than voice, biometrics features are fixed; for example, a fingerprint does not change unless it is accidentally cut or bruised. In voice biometrics, voice features change frequently.

Therefore, it is essential that the speech waveform be normalized, equalized, or conditioned prior to performing speaker verification. Speech pre-processes we will use include:

- Normalizing the peak amplitude of the speech waveform,
- Adaptively boosting high frequencies of speech for the spectral analysis,
- Cropping the speech waveform with a fixed rule,
- Using wider speech bandwidth to extract more voice features, and
- Removing speech distortion when a gas mask is worn.

Peak Amplitude Normalization

Loudness of speech is controlled by the amount of airflow from the lungs, and it is difficult to control speech loudness accurately based on auditory feedback of own voice. Furthermore, it is almost impossible to remember how loud one said the test phrase on previous occasions. Therefore, speech amplitude must be normalized before the speaker verification process. Amplitude levels must be matched because:

- Synchronization is essential for accurate comparison of two different voice features. Amplitude normalization makes it possible to crop the two different speech waveforms at the same speech onset locations.

- Speech level affects all voice features. Therefore, amplitude normalization reduces amplitude-dependent errors of voice features.

- Normalization of peak amplitude may also improve robustness in noisy environments, since it is well known that people talk more loudly in the presence of noise (this is part of the Lobaerd effect).

We normalize the amplitude during the enrollment session when the speech waveform is recorded. While the speech waveform is being recorded, the voice processor monitors the peak amplitude in real time (unless otherwise stated, the peak amplitude means the absolute value of the peak amplitude). For a 16-bit speech processor, the total dynamic range is from –32767 to 32767. If the measured peak amplitude is 32767, this is an indication that some speech samples have been clipped due to speaking too loudly. If so, the speech must be recorded again because clipped speech amplitudes are detrimental to extracted voice features. In this case, a pre-recorded voice message saying "Too loud. Record again," will be played back to the speaker.

When the test phrase has been recorded, the amplitude normalization process begins immediately and automatically. The speech waveform is amplified or attenuated by a fixed gain defined by the quantity in the bracket of Eq. (1):

$$\hat{e}(i) = \left[\frac{32000}{\hat{e}} \right] e(i), \tag{1}$$

where $\hat{e}(i)$ and $e(i)$ are the ith speech samples before and after amplitude normalization respectively, and $\hat{e}$ is the absolute value of the peak amplitude of $e(i)$. Note that amplitude normalization based on the root-mean-square (rms) level is not recommended because the normalized speech amplitude might be still clipped because of the presence of peaky speech waveforms (e.g., /i/, /ae/).

Adaptive High-Frequency Boost

If the glottis of a person closes slowly at the pitch rate, the glottis signal will lack high frequencies as an excitation signal of the vocal tract. The resultant speech signal likewise will lack high frequencies. If so, the speech spectral histogram is difficult to analyze. To alleviate this difficulty, high frequencies may be adaptively boosted in real time prior to spectral analysis. The simplest high-frequency booster is a differentiation network with a variable gain:

$$e_h(i) = e(i) - \mu e(i-1), \tag{2}$$

where $e(i)$ and $e_h(i)$ are the speech waveforms before and after the adaptive high-frequency boost, respectively, and the factor μ is a gain factor that controls the amount of high-frequency boost. The factor μ is determined by minimizing the mean-square value of the output of Eq. (2). Thus, the value of μ is obtained from

$$\frac{\partial \overline{e_h^2(i)}}{d\mu} = \overline{[e(i) - \mu e(i-1)][-e(i-1)]} = 0. \tag{3}$$

Hence,

$$\mu = \frac{\overline{e(i)e(i-1)}}{\overline{e^2(i-1)}}, \tag{4}$$

which is computed each frame (i.e., 50 times per second). The quantity μ will be larger for the voice that has a smaller quantity of high frequencies. Then, high frequencies will be boosted more, as indicated in Fig. 9.

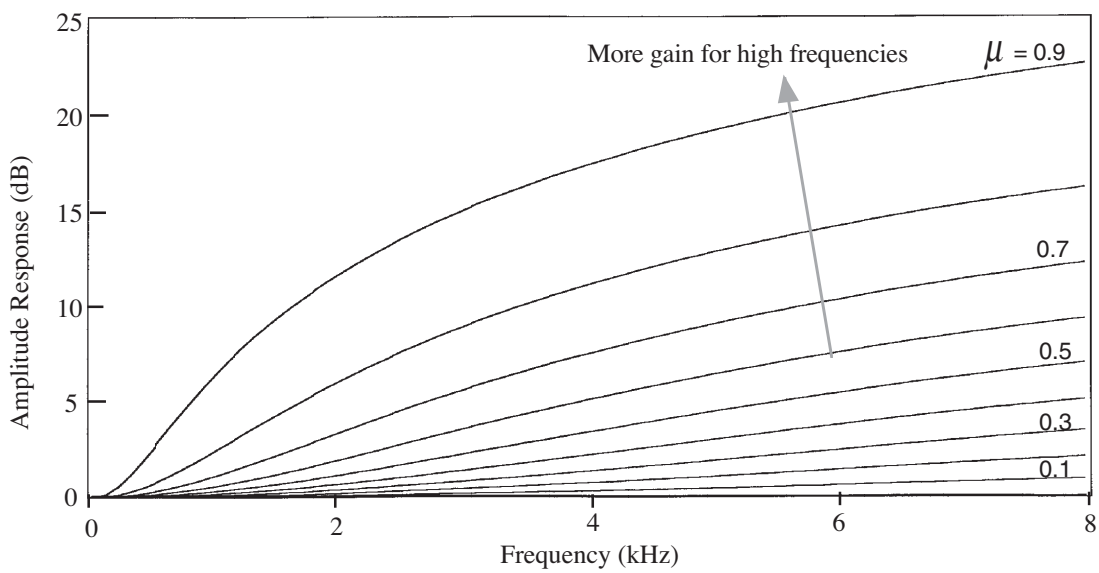


Fig. 9 — Frequency response of adaptive high-frequency booster. If the spoken voice has weak high-frequency components, they are automatically amplified in proportion to the value of μ, which is estimated at 50 times per second. According to the results of our test based on speaker verification scores, many speakers benefit from the use of adaptive high-frequency boosting.

Speech Waveform Cropping

In real time, speech flows. Once the recording process has started, the digitized speech samples will accumulate endlessly whether speech is present or absent. Therefore, the speech waveform must be cropped at speech onset so that two sets of speech features can be compared in phase. Since cropping is based on speech amplitude, waveform cropping must follow amplitude normalization.

Waveform cropping requires three frames of speech waveform and three frames of speech rms values (Fig. 10). If the rms values of all three consecutive frames are below a threshold level, the speech sample of the oldest frame (i.e., speech samples of the third frame, $\{x_3\}$) will be cropped. The threshold level is set to 256, which is less than 1% (–40 dB) of the normalized peak amplitude of 30,000. This process is repeated once every frame.

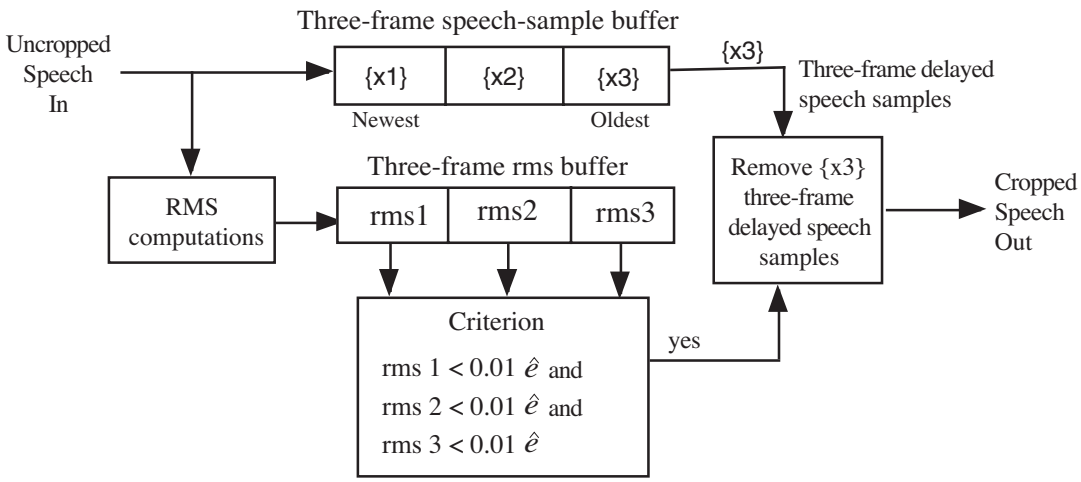


Fig. 10 — Speech waveform cropping. Each block represents one frame of speech samples. Each is 20 ms. If rms values of three consecutive frames are below 256 (40 dB below the peak denoted by $\hat{e}$), then the third frame (the oldest frame) of speech samples is cropped. (Figure 11 shows an example of good waveform cropping.)

Speech onset is often accompanied by two undesirable artifacts that are detrimental to waveform cropping:

- *Pre-onset burst of noise* — This relates to the presence of a "spit" noise generated when the mouth is opened. This feature is occasionally present, depending on the speaker and circumstance, and is detrimental to waveform cropping because this noise can be mistaken as a speech onset. The challenge is to isolate this artifact when enrolling and verifying a speaker.

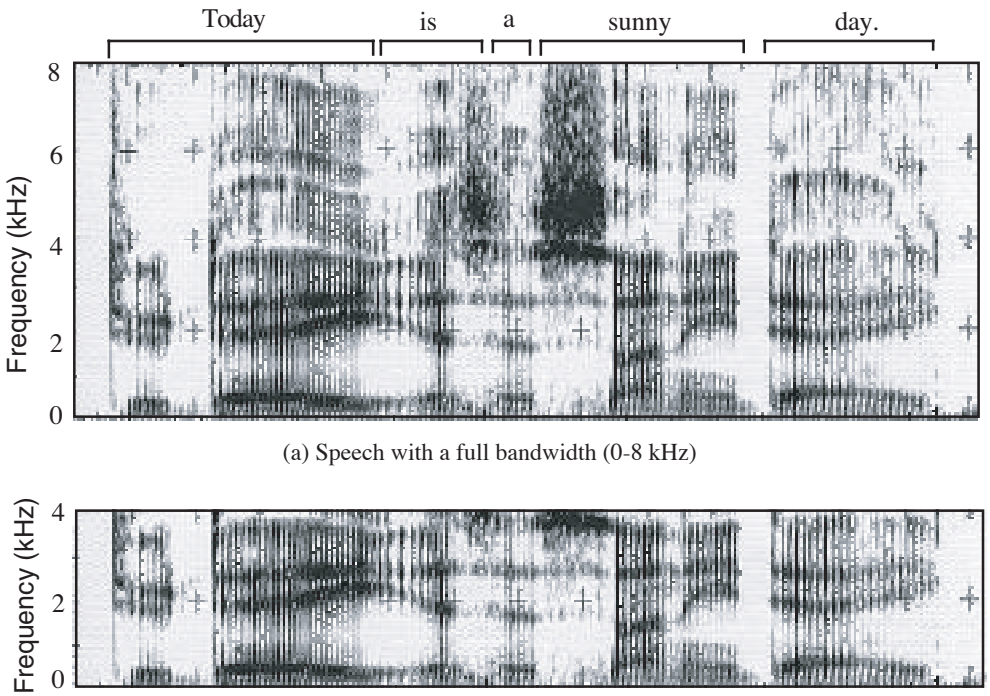
- *Puff noise* — The other artifact is noise caused by turbulent air accompanying fricatives. Such noise introduces an error in waveform cropping and also distorts the speech waveform near the onset. This artifact can be reduced by covering the microphone with a "puff screen."

Speech Bandwidth Expansion

Most people's voices have a speech bandwidth that extend up to 8 kHz or more. Yet speech bandwidth is often truncated to 4 kHz for speech transmission, speech recognition, speaker verification, etc. Unfortunately, this old practice has been rolled over to digital speech processing. In most cases, speech bandwidth is

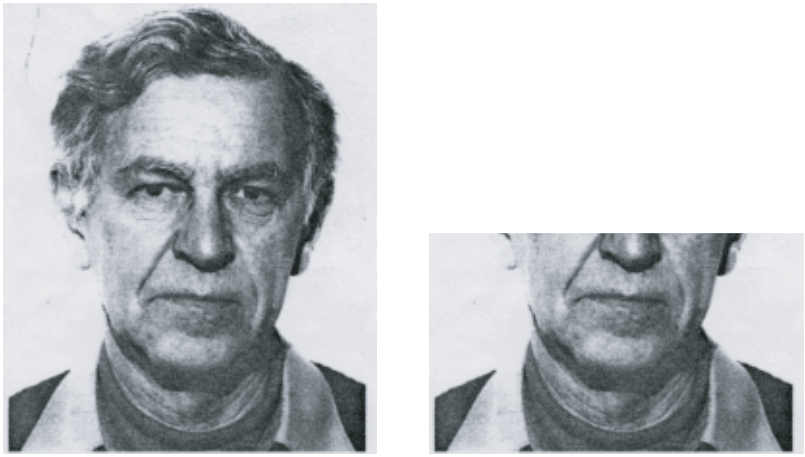
still restricted to 4 kHz. Figure 11 shows that a significant amount of speech information is lost by truncating the speech bandwidth to 4 kHz. In the NRL voice biometrics system, a speech bandwidth of 8 kHz is used.

We advocate the use of wideband speech (0-8 kHz) for voice encoding because it is useful in noisy environments, particularly for female speech [22]. We even generated a technique by which fricative spectra normally present above 4 kHz can be spread below 4 kHz to benefit narrowband users [22,23]. Recently, the use of wideband speech has been advocated for voice biometrics [24,25].



(a) Speech with a full bandwidth (0-8 kHz)

(b) Speech with the conventional, truncated bandwidth (0-4 kHz)



(c) Portrait in full view (d) Portrait in half view

Fig. 11 — Spectral comparison of full-band (0-8 kHz) and commonly used half-band (0-4 kHz) speech. Use of half-band speech signal for voice biometrics is like using a half face for the face-matching biometrics.

Gas Mask Speech Improvement

The gas mask is part of the nuclear, biological, and chemical (NBC) protective gear that also includes suits, gloves, and boots. In the military, the NBC protective gear is often referred as the Mission Oriented Protective Posture (MOPP). The MOPP has five levels, as shown in Fig. 12 [25].

The gas mask has continuously evolved since Leonardo da Vinci (1452-1519) first designed a simple protective mask to protect against toxic powder he himself developed. Reference 26 provides a comprehensive gas mask history.

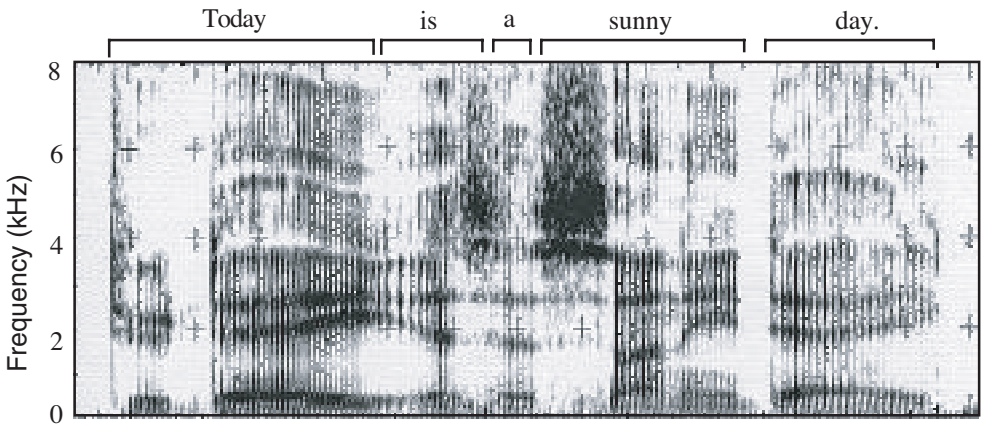
There are two speech paths from inside the mask to outside (Fig. 13). One speech path is through a vibrating membrane in the mask wall that receives speech sounds and re-emits toward the outside. This is called the "voicemitter," an abbreviation of the "voice emitter." The other speech path is a standard microphone installed inside the mask.

MOPP-0 MOPP-1 MOPP-2 MOPP-3 MOPP-4

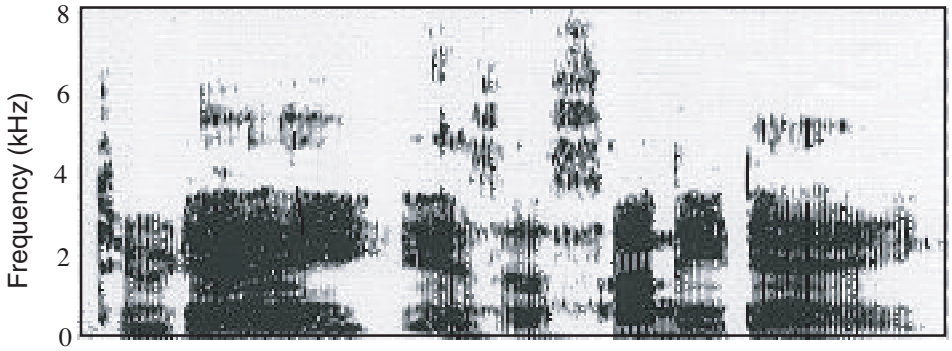
Fig. 12 — Five levels of MOPP. With MOPP-3 or MOPP-4, many biometrics techniques cannot be applied conveniently because of the gas mask and gloves. The most practical means for speaker verification is by the use of voice biometrics.

Fig. 13 — MOPP-4 outfit. The speech heard outside the mask is significantly muffled and tinny. As a result, the performance of a voice biometrics device will be degraded. Restoration of MOPP-4 speech to natural sounding speech is essential for voice biometrics.

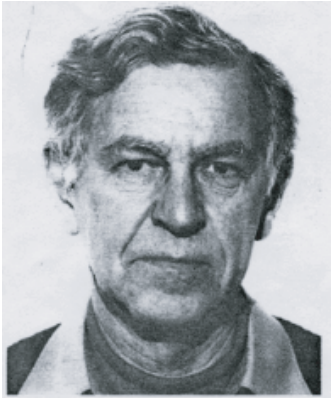
When the speaker wears a gas mask and gloves, the speaker's voice is the most convenient means for personnel verification. However, speech from the mask will sound fuzzy, muffled, or tinny (Fig. 14). The presence of speech distortion degrades voice biometrics performance. Therefore, equalization of speech from the mask must be one of the pre-processing functions.



(a) Speech uttered outside the mask (high-quality speech)



(b) Same phrase uttered inside the mask (fuzzy, muffled, and tinny)



(c) Clear photo

(d) Fuzzy photo

Fig. 14 — Comparison of clear speech (speech uttered in the free air) and distorted speech (speech uttered inside with the mask on). Use of distorted speech for voice biometrics is like using a fuzzy face for face-matching biometrics. We can remove speech distortion within ± 1 dB by the equalization technique presented here.

Restoring speech signals from the gas mask involves the following three steps (1) characterizing the frequency response of the mask; (2) developing an effective means for restoring speech from the mask; and (3) comparing the speech distortion before and after equalization. We discuss each step in the following.

Mask Characterization

The first step is to measure the frequency response of the mask when it is worn. The filter that eradicates speech distortion is the inverse of the mask frequency response. Note that we need to estimate the mask response at least once, sometime before the mask is actually being used. We, however, have to do it again if a new mask is to be used.

A commonly used method of measuring the frequency response of a network or transducer is to measure the output under the excitation of a sweeping sinusoidal signal at the input. Unfortunately, there is no acceptable way of injecting a sinusoidal signal into the microphone (or voicemitter) input when the mask is worn. Therefore, we must use an alternative method of measuring mask frequency response using actual spoken speech as a probe signal. This method is feasible because:

- The speech spectrum is broad over the frequency range of interest (0 to 8 kHz) if the spectrum is accumulated over several seconds.
- The spectral difference between the speech spoken with and without the mask approximates the frequency response of the mask.

The speech spectrum with the mask on in terms of the speech spectrum without the mask is

$$S_{inside}(\omega) = S_{outside}(\omega)M(\omega), \tag{5}$$

where $S_{inside}(\omega)$ is the spectrum of distorted speech originating from inside the mask, $S_{outside}(\omega)$ is the undistorted speech uttered outside the mask, and $M(\omega)$ is the mask frequency response. Figure 15 shows both speech spectra.

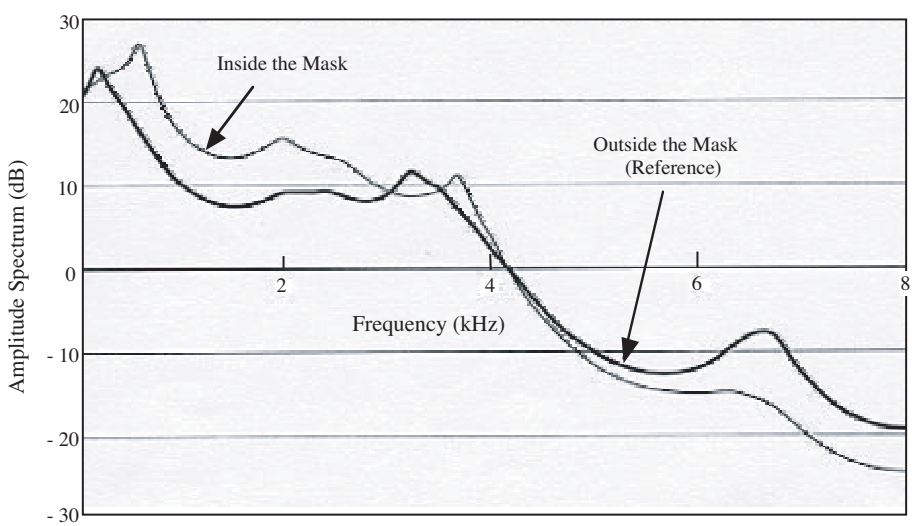


Fig. 15 — Spectra of the speech uttered outside the mask (undistorted) and inside the mask (distorted). The difference is the mask frequency response, as is shown in Fig. 16.

From Eq. (5), the mask frequency response is expressed as

$$M(\omega) = \frac{S_{inside}(\omega)}{S_{outside}(\omega)}. \tag{6}$$

The equalizer frequency response that removes mask distortion is the inverse of $M(\omega)$ expressed by Eq. (6). Thus, the equalizer frequency response is

$$E(\omega) = \frac{S_{outside}(\omega)}{S_{inside}(\omega)}. \tag{7}$$

Figure 16 shows the frequency responses of both the mask and the equalizer that removes mask distortion.

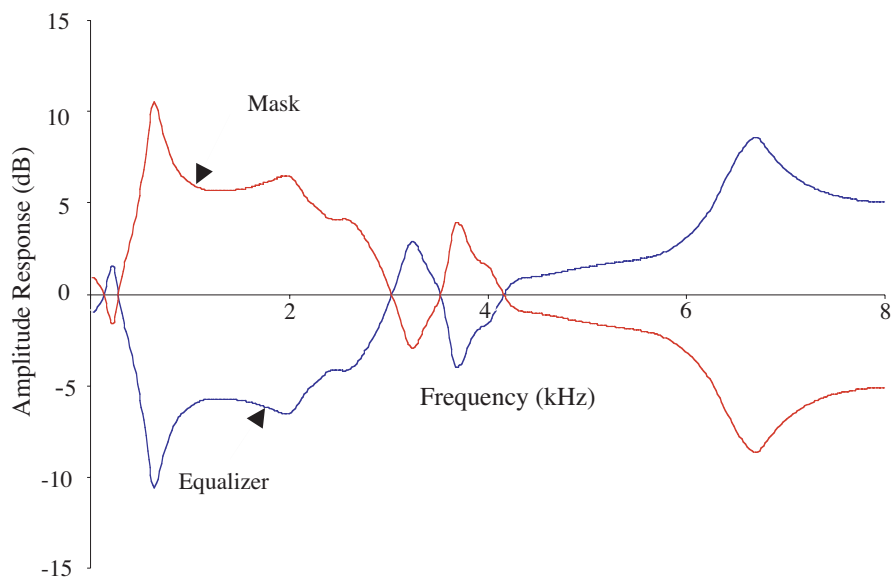
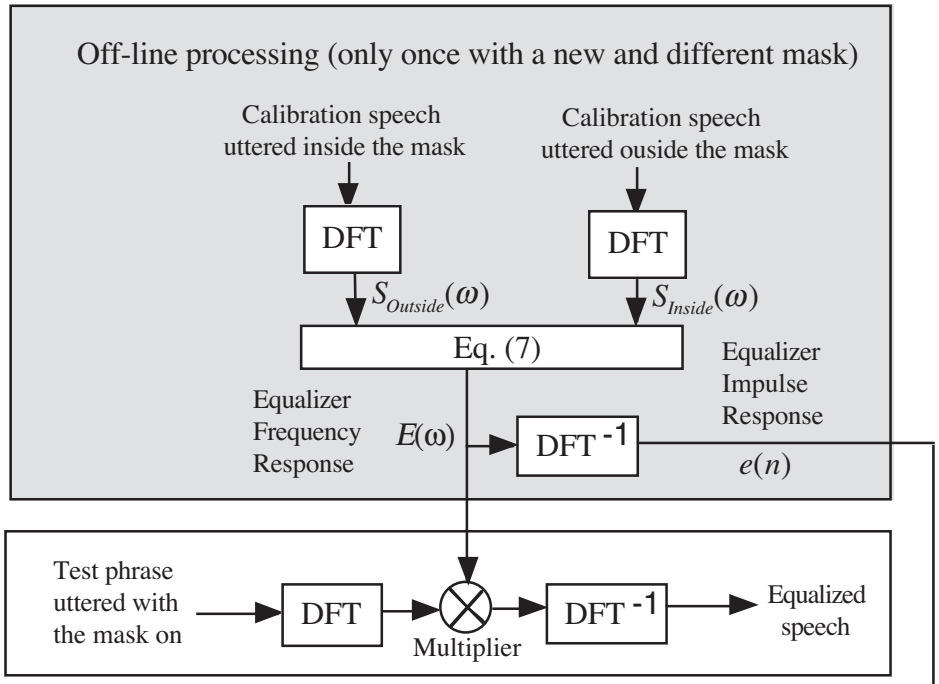


Fig. 16 — Frequency responses of the mask and the equalizer. These two frequency responses are the inverse of each other. It is interesting to note that the mask frequency response is a speaker-dependent voice biometrics feature because it is related to the speaker's facial contour for a given type of gas mask.

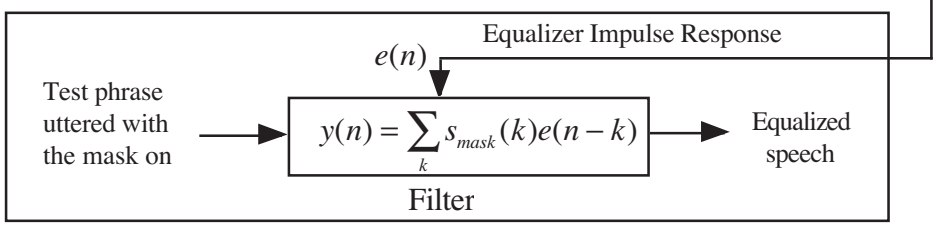
Three Approaches to Equalization

There are three different ways of obtaining the equalized speech output, depending on whether the equalization is performed in the frequency domain or in the time domain.

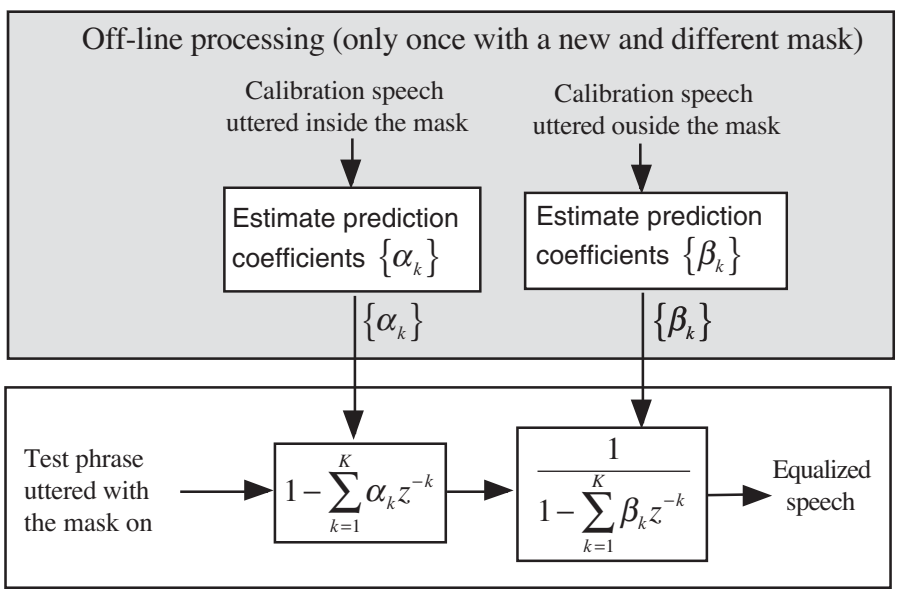
Approach #1: *Equalization in the frequency domain* — The speech waveform from the mask is Fourier transformed. Then the equalized speech spectrum is obtained by multiplying the mask speech spectrum by the equalizer frequency response expressed by Eq. (7). See Fig. 17(a). The equalized *spectrum* is then inverse Fourier transformed to generate the equalized speech *waveform*. Frames must be overlapped during the inverse Fourier transform. Otherwise, there will be audible clicks in the equalized speech.



(a) Equalization Method #1



(b) Equalization Method #2



(c) Equalization Method #3

Fig. 17 — Three approaches for equalizing distorted mask speech

Approach #2: *Frequency domain analysis and time domain synthesis (our preferred approach)* — The speech waveform is Fourier transformed as in Approach #1. The frequency response of the equalizer is inverse Fourier transformed to obtain the equalizer impulse response. See Fig. 17(b). Then the mask speech waveform is convolved with the mask impulse response to obtain the equalized speech output directly. An advantage of this approach is that frame overlapping is not needed.

Approach #3: *All-pole spectrum in place of Fourier transform (our alternate approach)* — The all-pole approximation to the speech spectrum has been used extensively by the Linear Predictive Coder (LPC) to encode speech [7]. In the framework of LPC, the z-transform of the speech originated from inside the mask is expressed by

$$S_{inside}(z) = \left. \frac{1}{1 - \sum_{k=1}^{K} \alpha_k z^{-k}} \right|_{z = j\omega\tau} , \tag{8}$$

where α_k is the kth prediction coefficients, K is the total number of prediction coefficients (usually $K = 10$), ω is frequency in radians/second, τ is the speech sampling time interval, and $j = \sqrt{-1}$. The method of computing prediction coefficients α_k ($k = 1,2,K$) is well established. Similarly, the z-transform of the speech uttered outside the mask (undistorted speech) can be expressed by

$$S_{outside}(z) = \left. \frac{1}{1 - \sum_{k=1}^{K} \beta_k z^{-k}} \right|_{z = j\omega\tau} , \tag{9}$$

where β_k is the prediction coefficient representing the undistorted speech. Therefore, the z-transform of the equalizer is

$$E(z) = \left. \frac{1 - \sum_{k=1}^{K} \alpha_k z^{-k}}{1 - \sum_{k=1}^{K} \beta_k z^{-k}} \right|_{z = j\omega\tau} . \tag{10}$$

The equalizer is a combination of a recursive filter (defined by the denominator) and a feed-forward filter (defined by the numerator) (Fig. 17(c)). Computationally, Approach #3 is the most efficient, and the quality of equalization is as good as the others. Any of these approaches improves mask speech significantly.

Speech Spectral Errors Before and After Equalization

The equalization process significantly removes speech distortion. As indicated in Fig. 18, the peak-to-peak speech spectral distortion prior to the equalization is approximately ±10 dB; this reduces to ±1 dB after equalization.

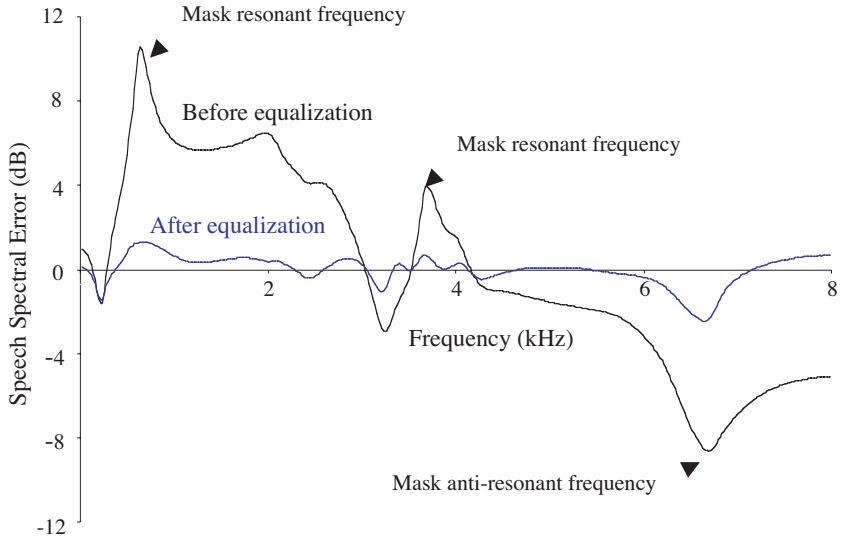


Fig. 18 - Gas mask speech distortion before and after equalization. Speech spectral distortion prior to equalization is ±10 dB, which reduces to ±1 dB after equalization. A frequency response error of ±1 dB is normally associated with communication-grade microphones.

VOICE FEATURE EXTRACTION

The voice biometrics process is significantly different from other biometrics processes in the sense that the human voice does not have stable biometrics features. We have to carefully design a voice biometrics process to overcome this inherent limitation.

The speech waveform is related to both the speech sounds (phonemes) and the speaker's voice timbre; these two properties have to be exploited for accuracy. We need to exploit the speech-related feature because two persons having similar voice timbre would pass speaker verification. Furthermore, we need to exploit the speaker's timbre so that imposters are prevented from passing speaker verification by uttering someone else's test phrase.

The use of only one voice feature is not satisfactory because many different voice characteristics must be discriminated from others. Therefore, we use five different voice features so that the deficiency of one voice feature will be made up by others. These features are:

1. *Speech spectral histogram* — This feature is in the form of 10-dimentional trajectories. They are obtained from the speech waveform through LPC analysis.

2. *Vocal tract area histogram* — This feature is also 10-dimensional trajectories and is obtained from the reflection coefficients generated by LPC analysis of the speech waveform. The initial speech parameters are the same as for the speech spectral histogram and the vocal tract area trajectories. The different transforms used in these features reveal different aspects of the voice.

3. *Speech loudness histogram* — This feature is a 1-dimensional trajectory. It is effective for detecting imposters uttering someone else's test phrase.

4. *Average pitch frequency* — This is a single-valued parameter. Although people can change pitch intentionally (this is what we call singing), the average value pitch frequency of normal conversation is relatively stable. This voice feature is effective for separating persons having large pitch differences.

5. *Speech energy distribution* — This is also a single-valued parameter. The speech energies contained in the low band (0-4 kHz) and the upper band (4-8 kHz) are compared. This feature is related primarily to the speaker's voice timbre.

Speech Spectral Histogram

The speech spectrum (unless stated otherwise, the speech spectrum is referred to as the speech spectral envelope — the former is a common misnomer) is widely used in all speech-related applications (speech encoding, speech recognition, speech alteration, etc.). The speech spectrum is conveniently obtained from the reflection coefficients through the LPC.

The speech spectrum is related to what is spoken and is also related to individual voice characteristics. The LPC analysis has already been discussed in connection with mask-speech equalization in Eqs. (8) and (9). The resultant LPC spectrum closely approximates the speech spectral envelope (Fig. 19). The histogram of the speech spectral envelope shown in Fig. 20 is one of the most important voice biometrics features.

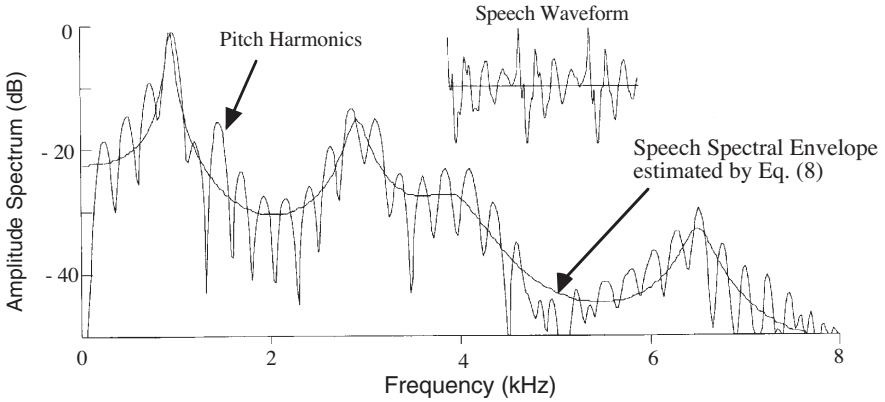


Fig. 19 — Speech spectrum observed in one frame. The speech spectrum has two components: one is the speech spectral envelope, and the other is pitch harmonics inscribed under the speech spectral envelope.

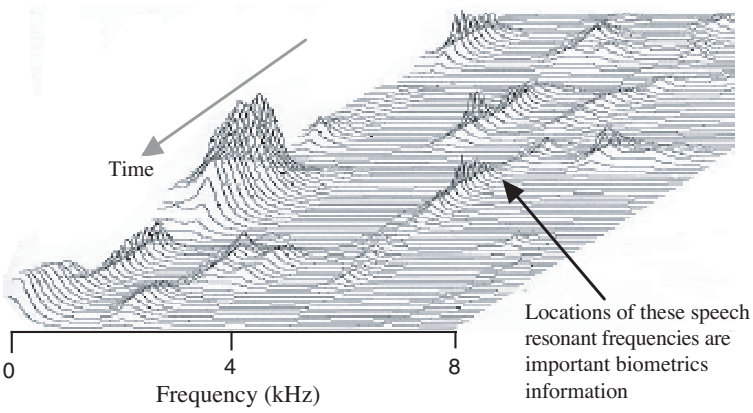


Fig. 20 — Spectral trajectories of speech. As will be shown in Fig. 21, resonant frequencies in the low-frequency region are more related to speech, whereas those in the high-frequency region are more related to the speaker's voice characteristics. As mentioned in the Background section, the speech spectral trajectory is one of the most widely used voice features.

To show the speaker characteristics exhibited in the speech spectrum, we plotted the speech spectra of the vowel /a/ in "art" spoken by three different people (Fig. 21). We note the following:

- The only commonality in the spectrum is the location of the first resonant frequency (see the arrow in Fig. 21)
- Even though everyone utters the same vowel /a/, higher resonant frequencies are located at different frequencies. It is not easy to alter only higher resonant frequencies while the first resonant frequency remains unchanged. The higher resonant frequencies are predetermined by the physical characteristics of the speaker. Hence, they are speaker-dependent.
- Therefore, the difference between two speech spectra over the entire test phase is predominantly due to physical differences between the speakers, making it a useful biometrics parameter. This difference is one of the five factors used in calculating our speaker verification score.

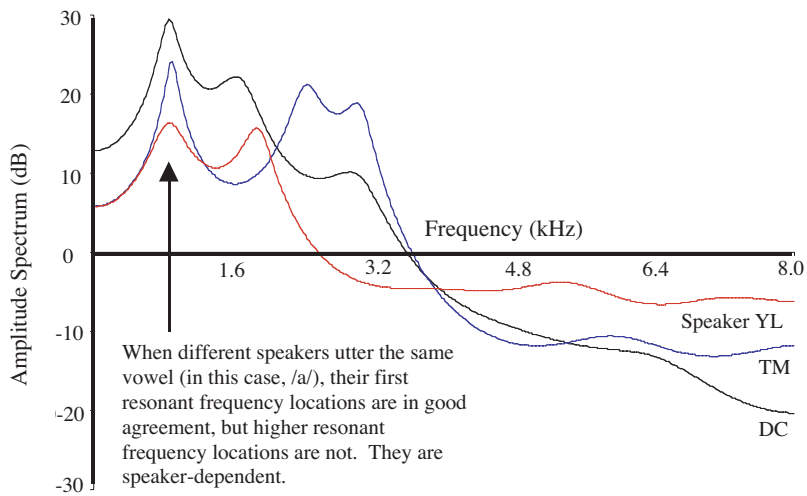


Fig. 21— Speech spectra of /a/ in "art" spoken by three different persons. The only commonality in the resultant speech spectra is the location of the first resonant frequency. The remaining resonant frequencies depend on the individual speaker's voice characteristics.

Vocal Tract Area Histogram

The human vocal tract can be approximated by a series of concentric tubes (Fig. 22), and the cross-sectional area of each section can be estimated by the reflection coefficients obtained through the linear prediction analysis of the speech waveform. The set of prediction coefficients from either Eq. (8) or Eq. (9) can be converted to a set of reflection coefficients through transformation [27]. This transform is reversible, and no information will be lost:

$$\alpha_{j|n+1} = \alpha_{j|n} - RC_{n+1}\alpha_{n+1-j|n} \qquad j = 1,2,\ldots,n, \tag{11}$$

with

$$\alpha_{n+1|n+1} = RC_{n+1}, \tag{12}$$

where $\alpha_{j|n+1}$ is the jth prediction coefficient for the $(n + 1)$ iteration, and RC_{n+1} is the $(n + 1)$ reflection coefficient.

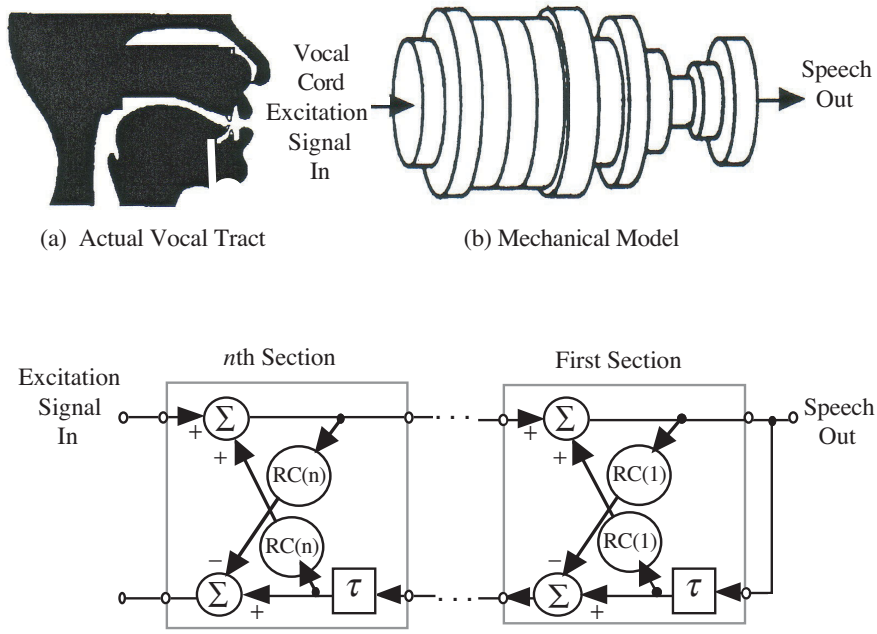


Fig. 22 — Vocal tract models. The mechanical model of the human vocal tract is a series of concentric tubes The electric model of the human vocal tract is a cascaded lattice filter as shown in Fig. 22(c), which is a realization of Eq. (8).

In physics, the area ratio of two adjacent sections of a concentric tube is expressed in terms of reflection coefficients,

$$\frac{A_j}{A_{j+1}} = \frac{1 - RC_j}{1 + RC_j}. \tag{13}$$

From Eq. (13), cross-sectional areas are successively computed by

$$A_i = \left[\frac{1 - RC_i}{1 + RC_i}\right] A_{i+1} \qquad i = 1,2,\ldots,I, \tag{14}$$

with the terminal section of $A_I = 1$.

Because the vocal tract areas are computed from reflection coefficients, there is a kinship between the LPC spectrum and the vocal tract area. Due to the different transformations of reflection coefficients, the transformed outputs also show different aspects of the speech signal.

To show speaker-dependency in the vocal tract area, as we demonstrated with the speech spectrum previously, we plotted three different vocal tract areas generated by three different persons uttering the vowel /a/ in "art." Similar to Fig. 21 with the LPC spectrum, the vocal tract area shows the commonality that the second section has the largest cross-sectional area (Fig. 23). Other cross-sectional areas have no similarities. Therefore, if the difference (or error) between two vocal tract areas is computed over the entire test phrase, the sum of the difference will be predominantly speaker-dependent.

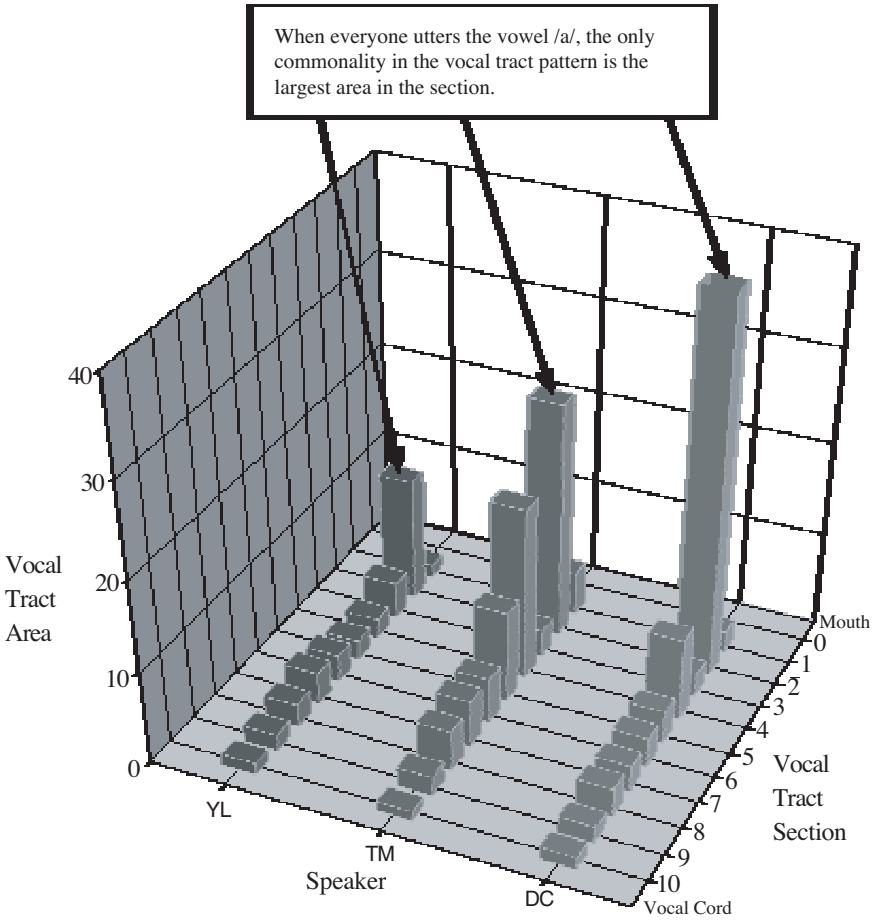


Fig. 23 — Vocal tract cross-sectional area when /a/ in "art" is spoken. As in the case of speech spectrum shown in Fig. 21, when people speak the same vowel, commonality is evident. In this case, the second cross-sectional areas are the largest. However, other cross-sectional areas are different, which is a phemomenon that will be exploited in voice biometrics.

Speech Loudness Histogram

The trajectory of speech loudness is useful for detecting an imposter uttering someone else's test phrase because the trajectory manifests the speaking rhythm. Figure 24 is an example of the speech loudness histogram. Because this trajectory clearly shows each vowel location, it is used later for time-aligning the voice feature trajectories to be compared, which is a critical step for speaker verification.

Average Pitch Frequency

The average pitch frequency over the entire test phrase is another voice feature used in voice biometrics. If we average pitch frequencies over the test phrase, any semantic context in the speech can be completely eliminated. As a result, the average pitch frequency is dependent on the speaker, and this is rather stable unless the speaker intentionally wants to change it.

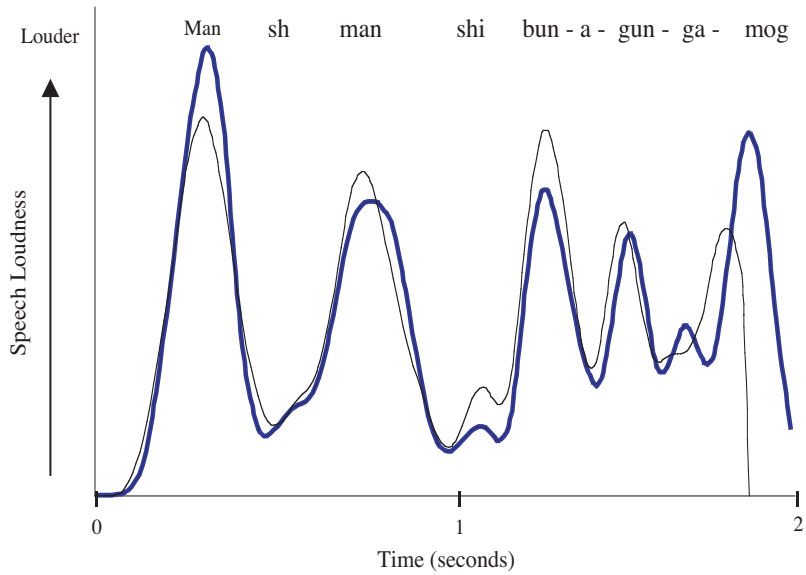


Fig. 24 — Loudness contours of speech uttered two weeks apart (one is in black, and the other is in blue). These well-overlapped trajectories signified that the same test phrase is spoken with virtually identical speaking rhythm (except near the end). This is one of the indications that the test phase is most likely uttered by the same speaker.

Pitch estimation has been greatly improved recently because of the widespread use of voice encoders that require pitch information. The Average Magnitude Difference Function (AMDF), defined in Eq. (15), is often used to estimate raw pitch:

$$AMDF(\tau) = \sum_j \left| e(j) - e(j+\tau) \right|, \tag{15}$$

where $e(j)$ is the jth speech sample, and τ is a delay. Figure 25 shows the AMDF profile. The most probable pitch period is the delay that corresponds to the minimum of the AMDF.

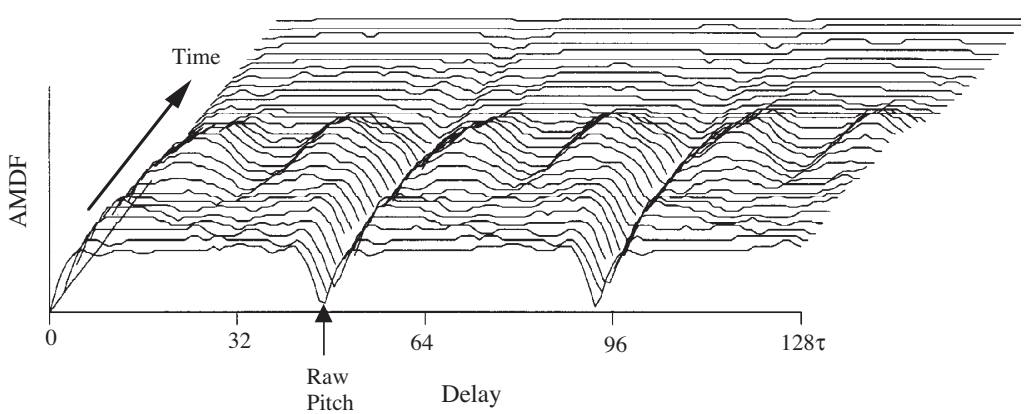


Fig. 25 — Histogram of the Average Magnitude Difference Function (AMDF). Since vowel waveforms are repetitive at the pitch rate, the AMDF shows nulls at multiples of the pitch period. The first null is the raw pitch period of that frame. The raw pitch obtained from a vowel waveform must be tracked (dead-reckoning) through a consonant waveform, which is aperiodic. Hence there is no pitch value.

Figure 26 shows pitch trajectories of four speakers uttering their voice biometrics test phrases. Pitch contour varies in accordance with context. For example, the test phrase uttered by TM is "Today is a sunny day," in which pitch frequency is higher for "today" and "sunny."

Note that there is no pitch for fricatives and other unvoiced speech because their waveforms are not repetitive in time. Hence, the pitch is difficult to track. Consequently, we used the average pitch frequency of the entire phrase (a single value) rather than the pitch trajectory.

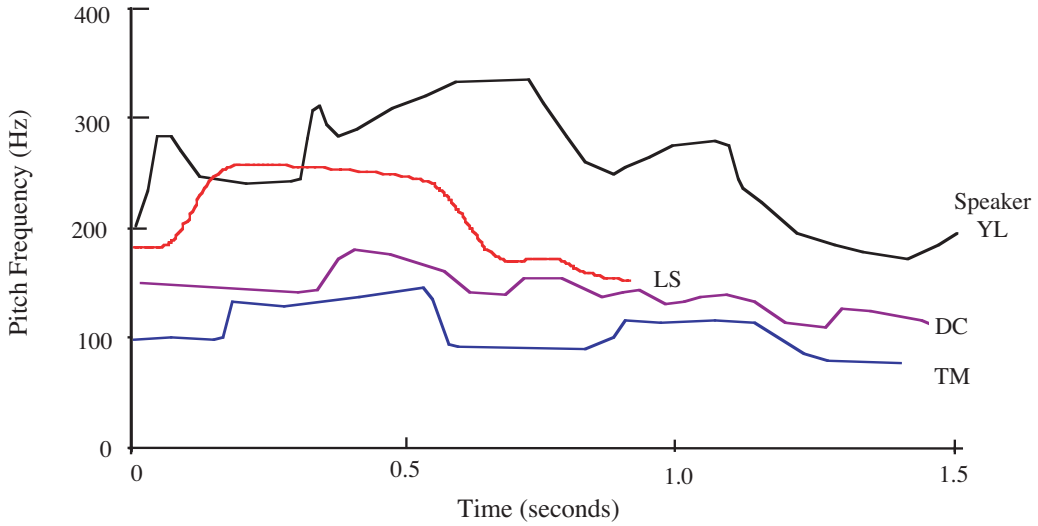


Fig. 26 — Pitch frequency. Pitch frequency is another voice feature for speaker verification. These four speakers have their own pitch ranges, which makes average pitch frequency useful for speaker verification.

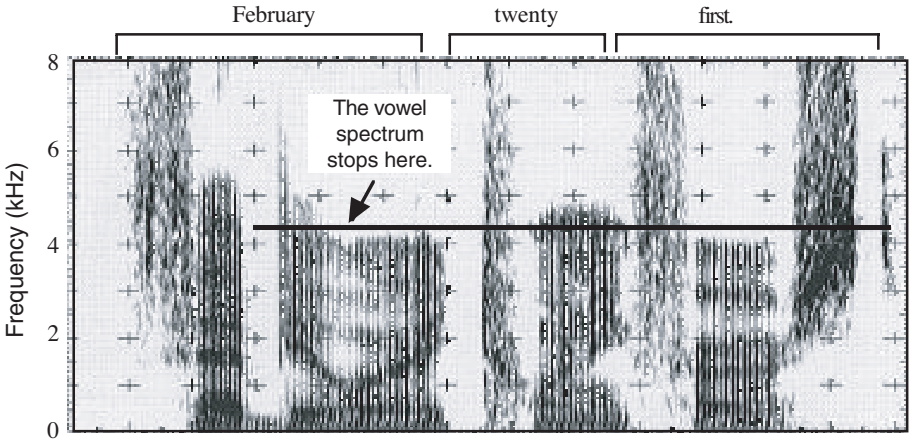
Speech Energy Distribution

For generating vowels, the glottis opens and closes periodically at the pitch rate. If the glottis closes instantly, it produces a wideband excitation signal. Likewise, the speech spectrum is broadband. On the other hand, if the glottis closes slowly (because of age, injury, smoking, etc.), then it has a narrowband excitation signal. The resultant speech spectrum is likewise narrowband.

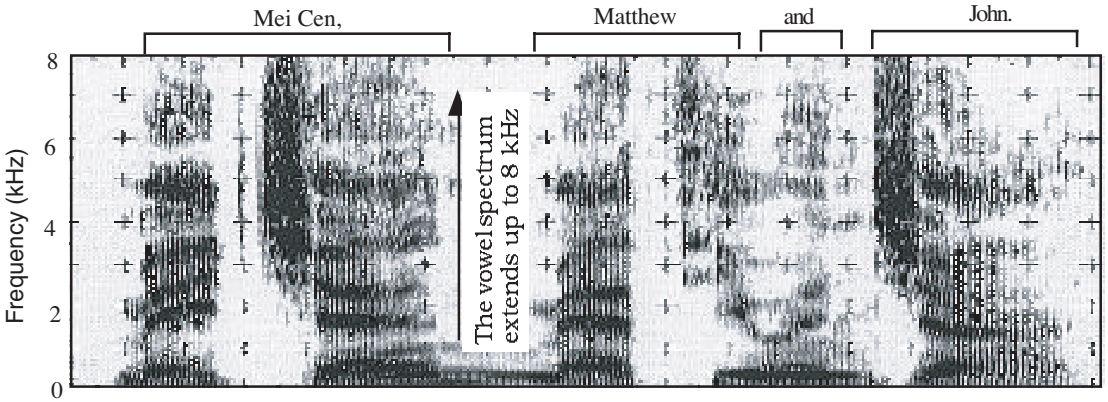
Figure 27 shows two examples of voice spectra. In Fig. 27(a), the vowel spectra are virtually cut-off at around 4 kHz, whereas Fig. 27(b) shows vowel spectra extending up to near 8 kHz. Therefore, the speech energy distribution is highly speaker-dependent. We use the ratio (a single value) of the average energies above and below 4 kHz over the entire test phrase.

To exploit the energy distribution for the voice biometrics process, speech energies in the low band (0-4 kHz) and in the high band (4-8 kHz) must be obtained. We average each speech energy over the entire test phrase. Hence, speech dependency is completely eliminated.

We use only the energies of vowels because excitation signal stems from the glottis in those sounds, and the vocal tract of the individual speaker is involved. Hence, the energy ratio is speaker-dependent (Fig. 28). The reason for excluding nonvowel energies is that their excitations originate from a constricted area somewhere between the teeth or the teeth and lips; they are spectrally similar, indicating that they are not speaker-dependent.

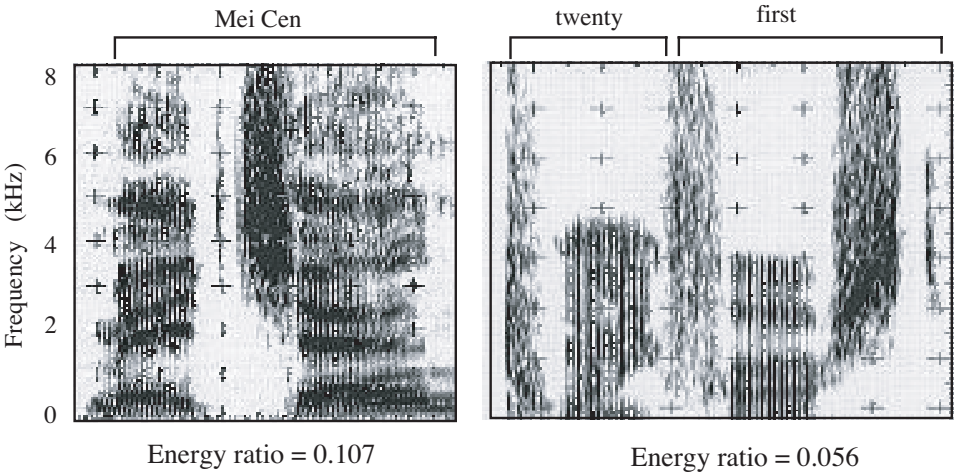


(a) Voice without high frequencies for vowels



(b) Voice with strong high frequencies for vowels

Fig. 27 — Spectrograms of a narrowband voice and wideband voice. The spectral distribution is the most explicit voice characteristic related to speakers.



(a) Voice without high frequencies for vowels (narrowband)

(b) Voice with strong high frequencies for vowels (wideband)

Fig. 28 — Highband-to-lowband speech energy ratio. A larger ratio means the presence of stronger high frequencies in vowels. This ratio is one of the five voice features we use in our voice biometrics systems.

SPEAKER VERIFICATION

Speaker verification is a process of comparing the voice features derived from two different speech waveforms; one is the stored template speech and the other is the live test speech (Fig. 29). The speaker verification device makes a decision as to whether the speaker under test is the same person as the speaker in the template. In addition, the speaker verification device recommends that the template be updated if the error is close to the accept-reject threshold level.

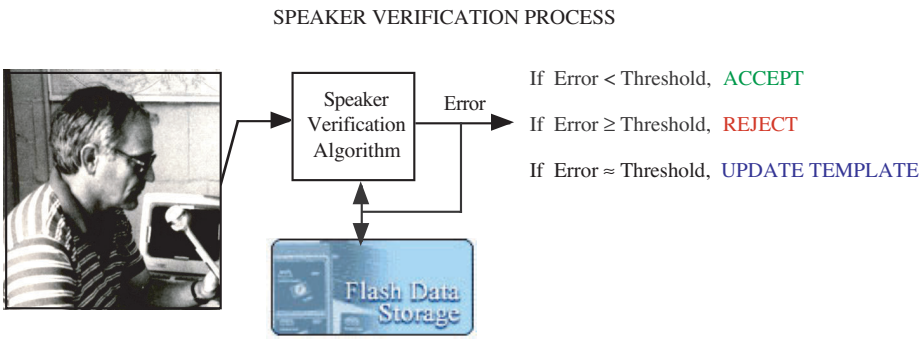


Fig. 29 — NRL speaker verification system. Ideally, the speaker's template should be stored at the speaker verification device to be used. For unseen circumstances, the voice biometrics device may not have a particular speaker template. In the NRL system, each speaker may feed his (or her) voice template into the device prior to speaker verification.

We have five voice features to compare; there are two steps in the comparison process:

- *Compute individual voice feature error* — Each of the five voice features is compared separately. If a voice feature is a single number (such as average pitch frequency), the error is the difference of the two. In this case, the problem of time alignment does not exist. On the other hand, if a voice feature is in the form of trajectory (such as the speech spectrum), then time alignment becomes a major problem. We present two approaches for solving this problem.
- *Combine individual errors to generate a total voice error* — The five individual voice feature errors are combined into a total error. They are combined in such way that when a speaker tests against himself (or herself), the total error will be a fixed value of 1.0. The fixed self-test error makes it easier to establish the accept-or-reject threshold level.

Generally, comparing two sets of voice features is not a simple problem because two sets of voice features may not be aligned properly. Figure 30 is an example of misaligned voice features. If misalignment is not eradicated, the correct speaker could be rejected by the speaker verification device. Therefore, we need a technique to align two voice features.

Time Alignment (Time Warping)

Before computing the error between individual voice features, they must be time-aligned. This is done by expanding or compressing the time axis for voice feature #2 according to whether it is leading or lagging behind voice feature #1. Consider a Cartesian coordinate system in which the abscissa is the time axis for voice feature #1 (reference) and the ordinate is the time axis for voice feature #2. The resultant time-mapping characteristic is the coordinate system, which we call the time transfer characteristic (TTC).

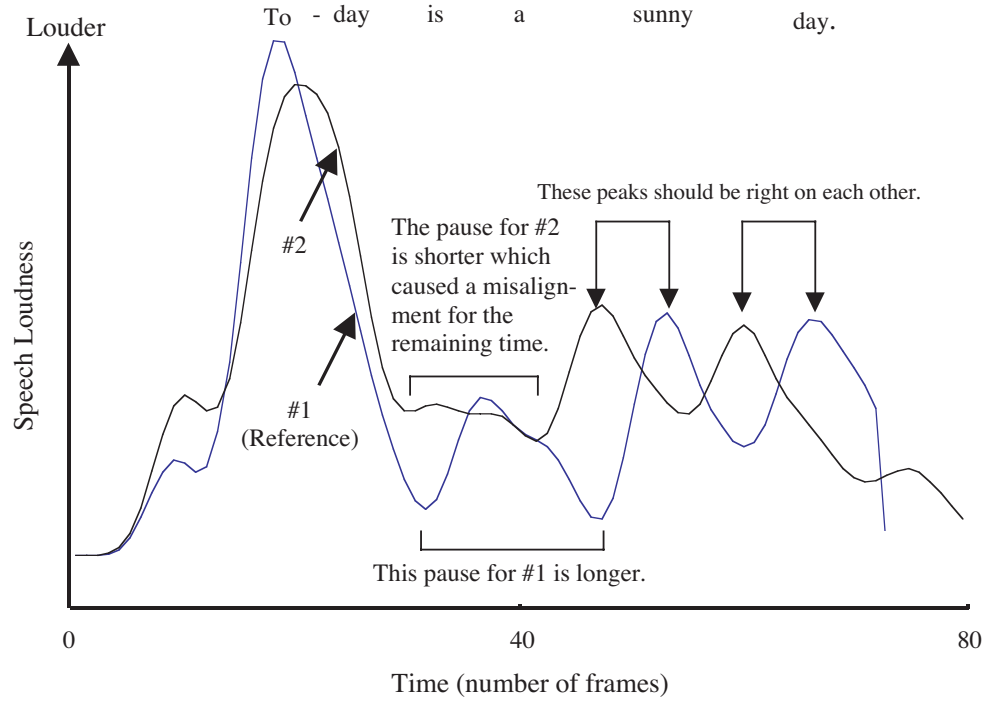


Fig. 30 — Example of misaligned voice features. This figure shows the loudness contours of the same phrase uttered by the same speaker only two weeks apart. The misalignment is caused by a shorter pause after "Today is" when the speaker uttered it for the second time (black line). The mutual separation is best determined at the local peaks. Peak location of voice features #1 and #2 are, respectively, (22,24), (57,49), and (71,63). These three time coordinates are critical for restoring time alignment.

If the two voice features are perfectly aligned, both time axes have an equal metric, namely, one time division equals one frame of 20 ms. Therefore, the TTC for the ideal case is a 45º line. If the voice features are not aligned, the TTC will be a curve. In this case, the TTC is constructed by measuring the mutual separation between the two voice features at each frame epoch.

The mutual separation between two voice features (see Figs. 24 or 30) is not easy to measure in the presence of an amplitude mismatch. The only locations where the mutual separation is certain are at the local peaks. Referring to Fig. 30, the time axes corresponding to the three local peaks are

$$\begin{bmatrix} (t_1,t_2)_1 \\ (t_1,t_2)_2 \\ (t_1,t_2)_3 \end{bmatrix} = \begin{bmatrix} (22,24) \\ (57,49) \\ (71,63) \end{bmatrix}. \tag{16}$$

These coordinate points are the three most reliable points on the TTC. The rest of the points are derived by the interpolation (Fig. 31).

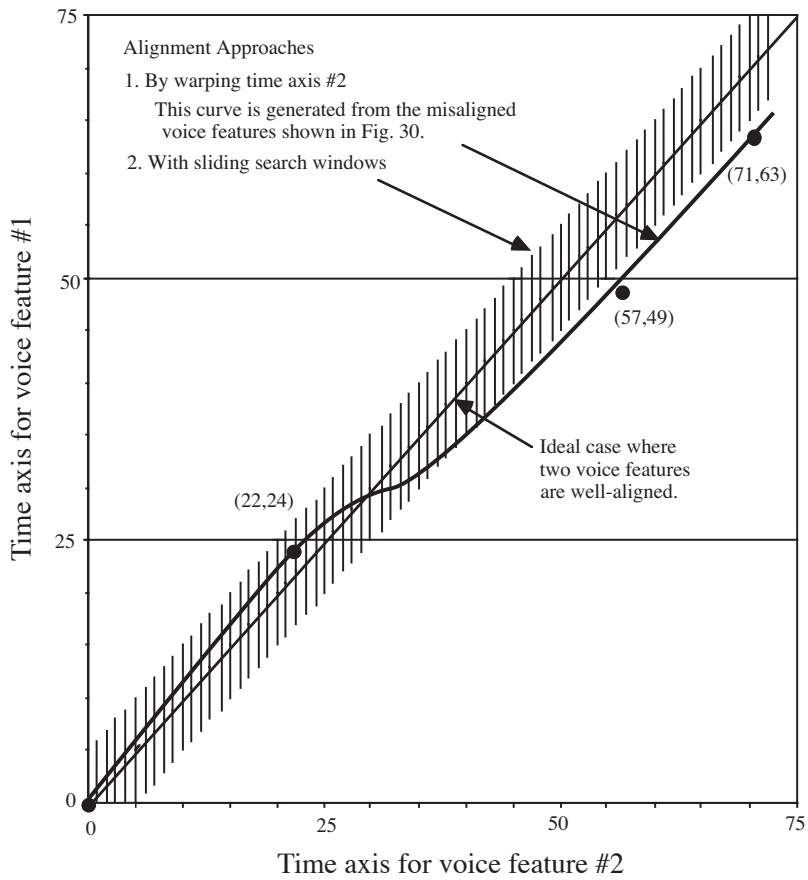


Fig. 31 — TTCs for two different approaches for comparing voice features. In Approach #1, the TTC is the interpolated line based on times corresponding to the speech energy peaks (see Fig. 30). In Approach #2, we do not warp the time axis for feature #2. At each operating point along the ideal TTC characteristic (the 45º line), the best-aligned condition is searched by sliding the window as much as 10 frames.

Individual Voice Feature Errors

To facilitate the discussion, we begin with the ideal case in which voice features are well aligned. We note that approximately 20% of speakers uttering their test phrases can generate nearly time-aligned voice features (see Fig. 24). In this case, the voice feature error can be obtained by direct subtraction (i.e., without warping the time axis):

$$\varepsilon = \sum_{n=1}^{N} \sum_{k=1}^{K} \left| \lambda_1(n,k) - \lambda_2(n,k) \right| \qquad \text{(ideal case)}, \tag{17}$$

where $\lambda_1(n,k)$ and $\lambda_2(n,k)$ are the voice features in the template and the sample under test, respectively; k is the time index, and n is the number of variables in the voice feature.

Generally, however, voice feature trajectories will not be aligned properly (see Fig. 30). There are two approaches to eradicate time misalignment.

Approach #1: *With time warping* — The time-warping technique mentioned previously can be used. A weakness of this approach is that the construction of TTC is rather difficult if voice features have a mutual delay but also have amplitude mismatch. The only location where mutual delay can be estimated reliably is at the local peaks. The rest of the TTC must be interpolated from the time coordinates at the local peaks. Let the resultant TTC be denoted by

$$y_t = f(x_t). \tag{18}$$

Then, the individual voice feature error is obtained from

$$\varepsilon = \sum_{n=1}^{N} \sum_{k=1}^{K} \left| \lambda_1(n,k) - \lambda_2(n, f(k)) \right|. \tag{19}$$

Approach #2: *With a sliding window (our approach)* — This approach does not warp one of the time axes. Rather, it uses a sliding window, the ideal TTC (a 45º line), to accommodate timing misalignment between voice features. The window moves vertically from –5 frames to +5 frames (Fig. 31). The best error is selected from the minimum of the 11 errors generated at each frame epoch,

$$\varepsilon = \sum_{n=1}^{N} \min_{j} \left[\sum_{k=1}^{K} \left| \lambda_1(n,k) - \sum_{j=-5}^{5} \lambda_2(n, k-j) \right| \right]. \tag{20}$$

The advantage of this approach is that it gives a more accurate feature error because it looks for an error at every frame. However, this approach requires more computations.

Total Voice Feature Error and Weighting Factors

To perform voice biometrics, all five individual voice feature errors must be combined into the total error. Thus,

$$\varepsilon_T = \frac{1}{M} \sum_{m=1}^{M} w(m) \varepsilon_m , \tag{21}$$

where ε_T is the total error, ε_m is the individual error ($m = 1,2, \ldots, M$). In our case, M is 5, and $w(m)$ is the mth weighting factor. The weighting factors are important because physically diverse individual errors are being combined into a total error of a single dimension.

We combine the individual errors in such a way that the total error becomes a fixed value when any speaker tests against himself (or herself). Without this property in the speaker verification, it is difficult to define a fixed threshold for acceptance or rejection applicable to anyone. Therefore, each weighting factor must be equal to the reciprocal of the respective individual voice feature error. Thus, the mth weighting factor (where the total number of m is 5 because we have 5 voice features) is

$$w(m) = \frac{1}{\varepsilon_m} \qquad m = 1,2,3,\ldots,5, \tag{22}$$

which means the weighting factor is inversely proportional to the respective voice feature error. Thus, if a speaker speaks his (or her) phrase consistently, there is less chance for imposters to be verified incorrectly. Most importantly, the self-test score for any speaker will be around 1.0. Because of this property, the

threshold level for acceptance or rejection can be set to a fixed value for all speakers (in our case, it is 2.0.) The verification score also indicates the confidence level of acceptance, which indicates when the template should be updated.

Self-Test Score

The self-test score generated by the speech samples uttered for enrollment is 1.0. The self-test scores after enrollment, however, will be clustered around unity. The scores will be scattered because we are not able to utter the same phrase similarly at subsequent verification sessions. Figure 32 shows typical self-test scores with a standard deviation.

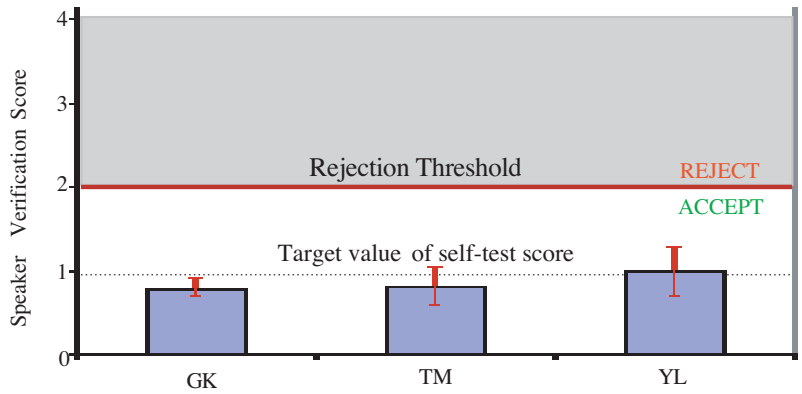


Fig. 32 — Self-test scores. The NRL voice biometric system is calibrated so that the self-test score is centered around 1.0, independent of speakers. This figure shows self-test scores of three individuals after enrollement. The average score is indicated by a bar, and the standard deviation is indicated by a short line segment around the average. The fixed level of self-test score makes it possible to assign a fixed level of the acceptance-or-rejection threshold, which is 2.0.

Inter-Speaker Scores

Figure 33 shows the inter-speaker verification score in which speaker LS in the template is compared with LS under verification. LS scored around 1.0; therefore, LS is verified as being a correct person. Figure 33 also plots verification scores between LS and the other speakers. As expected, the inter-speaker scores are far above the acceptance-or-rejection threshold. All except LS are rejected as imposters.

Template Update

In other biometrics, a template update is not needed because, for example, fingerprint patterns or iris patterns are difficult to change. In voice biometrics, however, template update is mandatory because voice features can easily change. Fortunately, the NRL voice biometrics system provides an indication as to when the template must be updated (Fig. 33).

If the verification score approaches 1.0, the confidence level of acceptance is high, but as the score approaches 2.0, the confidence level of acceptance is low. If the score is somewhat below 2.0 (i.e., between 1.7 and 2.0), then the present speech waveform should be included in the template because the speech waveform in the template is becoming dated.

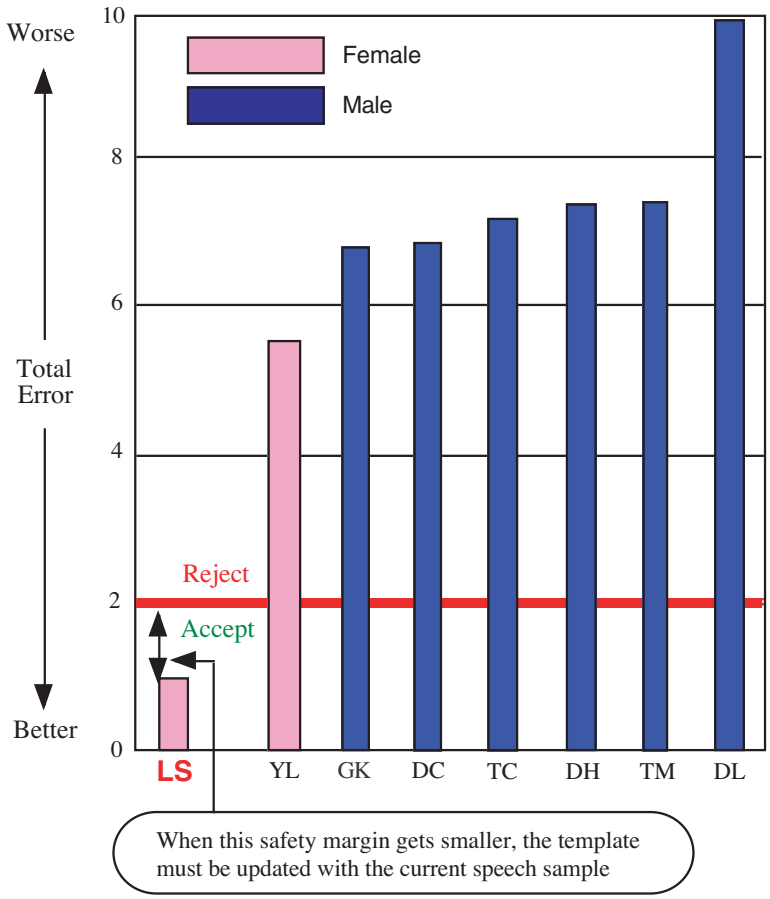


Fig. 33— Example of inter-speaker verification scores. Because each speaker has chosen his (or her) own test phrase, the verification score between two different speakers is very large. It is almost impossible to have a person erroneously accepted when he (or she) utters his (or her) own chosen phase.

Test Scores with Tape-Recorded Input Speech

One of the frequently asked questions with respect to voice biometrics is "what happens if someone uses tape-recorded speech?" This is a valid question for the NRL voice biometrics system because the test phrase is pre-selected by the speaker, and the speaker will use it at every speaker verification session. Therefore, imposters could record someone saying his (or her) test phrase at a speaker verification session.

Accordingly, we tested our voice biometrics system by five speakers using their chosen test phrase recorded on the tape. As noted in Fig. 34, they are all rejected by a good margin. The tape-recorded speech is different from the live speech, even if the best-quality digital audio tape (DAT) is used for recording,

Test Scores of Imposters

Figure 35 shows speaker verification scores of three speakers who imitated someone else's test phrase. They all failed. The choice of test phrase and the imposter's ability to replicate it has a large influence on the test scores.

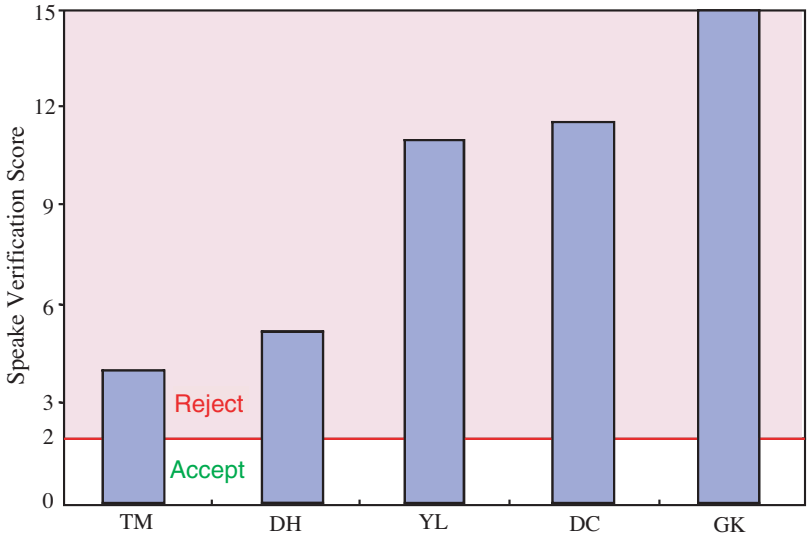
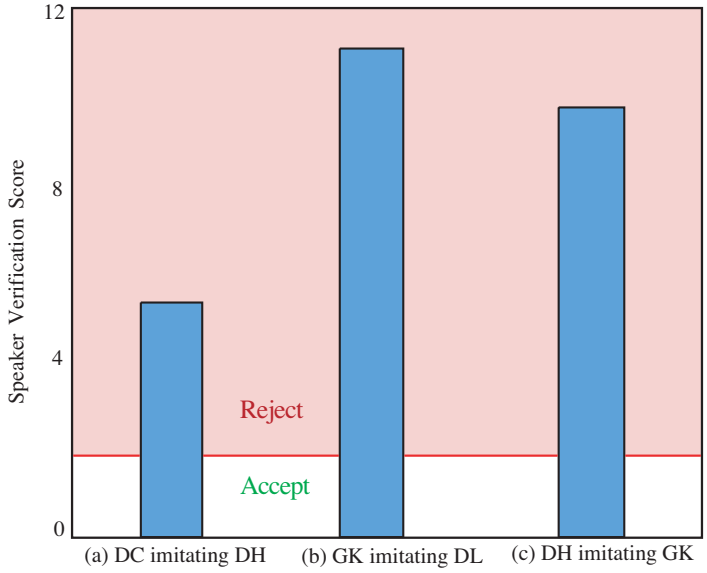


Fig. 34 — Speaker verification score with tape-recorded voices. The voice biometric process has a much better capability to analyze voice timbre than even humans through their own ears. Our voice biometrics devices will not be deceived by tape-recorded voices.



Case (a): The test phrase by DH is "Mei Cen, Matthew and John," which is imitated reasonably by DC because the text is in English, but the score (error) is far above the acceptance level.

Case (b): The test phrase by DL is "man sh man shi buna gungamog," which is hard to pronounce by other persons. The imposter is rejected by a good margin.

Case (c): The test phrase by GK is in a foreign language, which is difficult to imitate. The score is well above the acceptance level.

Fig. 35 — Test scores with imposters. All three imposters did not pass the verification test even though each practiced someone else's test phrases. It can be concluded that the choice of an unfamiliar phrase is an effective means for rejecting imposters.

Test Scores with Gas Mask

When the mask is worn, the speech waveform is distorted because of the resonant frequencies generated by the gas mask. On the other hand, the speech waveform stored in the template is free from similar distortion. Hence, when the speaker verification is performed with the mask on, the score is expectedly low, as indicated by Fig. 36.

With the equalization method illustrated in Fig. 18, however, speaker verification performance is significantly improved. As shown in Fig. 19, all speakers correctly participated in this experiment.

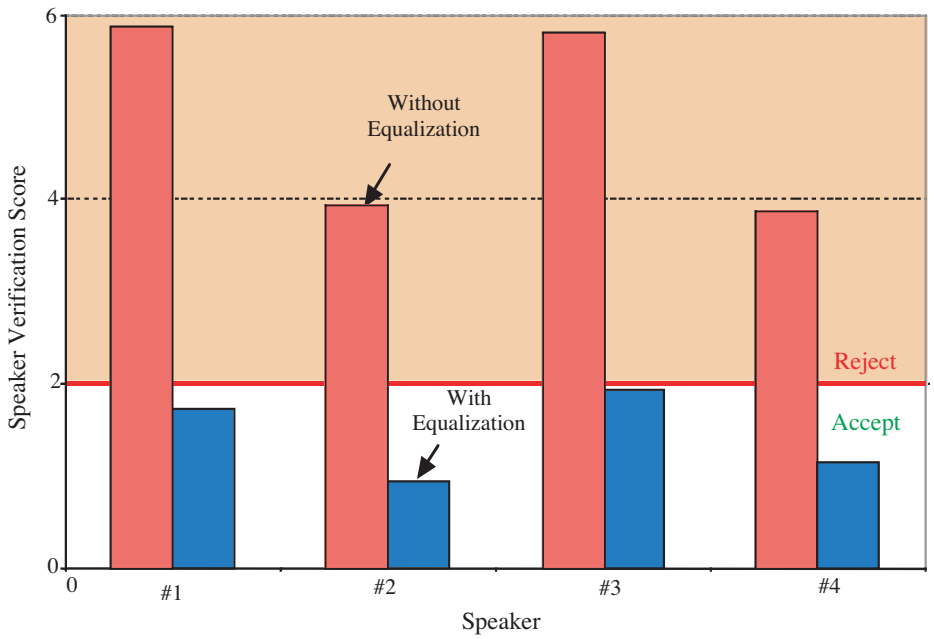


Fig. 36 — Test scores when the gas mask worn with and without equalization process discussed in Fig. 18. As shown in this figure, this equalization process enables correct speaker verification.

CONCLUSIONS

In this report, we have described a voice biometrics system developed at NRL for controlling access to classified information, information systems, facilities, and weapon systems.

The NRL voice biometrics system was designed to improve reliability and user-friendliness. To achieve these goals, we have developed the following steps:

- *Select a test phrase by each speaker* — Each speaker selects his (or her) own test phrase that can be uttered consistently and uniquely, while at the same time making it difficult to impersonate.

- *Carry own speech template* — Each test phrase is recorded, encrypted, and stored in a flash memory card. Each speaker carries this in case the voice biometrics device does not have the speaker's template; this situation may be encountered in battlefield situations.

- *Pre-process the speech waveform for normalization* — People cannot repeat the same phrase identically. The speech may be degraded (for example, the speaker wears a gas mask). Therefore, the biometrics device must be designed to normalize, equalize, and condition the speech waveform prior to the voice biometrics process.

- *Optimize the voice biometrics performance by selecting and utilizing the correct combination of speaker–related parameters* — Use as many different voice features as practical because the speech waveform is dependent on both speech and the speaker. These parameters include: (1) speech spectral histograms, (2) vocal tract area histograms, (3) speech loudness histograms, (4) average pitch frequency, and (5) speech spectral energy distribution.

- *Calibrate the self-test score to cluster around a fixed value* — We developed a method of calibrating the verification score so that when the speaker is tested against himself (or herself), the verification score is a fixed value regardless of the speaker. Hence, the magnitude of the score can indicate the confidence level of verification that leads to three possible decisions: (1) accept, (2) reject, and (3) accept, but the template must be updated.

In conclusion, we have made significant progress for making our voice usable for our password.

ACKNOWLEDGMENTS

This project was jointly sponsored by NRL and by SPAWAR. The authors thank Dr. John McLean, Superintendent, and Mr. Dale Long, Associate Superintendent, of the Information Technology Division, at NRL.

The authors express their appreciation to Vanessa Hallihan, Program Manager of Navy Information Security Office (SPAWAR PMW-161); Mr. Earle Kirkey, Director of the same organization, who oversees the Navy biometrics programs, and Dr. Jim Davies of the same organization who has been the project manager of our biometrics effort. The authors also express their gratitude to Mr. Dave Guerrino of the same organization who has been the Navy representative to the DoD Biometrics Management Office.

The authors appreciate Ms. Stephanie Everett of the NRL Information Technology Division who read the draft copy of this report. As a research linguist who once evaluated a voice biometrics device using vocoded speech (which was a pioneering effort in voice biometrics research), her comments are most valuable.

Finally, the authors thank NRL members Dave Heide, Tom Mora, and their associates, Dan Latham and Toby Cotten. They assisted the authors in making this project successful.

REFERENCES

1. DoD Biometrics Management Office, "Beyond Passwords, Keys, Locks, and Pins," http://www.c3i.osd.mil/biometrics/

2. P.D. Bricker, R. Gnanadesikan, M.V. Matthews, S. Pruzansky, P.A. Tucky, K.W.Wachter, and J.L. Warner, "Statistical Techniques for Talker Identification," *Bell Syst. Tech. J.* **50**, 1427-1454 (1971).

3. W.S. Mohns, "Statistical Features Evaluation in Speaker Identification," Ph.D. dissertation, North Carolina State University, Raleigh (1969).

4. G.R. Doddington,""A Method of Speaker Verification," Ph.D. dissertation, University of Wisconsin, Madison (1970).

5. B.S. Atal, "Automatic Recognition of Speakers from Their Voices," *Proc. IEEE*, **64**(4), 460-475 (1976).

6. A.E. Rosenberg, "Automatic Speaker Verification," *Proc. IEEE*, **64**(4), 475-487 (1976).

7. B.S. Atal and S.L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *J. Acoust. Soc. Amer.* **50**(2), 637-655 (1971).

8. B.S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification," *J. Acoust. Soc. Am.*, **55**(6), 1304-1312 (1974).

9. M.R. Sambur, "Speaker Recognition Using Orthogonal Linear Prediction," *IEEE Trans. Acoustics, Speech, Sig. Proc.* **ASSP-24**(4), 283-289 (1976).

10. G.A. Mian, "Some Factors Influencing the Performance of a Speaker Recognition System Based on LPC," *IEEE ICASSP,* 781-784 (1979)

11. R.C. Lummis, "Speaker Verification by Computer Using Speech Intensity for Temporal Registration," *IEEE Trans. Audio Electroacoust.* **AU-21**(2), 80-89 (1973).

12. J.E. Atkinson, "Inter- and Intraspeaker Variability in Fundamental Voice Frequency," *J. Acoust. Soc. Amer.* **60**(2), 440-445 (1976).

13. J.D. Markel, B.T. Oshika, and A. Gray, "Long-Term Feature Averaging for Speaker Recognition," *IEEE Trans. Acoustics, Speech, Sig. Proc.* **ASSP-25**(4), 330-337 (1977).

14. J.T. Buck, D.K. Burton, and J.E. Shore, "Text-dependent Speaker Recognition Using Vector Quantization," *IEEE ICASSP,* 391-394 (1985).

15. F.K. Soong, A.E. Rosenberg, L.R. Labiner, and B.H. Juang, "A Vector Quantization Approach to Speaker Recognition," *IEEE ICASSP,* 387-402 (1985).

16. H. Ney and R. Gierloff, "Speaker Recognition Using a Feature Weighting Technique," *IEEE ICASSP*, 1645-1648 (1982).

17. T. Matsui, T. Nishitani, and S. Furui, "Robust Methods of Updating Model and A Priori Threshold in Speaker Verification," *IEEE ICASSP*, 97-100 (1996).

18. P. Jesorsky, U. Höfker, and M. Talmi, "Extraction of Speaker-specific Features from Spoken Code Sentences," *IEEE ICASSP*, 279-282 (1978).

19. H. Ney, "Telephone-line Speaker Recognition Using Clipped Autocorrelation Analysis," *IEEE ICASSP*, 188-192 (1981).

20. NATO, "The Impact of Speech Under "Stress" on Military Speech Technology," RTO-TR-10, NATO Research and Technology Organization, BP 25, 7 rue Ancelle, F-92201 Neuilly-sur-seine cedex, France. (2000)

21. G.S. Kang and D.A. Heide, "Acoustic Noise Reduction for Speech Communication (Second-order Gradient Microphone)," *IEEE Int. Conf. on Circuit Theory*, IV-556-559 (1999).

22. G.S. Kang, L.J. Fransen, T.M. Moran, and D.A. Heide, "Voice Processing Techniques for C^4I Applications," NRL/FR/5550--97-9852 (1997).

23. G.S. Kang and S.S. Everett, "Improvement of the Narrowband Linear Predictive Coder, Part I — Analysis Improvement," NRL Report 8645 (1982).

24. S. Hayakawa and F. Itakura, "Text-Dependent Speaker Recognition Using the Information in the Higher Frequency Band," *IEEE ICASSP*, I-137-140 (1994).

25. http://www.gulflink.osd.mil/mopp/index.html

26. J.K. Smart, "History of the Army's Protective Mask," NBC Systems, ATTN: AMSSB-SCI-H E5232, Aberdeen Proving Ground, MD 21010-5424 (Undated).

27. J.G. Proakis and D.G. Manolakis, *Introduction to Digital Signal Processing* (Macmillan Publishing Company, Old Tappan, NJ, 1988).