ELSEVIER

Speech Communication 20 (1996) 151–173

SPEECH COMMUNICATION

Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition [1]

John H.L. Hansen *

Robust Speech Processing Laboratory, Department of Electrical Engineering, Box 90291, Duke University, Durham, NC 27708-0291, USA

Received 29 January 1996; revised 15 June 1996

Abstract

It is well known that the introduction of acoustic background distortion and the variability resulting from environmentally induced stress causes speech recognition algorithms to fail. In this paper, several causes for recognition performance degradation are explored. It is suggested that recent studies based on a *Source Generator Framework* can provide a viable foundation in which to establish robust speech recognition techniques. This research encompasses three inter-related issues: (i) analysis and modeling of speech characteristics brought on by workload task stress, speaker emotion/stress or speech produced in noise (Lombard effect), (ii) adaptive signal processing methods tailored to speech enhancement and stress equalization, and (iii) formulation of new recognition algorithms which are robust in adverse environments. An overview of a statistical analysis of a Speech Under Simulated and Actual Stress (SUSAS) database is presented. This study was conducted on over 200 parameters in the domains of pitch, duration, intensity, glottal source and vocal tract spectral variations. These studies motivate the development of a speech modeling approach entitled *Source Generator Framework* in which to represent the dynamics of speech under stress. This framework provides an attractive means for performing feature equalization of speech under stress. In the second half of this paper, three novel approaches for signal enhancement and stress equalization are considered to address the issue of recognition under noisy stressful conditions. The first method employs (Auto:I,LSP:T) constrained iterative speech enhancement to address background noise and maximum likelihood stress equalization across formant location and bandwidth. The second method uses a feature enhancing artificial neural network which transforms the input stressed speech feature set during parameterization for keyword recognition. The final method employs morphological constrained feature enhancement to address noise and an adaptive Mel-cepstral compensation algorithm to equalize the impact of stress. Recognition performance is demonstrated for speech under a range of stress conditions, signal-to-noise ratios and background noise types.

Zusammenfassung

Es ist wohlbekannt, dass die Einführung von Hintergrundgeräuschen und von Variabilität der Umgebung dazu führen, dass Spracherkennungsalgorithmen versagen. In diesem Paper werden verschiedene Fälle untersucht, die zu einer Minderung des Erkennungsgrades führen. Es wird vorgeschlagen, dass gegenwärtige Untersuchungen, basierend auf *Source Generator Framework*, eine variable Grundlage bilden, in der robuste Spracherkennungstechniken aufgebaut werden können. Diese

* E-mail: jhlh@ee.duke.edu.

[1] Audiofiles available. See http://www.elsevier.nl/locate/specom .

PII S0167-6393(96)00050-7

Untersuchung schliesst drei Punkte mit ein, die damit in Beziehung stehen: (i) Analyse und Modellierung von Sprachcharakteristika, die durch Stress, Emotionen oder Sprache in einer lauten Umgebung (Lombard Effekt), herinführen, (ii) adaptive Signalverarbeitungsmethoden, angepasst an den Ausgleich von Betonungen und (iii) Formulierung neuer und robuster Spracherkennungsalgorithmen. Ein Überblick über eine statistische Analyse von Sprache unter simulierten und aktuellen Stressdatenbanken (SUSAS) wird gegeben. Diese Untersuchung wurde an mehr als 200 Parametern ausgeführt in den Bereichen Länge, Intensität und vokal spektrale Variationen. Diese Untersuchungen motivieren die Entwicklung eines Sprachmodellierungsansatzes, genannt *Source Generator Framework*, bei dem die Dynamik der Sprache unter Stress dargestellt wird. In der zweiten Hälfte des Papers werden drei Ansätze zum Stressausgleich vorgestellt, um auch den Punkt der Spracherkennung in einer verrauschten Umgebung anzusprechen. Die erste Methode beinhalten (Auto:I,LSP:T) beschränkte iterative Sprachzusätze, um Hintergrundgeräusche zu erfassen sowie mit höchster Wahrscheinlichkeit einen Stressausgleich über Bandbreiten und Ort hinweg zu erreichen. Die zweite Methode benutzt die Eigenschaft, künstliche neuronale Netze durch Eigenschaften zu erweitern, welche verrauschte Eingaben (die während der Parametrierung für Schlüsselworterkennungen entstehen) transformiert. Die letzte Methode beinhaltet morphologisch beschränkten Zusatz von Eigenschaften, um Rauschen zu betrachten sowie einen adaptiven Mel-cepstral Kompensationsalgorithmus, um den Einfluss von Stress auszugleichen. Der Grad der Erkennung wird demonstriert für Sprache unter einem grossen Bereich von Stressbedingungen, Signal-Rauschen Verhältnis sowie Hintergrundgeräuschen.

Résumé

Il est connu que la distorsion acoustique introduite par l'environnement ambiant ainsi que la variabilité résultant du stress induit détériorent énormément les performances des algorithmes de reconnaissance. Dans cet article, on explore les diverses causes de dégradation de ces performances. On suggère que les études récentes effectuées sur l'approche appelée *Source Generator Framework* produisent un fondement viable pour développer des techniques robustes de reconnaissance de la parole. L'étude décrite s'articule autour de trois axes corrélés: (i) l'analyse et la modélisation de la parole produite soit sous l'effet de stress du à la charge de travail et/ou à l'émotion, soit dans le bruit, (ii) les méthodes de traitement adaptatif du signal pour le débruitage de la parole et la réduction de l'effet du stress, et (iii) la formulation de nouveaux algorithmes robustes de reconnaissance. Une analyse statistique d'une base de données (SUSAS) de parole sous stress simulé et réel est présentée. Cette analyse a été menée sur plus de 200 paramètres relatifs au pitch, à la durée, à l'intensité, à la source glottique et aux variations des spectres du conduit vocal. Ces études ont motivé le développement de l'approche appelée *Source Generator Framework* qui permet de modéliser la dynamique de la parole sous stress. Ce cadre offre des moyens intéressants pour effectuer l'égalisation des paramètres de la parole sous stress. Dans la seconde moitié de l'article, trois nouvelles approches pour le débruitage de la parole et la réduction de l'effet du stress sont considérées. La première méthode utilise la technique itérative contrainte (Auto:I,LSP:T) de débruitage et une égalisation par maximum de vraisemblance de la parole à travers la localisation des formants et leurs bandes passantes. Pour la reconnaissance de mots clés, la seconde méthode utilise un réseau de neurones qui transforme les vecteurs de paramètres de la parole sous stress pendant la phase de paramétrisation. La dernière méthode applique une technique de rehaussement des paramètres basée sur des contraintes morphologiques pour effectuer le débruitage et utilise un algorithme adaptatif sur les cepstres-Mel pour égaliser les effets du stress. Les performances de reconnaissance sont données pour la parole produite dans plusieurs conditions de stress, avec plusieurs rapports signal/bruit, et pour différents types de bruit ambiant.

Keywords: Speech under stress; Lombard effect; Robust speech recognition; Noise suppression

1. Introduction: why recognizers break

The issue of robustness in speech recognition can take on a broad range of problems. A speech recognizer may be robust in one environment and inappropriate for another. The main reason for this is that performance of existing recognition systems which assume a noise-free tranquil environment, degrade rapidly in the presence of noise, distortion and stress. In Fig. 1, a general speech recognition scenario is presented which considers a variety of speech signal distortions. Here, the index *n* represents time. For this scenario, we assume that a speaker is exposed to some adverse environment, where ambient noise is present and a stress induced task is required (or the speaker is experiencing emotional stress). The ad-

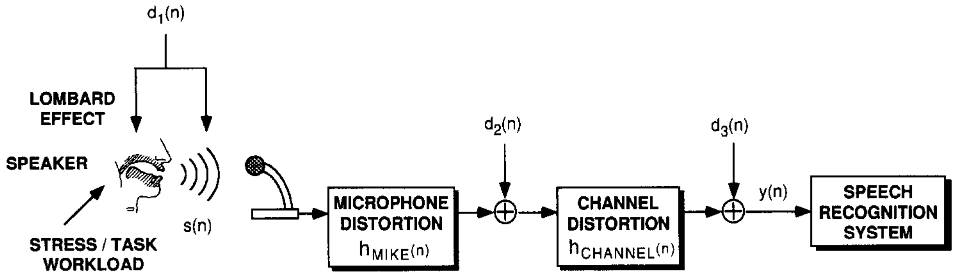


Fig. 1. Types of distortion which can be addressed for robust speech recognition.

verse environment could be a noisy automobile where cellular communication is used, high-stress noisy helicopter or aircraft cockpits, factory environments, and others. Since the user task could be demanding, the speaker is required to divert a measured level of cognitive processing, leaving formulation of speech for recognition as a secondary task.

Workload task stress has been shown to significantly impact recognition performance (Chen, 1988; Hansen, 1988, 1989; Paul, 1987; Rajasekaran et al., 1986). Since background noise is present, the speaker will experience the *Lombard* effect (Lombard, 1911), a condition where speech production is altered in an effort to communicate more effectively across a noisy environment. The level of *Lombard* effect may depend on the type and level of ambient noise $d_1(n)$ (though no studies have considered this), and has been shown to vary between male and female speakers (Junqua, 1993). In addition, a speaker may also experience situational stress (i.e., anger, fear, other emotional effects) or workload task stress (i.e., flying an aircraft) which will alter the manner in which speech is produced. If we assume $s(n)$ to represent a *Neutral*, noise-free speech signal, then the acoustic signal at the microphone will include distortion due to stress, workload task, *Lombard* effect and additive noise. The acoustic background noise $d_1(n)$ will also degrade the speech signal as illustrated in Fig. 1. Next, if the speech recognition system is trained with one microphone and another is used for testing, then distortion due to microphone mismatch can be modeled with a frequency distortion impulse response $h_{\text{MIKE}}(n)$. If the speech signal is transmitted over a telephone or cellular channel, further distortion is introduced (modeled as either additive noise $d_2(n)$, or a frequency distortion with impulse response $h_{\text{CHAN}}(n)$). Furthermore, noise could also be present at the receiver $d_3(n)$. Therefore, the *Neutral* noise-free distortionless speech signal $s(n)$, having been produced and transmitted under adverse conditions, is transformed into the degraded signal $y(n)$.

$$y(n) = \left(\left(\left\{\left[s(n) \;\middle|\; \begin{array}{l}\text{Workload task}\\ \text{Stress}\\ \text{Lombard effect}\{d_1\}\end{array}\right] + d_1(n)\right\} * h_{\text{MIKE}}(n) + d_2(n)\right) * h_{\text{CHAN}}(n)\right) + d_3(n). \qquad (1)$$

We should emphasize that all forms of distortion identified in Eq. (1) and Fig. 1 may not exist simultaneously. In this study, the primary focus will be on speech under stress (including Lombard effect), with secondary emphasis on speech under stress with additive background noise distortion.

2. Recent methods and studies

Approaches for robust recognition can be summarized under three areas: (i) better training methods, (ii) improved front-end processing, and (iii) improved back-end processing or robust recognition measures. These approaches have been used to address improved recognition of speech in (a) noisy environments, (b) *Lombard* effect, (c) workload task stress or speaker stress, and (d) microphone or channel mismatch.

To formulate automatic speech recognition algorithms which are more effective in changing environmental conditions, it is important to understand the effects of noise on the acoustic speech waveform, the acoustic-phonetic differences between normal speech and speech produced in noise, and the acoustic-phonetic differences between normal speech and speech produced under stressed conditions. Several studies have shown distinctive differences in phonetic features between normal and *Lombard* speech (Hanley and Harvey, 1965; Hansen, 1988; Junqua, 1993; Stanton et al., 1988), and speech spoken in noise (Gardner, 1966; Pisoni et al., 1985; Summers et al., 1988). Further studies have focused on variation in speech production brought on by task stress or emotion (Bou-Ghazale and Hansen, 1995; Hansen, 1988, 1989, 1993, 1995; Murray, 1993). The primary purpose of these studies has been to improve the performance of recognition algorithms in noise (Hansen and Clements, 1991; Juang, 1991; Alexandre and Lockwood, 1993), *Lombard* effect (Junqua, 1993; Hanson and Applebaum, 1990, 1993; Stanton et al., 1989), stressed speaking styles (Lippmann et al., 1987; Paul, 1987; Chen, 1988; Lockwood and Boudy, 1992; Lockwood et al., 1992), noisy *Lombard* effect (Hansen, 1988, 1994; Hansen and Bria, 1990) and noisy stressful speaking conditions (Rajasekaran et al., 1986; Hansen, 1988, 1989, 1993; Hansen and Clements, 1989). Other studies have also considered feature analysis methods for classification of speech under stress (Cairns and Hansen, 1994; Hansen and Womack, 1996).

Approaches based on improved training methods include multi-style training (Lippmann et al., 1987; Paul, 1987), simulated stress token generation (Bou-Ghazale and Hansen, 1994, 1995), training/testing in noise (Dautrich et al., 1983), and others (Juang, 1991). Improved training methods can increase recognition performance; however, results degrade as test conditions drift from the original training data. A solution which has been suggested is fast update methods for recognition models under varying noise environments. While it may be possible to show that training a recognizer on noise-corrupted speech databases leads to higher performance than attempting to improve input SNR of test utterances (Mokbel and Chollet, 1995), this result ignores the devastating impact of Lombard effect for high noise environments. In fact, even if background noise could be addressed in this manner, poor recognition performance will persist due to changing speech characteristics caused by stress and Lombard effect.

Another area which has received much attention is front-end processing/speech feature-estimation for robust recognition. Here, many studies have attempted to uncover a speech representation which is less sensitive to various levels and types of additive, linear filtering or convolutional distortion. For example, some studies focus on identifying better speech recognition features (Hanson and Applebaum, 1990, 1993), or estimation of speech features in noise (Hansen and Clements, 1991; Lockwood and Boudy, 1992), or processing to obtain better speech representations (Hermansky and Morgan, 1994; Hunt and Lefebvre, 1989). If the primary distortion is additive noise, then a number of speech enhancement algorithms exist (Ephraim, 1992; Hansen and Clements, 1991; Lockwood et al., 1992; Nandkumar and Hansen, 1995; Hansen and Nandkumar, 1995), while other front-end processing methods incorporate feature processing for noise reduction and stress equalization [2] (Hansen, 1993; Hansen and Clements, 1989; Hansen and Cairns, 1995), or additive/convolutional noise (Hermansky et al., 1993; Gales and Young, 1995).

The last area is improved back-end processing or robust recognition measures. Such processing methods refer to changes in the recognizer formulation such as hidden Markov model structure, or developing better models of noise within the recognizer (Wang and Young, 1992). Robust recognition measures seek to project either the test data space closer to the trained recognition space, or trained space towards test space (Mansour and Juang, 1988; Carlson and Clements, 1992). Studies relating to robust metrics include linear filtering or microphone mismatch distortion processing (Liu et al., 1992).

[2] The concept of stress equalization is based on a processing scheme which operates on a parameter sequence which is extracted from input speech under stress. The stress equalization algorithm attempts to normalize the variation of the parameter sequence due to the presence of stress on the input speech signal.

3. Analysis and modeling of speech under stress

Stress is a psychological state that is a response to a perceived threat or task demand and is normally accompanied by specific emotions. Psychiatrists agree that verbal markers of stress range from highly visible to invisible markers (Goldberger and Breznitz, 1982; Darby, 1981). Researchers have also considered the effects of aircraft pilot stress (Flack, 1918; Williams and Stevens, 1969) and its impact on speech data for recognition (Russell and Moore, 1983). Still others have considered speech and emotion (Lieberman and Michaels, 1962; Williams and Stevens, 1972), workload (Lively et al., 1993) and Lombard effect (Lombard, 1911; Bond and Moore, 1990; Junqua, 1993).

In this section, we present results from an investigation of how stress affects speech characteristics with specific application to improving automatic speech recognition. Past stress studies have been limited in scope, often using only one or two subjects and analyzing only one or two parameters (typically pitch). A comprehensive speech under stress database has been established for the purposes of stress research. Analysis was first performed on (i) *speech with simulated stress, workload tasks, or speech in noise*. Statistically significant parameters were established, and an equivalent analysis performed with (ii) *speech produced under actual stress or emotion*. This scheme was chosen since simulated conditions allowed for careful control of vocabulary, task requirements and background noise characteristics. Evaluation over actual stress conditions was used to verify results established under simulated conditions (see Hansen, 1988, 1989, 1994, 1995; Hansen and Bria, 1990; Hansen and Clements, 1987 for further details).

3.1. SUSAS: speech under stress database

The studies conducted in this research were based on data previously collected for analysis and algorithm formulation of speech recognition in noise and stress. This database, called *SUSAS*, refers to *Speech Under Simulated and Actual Stress*, and has been employed extensively in the study of how speech production and recognition varies when speaking during stressed conditions (Hansen, 1988, 1989, 1994, 1995; Hansen and Bria, 1992). SUSAS consists of five domains, encompassing a wide variety of stresses and emotions. A total of 44 speakers were employed to generate in excess of 16,000 isolated-word utterances. The five stress domains include (i) psychiatric analysis data (speech under depression, fear, anxiety), (ii) talking styles [3] (slow, fast, soft, loud, angry, clear, question), (iii) single tracking computer response task or speech produced in noise (Lombard effect), (iv) dual tracking computer response task, (v) subject motion-fear tasks (G-force, Lombard effect, noise, fear). A common highly confusable vocabulary set of 35 aircraft communication words make up the database (e.g., /go–oh–no/, /wide–white/, etc.). A more complete discussion of SUSAS can be found in the literature (Hansen, 1995, 1994, 1988; Hansen and Bria, 1990; Hansen and Cairns, 1995) [4]. The subset of data for this study consists of neutral training and test data, and speech from ten stressed styles (talking styles, single tracking task and Lombard effect domains). For talking styles, speakers were asked to speak as if they were producing speech under that style. Speech data under Lombard effect was produced by having speakers listen to 85 dB SPL pink noise binaurally while uttering test tokens (i.e., all tokens are noise-free). Speech under task condition required talkers to produce speech while performing a single workload tracking task on a computer screen. All speech tokens were sampled using a 16-bit A/D converter at a sample rate of 8 kHz.

To illustrate the problem of speech recognition in stress and noise, a baseline speech recognizer (VQ-HMM) [5] was employed on noise-free and noisy

[3] Approximately half of the SUSAS database consists of style data donated by Lincoln Laboratories (Lippmann et al., 1987; Paul, 1987; Chen, 1988; Hansen, 1988; Hansen and Clements, 1989).

[4] An audio demonstration of speech data from SUSAS is available at http://www.elsevier.nl/locate/specom. A brief summary of the demonstration is included in Appendix A.

[5] This baseline speech recognizer VQ-HMM is a speaker dependent, isolated word system, which uses discrete observations from a 64-entry vector quantizer observation codebook and 5-state left-to-right hidden Markov models which are fully connected (i.e., all state transitions from left-to-right are possible). Further details regarding this baseline recognizer can be found in previous studies (Hansen, 1994, 1993, 1988; Hansen and Arslan, 1995a).

Table 1
Recognition performance of neutral and stressed type speech in noise-free and noisy conditions

Condition	Stressful speech recognition results												
	N	Sl	F	So	L	A	C	Q	C50	C70	Lom	Avg10	StDev10
Stressful, noise-free	88.3%	60%	65%	48%	50%	20%	68%	75%	63%	63%	63%	57.5%	15.35
Stressful, noisy [a]	49%	45%	28%	33%	18%	15%	40%	28%	35%	33%	28%	30.3%	9.12

[a] Additive white Gaussian noise, SNR = +30 dB.

Stressed speech style key: N: neutral; Sl: slow; F: fast; So: soft; L: loud; A: angry; C: clearly spoken; Q: question; C50: moderate computer workload task condition; C70: high computer workload task condition; Lom: Lombard effect noise condition.

stressed speech from SUSAS. Table 1 shows that when stress and noise are introduced, recognition rates decrease significantly. When white Gaussian noise is introduced, noisy stressed speech rates varied, with an average rate of Avg10 = 30.3% (i.e., a 58% decrease from the 88.3% noise-free neutral rate). Recognition performance also varies considerably across stressed speaking conditions as reflected in the large standard deviation in rate of recognition. (StDev10 = 15.35, 9.12 for noise-free and noisy stressed conditions).

3.2. Source generator framework

Since noise, stress and *Lombard* effect have been shown to disrupt speech recognition, we consider the following *Source Generator Framework* as a means of representing the variation of speech production in noise and stress. Source generator theory was first presented in (Hansen, 1994), and later employed in other robust recognition algorithms (Hansen, 1993; Bou-Ghazale and Hansen, 1994; Hansen and Cairns, 1995; Hansen and Clements, 1995).

Let $\vec{s}$ be a sample vector of clean *Neutral* speech $s(n)$ in a sample space $\mathcal{T}_s$. Also, let the sample space $\mathcal{T}_s$ consist of J independent and mutually exclusive random speech type sources

$$\vec{s} \in \mathcal{T}_s : \{\gamma_j;\ j = 1,2,\ldots,J\}. \qquad (2)$$

Here, the collection of generators $\vec{\gamma}$ span the entire source generator space, and each generator γ_j could represent an isolated phoneme, a transition between pairs of phonemes, or some other temporal partition of how the speech signal is produced. It is known that the presence of stress will effect how the

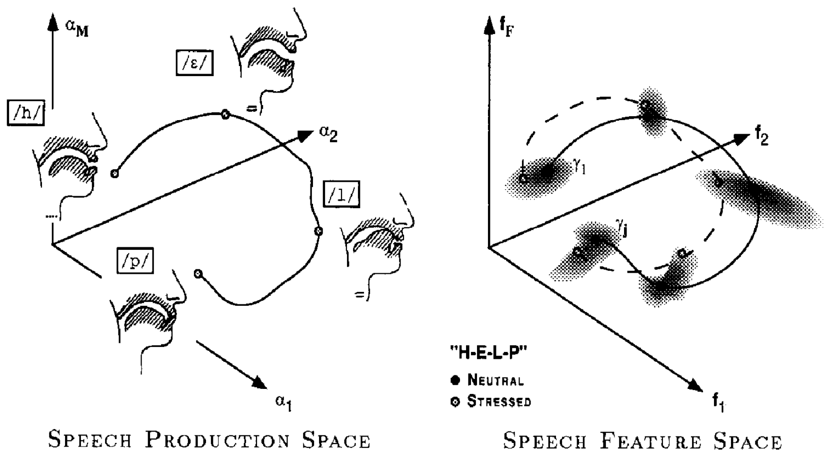


Fig. 2. Proposed source generator framework for modeling speech under stress. (a) Speech production space; (b) speech feature space.

speech production system produces the observation vector $\vec{s}$. In Fig. 2, production variation of the utterance "help" is illustrated for neutral and stressed speech. For the production of this word, we assume that a sequence of coordinated movements of the vocal system articulators and excitation controls are needed (represented in the multi-dimensional speech production space as $\alpha_1, \ldots, \alpha_M$). The coordinated sequence of excitation and articulatory controls are modeled as a smooth path in this speech production space. It is hypothesized that vocal system controls $\vec{\alpha}$ (i.e., articulators, etc.) will be perturbed under stressed speaking conditions resulting in deviations from this "neutral" production space path. From previous studies, it is known that the presence of stress will cause changes in phoneme production with respect to glottal source factors, pitch, intensity, duration and spectral shape (Hansen, 1988). It is proposed that the perturbation of these vocal system controls can be modeled by a change in the speech source generator γ_{j} in some F-dimensional feature space. As Fig. 2(b) illustrates, each source generator γ_i will occupy some volume in the multi-dimensional feature space, and that deviations in speech production under stress will result in a feature sequence which deviates from the mean "neutral" path. Let each change be represented by a mapping of the neutral speech samples $s_{\gamma_j}^{\mathrm{NEUTRAL}}$ for source generator γ_{j} to stressed ($c =$ STRESSED CLASS) speech samples as follows:

$$\Psi_{D(k),c}[\gamma_j] : s_{\gamma_j}^{\mathrm{NEUTRAL}} \to s_{\gamma_j}^{\mathrm{STRESSED}}, \tag{3}$$

$$D \in (\text{PITCH, DURATION, INTENSITY,}$$
$$\text{GLOTTAL SOURCE,}$$
$$\text{VOCAL TRACT SPECTRUM}),$$

$$c \in (\text{NEUTRAL, SLOW, FAST, SOFT, LOUD,}$$
$$\text{ANGRY, CLEAR, QUESTION, C50 TASK,}$$
$$\text{C70 TASK, LOMBARD}),$$

where the particular speech production domain is represented as D, corresponding to the five speech feature domains. Here, the speech feature domain is represented as $D(\vec{k})$, which corresponds to a set of $\vec{k}$ features per domain which could be mapped due to stress (i.e., if D is the SPECTRAL domain, then $D(\vec{k})$ could be the four mean formant locations for generator γ_j). In addition, it is assumed that the mapping for each generator for an input word or phrase is performed over the same stress class c in $\Psi_{\mathrm{D(k),c}}[\gamma_j]$ (i.e., for the input word "help", we assume that the particular stress class c is the same for each generator for the entire word). We therefore do not allow individual generators to be under different stress conditions. Here, the stress generator class c corresponds to one of the eleven speaking styles $c \in$ (SLOW,...,LOMBARD) (also summarized in Table 1). Given that the feature domain D consists of a multi-dimensional production space $D(\vec{k})$, the general *Neutral* speech vector $\vec{s} \in \Upsilon_s$ is modeled under stressed conditions as follows:

$$\Psi_{\mathrm{PITCH}(\vec{k}),c}\Big[\Psi_{\mathrm{DURA}(\vec{k}),c}\Big[\Psi_{\mathrm{INTEN}(\vec{k}),c}\Big[\Psi_{\mathrm{GLOT}(\vec{k}),c}\Big[\Psi_{\mathrm{SPEC}(\vec{k}),c}[\gamma_J]\Big]\Big]\Big]\Big] : s_{\gamma_j}^{\mathrm{NEUTRAL}} \to s_{\gamma_j}^{\mathrm{LOMBARD}}, \tag{4}$$

where for this case we let the stress class c be LOMBARD, $\vec{k}$ spans the number of features needed for each speech production domain, and j spans the number of possible source generators. The model suggests that production of the sample speech vector $\vec{s}$ in the sample space Υ_s is achieved by transforming the speech source generator γ for the jth speech type across each of the five production feature domains.

Next, let $\vec{y}$ be a sample vector from some *Neutral* source generator γ_j which is corrupted by an additive noise vector $\vec{d}$. The resulting noisy, stressed induced speech vector from source generator γ_j is written as

$$\vec{y}_{\vec{\Psi}_{D(k),c}[\gamma_j]} = \vec{s}_{\vec{\Psi}_{D(k),c}[\gamma_j]} + \vec{d}, \tag{5}$$

where the vector of feature domain transformations is written as

$$\vec{\Psi}_{D(k),c}[\gamma_j] = \Psi_{\mathrm{PITCH}(\vec{k}),c}\Big[\Psi_{\mathrm{DURA}(\vec{k}),c}\Big[\Psi_{\mathrm{INTEN}(\vec{k}),c}\Big[\Psi_{\mathrm{GLOT}(\vec{k}),c}\Big[\Psi_{\mathrm{SPEC}(\vec{k}),c}[\gamma_j]\Big]\Big]\Big]\Big], \tag{6}$$

which implies application of the speech production feature transformations to the input source generator γ_j in a serial manner. This equation represents a general source generator framework in which to investigate speech recognition under stressful conditions. Next, we consider analysis of the five speech feature domains in order to establish useful features $\vec{k}$ in each domain for stress normalization.

3.3. Speech production feature analysis

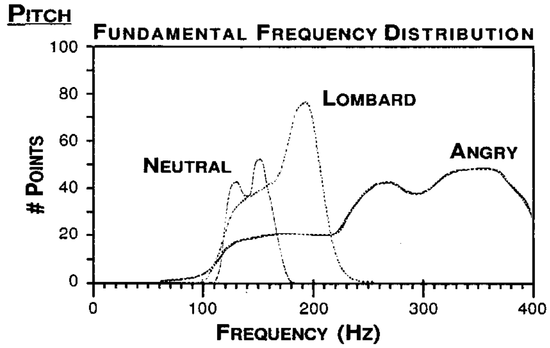
Due to difficulty in experimental design and limited research efforts, changes in the characteristics of speech produced under workload stress remain unclear. Thusfar, research has been limited in scope, often using only one or two subjects and analyzing a single parameter (often f_0). It is not unusual for researchers to report conflicting results, due to differences in experimental design, level of actual or simulated stress, or interpretation of results. For example, some studies concentrate on analysis of recordings from actual stressful situations (Kuroda et al., 1976; Simonov and Frolov, 1977; Streeter et al., 1983; Williams and Stevens, 1969, 1972), while others use simulated stress or emotions (Hecker et al., 1968; Hicks and Hollien, 1981; Williams and Stevens, 1972). This offers the advantage of a controlled environment, where a single emotion can be examined with little background noise. This allows results to be based on general speaker characteristics instead of possibly particular characteristics of an individual speaker in conveying emotion. The major disadvantage in these studies has been the reduction in task stress levels. In addition, studies using actors may produce exaggerated caricatures of emotions in speech.

Here, we discuss several results from a comprehensive investigation of acoustic correlates of speech under stress (Hansen, 1988, 1995). In these studies, well over 200 parameters and 10,000 statistical tests were considered in evaluating the following parameter areas of speech production: (i) pitch, (ii) duration, (iii) intensity, (iv) glottal source, and (v) vocal tract spectrum.

3.3.1. Pitch

The most widely considered area of stress evaluation are characteristics of pitch. In our studies, we have considered subjective assessment of pitch contours, statistical analysis of pitch mean, variance and distribution (see Fig. 3). A partial list of conclusive points are:

- Mean pitch values may be used as significant indicators for speech in soft, fast, clear, Lombard, question, angry or loud styles when compared to neutral conditions.



	NEUTRAL	LOMBARD	ANGRY
MEAN (Hz)	145	160	253
STAND. DEV.	15	24	95

Fig. 3. Sample pitch (fundamental frequency) variation for speech under neutral, Lombard and angry stress conditions.

- Loud, angry, question and Lombard mean pitch are all significantly different from all other styles considered.
- Speech produced under Lombard effect gave mean pitch values most closely associated with pitch from fast and clear conditions.
- Soft and loud pitch variance are significantly different from all styles considered.
- Pitch variance for clear and Lombard conditions are similar, but different from all other styles considered.

3.3.2. Duration

Previous studies of speech under stress have not considered statistical evaluations of individual phoneme class duration. Duration analysis was conducted across (i) whole words, and (ii) individual phoneme-classes (vowel, consonant, semivowel and diphthong). An analysis was also conducted on inter-class duration movement to determine if speakers increased duration of certain phoneme classes at the expense of others (see Fig. 4). A partial list of duration conclusions are:

- Mean word duration may be used as significant indicators for speech in slow, clear, angry, loud, Lombard or fast styles when compared to neutral.
- Slow and fast mean word duration are all significantly different from all other styles considered.
- Clear mean consonant duration was significantly different from all styles except slow.

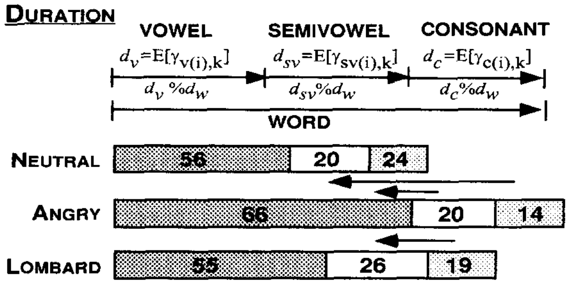


Fig. 4. Sample duration variation for speech under neutral, Lombard and angry stress conditions. Statistically significant duration shifts between vowel, semivowel and consonant classes are indicated by arrows.

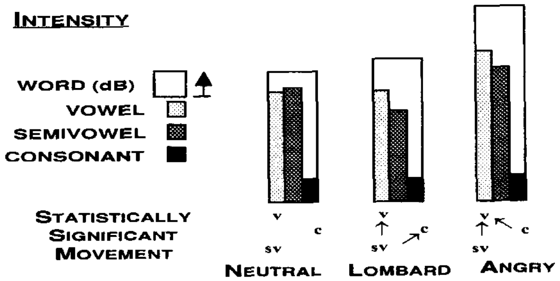


Fig. 5. Sample RMS intensity (dB) variation for speech under neutral, Lombard and angry stress conditions. Statistically significant energy shifts between vowel, semivowel and consonant classes are indicated by arrows.

- Duration variance increased for all domains (word, vowel, consonant, semivowel, diphthong) under slow stress.
- Duration variance decreased for most domains under fast stress condition.
- Duration variance significantly increased for angry speech.
- Clear consonant duration variance was significantly different from all styles.

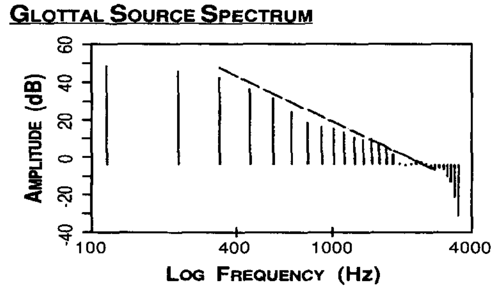
3.3.3. Intensity

An analysis was conducted on (i) whole word intensity, and (ii) speech phoneme-class intensity (vowel, semivowel, diphthong, consonant). Statistical tests were performed on mean, variance and distribution across the database. The shift in available energy between speech classes was also considered to determine if speakers reduce intensity in some classes in order to increase others (Fig. 5). A partial list of conclusions are:

- Average RMS word intensity values may be used as significant indicators for speech in angry, loud and high workload task styles when compared to neutral conditions.
- Loud & angry RMS word intensity are significantly different from all other styles considered.
- Loud & angry RMS vowel and diphthong intensities were significantly different from all other styles considered.
- RMS consonant & semivowel intensity are not significant stress indicators for any of the styles considered.
- Variance of average RMS word intensity values may be used as significant indicators for speech in angry and loud styles when compared to neutral.
- Variance of loud and angry average RMS word intensity is significantly different from most other styles considered.

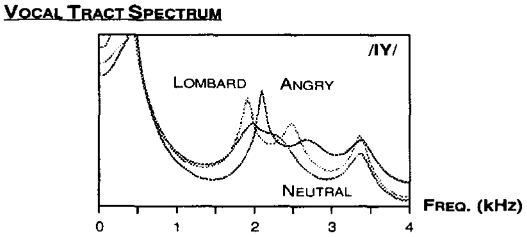
3.3.4. Glottal source

Aspects such as duration of each laryngeal pulse (open/closed periods), instant of glottal closure, spectral structure of each glottal pulse, or pulse shape play important roles in conveyance of stress state (Hansen, 1988; Cummings and Clements, 1990). Due to limitations of glottal inverse filtering techniques in stress evaluation, this portion focused on direct estimation of the glottal flow spectrum. Examples of spectral structure, average spectral value and



	NEUTRAL	LOMBARD	ANGRY
AVG. SPEC. (dB)	15.2	20.8	23.0
SLOPE (dB/OCT.)	-12.1	-9.2	-9.4

Fig. 6. Sample glottal source spectra for speech under neutral, Lombard and angry stress conditions.



	F1 (Bw)	F2 (Bw)	F3 (Bw)	F4 (Bw)
NEUTRAL	411 (52)	1970 (222)	2607 (496)	3368 (366)
LOMBARD	412 (73)	2006 (139)	2644 (250)	3376 (185)
ANGRY	586 (102)	2078 (166)	2661 (464)	3357 (392)

Fig. 7. Sample vocal tract spectra variation for speech under neutral, Lombard and angry stress conditions.

spectral slope (in decibles/octave) are shown in Fig. 6. The present analysis of glottal source spectrum revealed that parameters such as spectral slope and the distribution of energy are important for relaying stress.

3.3.5. *Vocal tract spectrum*

Analysis of vocal-tract spectrum focused on formant location and bandwidth for selected phonemes across the SUSAS database. Mean and variance estimates for specific phonemes were analyzed for the 11 stress conditions. Statistical evaluations showed that of the 400 Student *t*-tests, 166 were statistically different from neutral. Most of these involved loud, angry or Lombard effect formant information. A majority of the significant comparisons involved mean and variance of formant location and bandwidth for F1 and F2. Of the ten stress conditions, average formant information for loud and angry were the most consistent across the phonemes tested (Fig. 7).

In this section, we have considered a brief discussion of an analysis of speech under stress. The focus was on speech from simulated stressed conditions. A similar evaluation was also conducted on speech from actual stress conditions to confirm direction and degrees of speech parameter variation. A complete discussion of these results cannot be fully addressed here, and therefore the interested reader may consider the following references (Hansen, 1988, 1989, 1995). However, these results serve to motivate the type of speech processing needed to address recognition of speech under stress.

3.4. *Robust feature enhancement in noise*

Though a number of speech enhancement algorithms have been proposed in the past, the majority are actually *noise cancelers* as opposed to *speech enhancers*. As a result, these systems have limited success in suppressing noise or improving intelligibility in low-energy and transitional regions of speech. Recently, a new class of constrained iterative algorithms have been formulated which focus on developing a better physical characterization of speech production for improved "speech enhancement" by adapting the enhancement process to varying types of input speech. This represents an innovative departure from classical noise filtering schemes, since spectral characteristics caused by changes during speech production are used during single or dual-channel enhancement. This has resulted in substantial and consistent quality improvement for human listeners in both white and colored noise environments. The following techniques are suggested in order to address recognition of noisy speech under stress.

3.4.1. *(Auto:I,LSP:T) enhancement*

In an earlier study (Hansen and Clements, 1991), a new form of iterative speech enhancement was developed for single channel inputs. The basis of the procedure is sequential maximum a posteriori (MAP) estimation of the speech waveform and its all-pole parameters, followed by imposition of constraints upon the sequence of speech spectra. The new approach imposes intra- and inter-frame constraints on the input speech signal to ensure more speech-like formant trajectories, reduce frame-to-frame pole jitter and effectively introduce a relaxation parameter to the iterative scheme. Intra-frame constraints are applied to the *Auto*correlation coefficients, and inter-frame constraints are applied to the *L*ine-*S*pectral-*P*air parameters (Auto-LSP). The algorithms have also been generalized and successfully tested for non-white and slowly varying noise. The current systems result in substantially improved speech quality and parameter estimation in this context with only a minor increase in computational requirements. The Auto-LSP method has also been employed for robust speech parameter estimation with application to speech coding and recognition in actual noisy environments (Hansen and Arslan, 1995a,b).

3.4.2. ACE enhancement

Building on the success of single-channel Auto-LSP, two additional iterative frequency-domain algorithms have recently been formulated, employing auditory constraints for dual-channel speech enhancement (*A*uditory *C*onstrained speech *E*nhancement techniques: ACE-I, ACE-II) (Nandkumar and Hansen, 1995; Hansen and Nandkumar, 1995). The dual-channel enhancement schemes are shown to follow the iterative Expectation-Maximization (EM) algorithm, resulting in a two-step dual-channel Wiener filtering scheme. New techniques for applying constraints during the EM iterations were developed which take advantage of auditory and perceptual properties of speech. The algorithms have been shown to produce improved global objective speech quality measures for continuous speech (TIMIT speech database) degraded by additive white Gaussian noise, aircraft cockpit noise and computer cooling-fan noise. These processors have also been shown to improve speech quality as a tandem processor for a 4.8 KBPS CELP vocoder operating in noisy environments (Nandkumar et al., 1992).

3.4.3. MCE: morphological constrained enhancement

Finally, a new speech enhancement algorithm has also been formulated which employs noise adaptive boundary detection and morphological based spectral constraints (Hansen, 1994). The technique is developed in the frequency domain, and uses a speech specific weighted subtraction factor and power exponent, followed by the application of morphological based smoothing constraints. Source generator boundary information allow the enhancement procedure to adapt and thereby track changing speech characteristics. The new method provides superior speech quality for all speech sound classes, comparable to intra and inter-frame constrained methods (Hansen and Clements, 1991), without the requirement of iterative processing.

4. Robust speech recognition

A number of studies have shown that front-end signal processing methods can improve speech recognition robustness (Lockwood et al., 1992; Lockwood and Boudy, 1992; Hansen and Bria, 1990, 1992; Hansen and Clements, 1991; Junqua, 1993). Next, we discuss three front-end processing approaches to speech recognition under stress (and/or noise). The formulated methods are based on innovative speech parameter estimation schemes which are less sensitive to varying levels and types of background noise, as well as accurate modeling of the human speech production under stress to improve recognition in adverse environments. These methods employ robust speech feature estimation algorithms, as well as stress equalization techniques based on source generator theory. A comparison of how the stress equalization methods are applied to the extracted speech feature sequence is shown in Fig. 8.

4.1. Stress equalization and noise suppression

The first front-end approach employs feature enhancement and production equalization algorithms under the framework of source generator theory (see Fig. 9). The intent here is to demonstrate that a source generator based approach can reduce the effects of stress for robust recognition in diverse environmental conditions. Therefore, though the choice of source generator type is arbitrary, hand labeled phoneme partitions were employed here (see Fig. 8(a)). The feature enhancement algorithm is formulated based on a class of constrained iterative techniques previously derived for automatic enhancement of speech in varying background noise environments. The present technique (see Section 3.4.1) employs speech specific inter and intra-frame spectral constraints applied to line-spectral-pair parameters and autocorrelation estimates. Next, a multi-dimensional stress equalization approach is formulated which produces recognition features which are suggested to be less sensitive to varying factors caused by stress and noise. The stressed based equalization domain is restricted to be the SPECTRAL domain in an eight-dimensional feature space ($\vec{k} = d_1, \ldots, d_8$).

$$\Psi_{\mathrm{SPECTRAL}(\vec{k}),c}[\gamma_j] : s_{\gamma_j}^{\mathrm{NEUTRAL}} \rightarrow s_{\gamma_j}^{\mathrm{STRESSED}}, \qquad (7)$$

$$c \in (\text{NEUTRAL, SLOW, FAST, SOFT, LOUD,}$$
$$\text{ANGRY, CLEAR, QUESTION, C50 TASK,}$$
$$\text{C70 TASK, LOMBARD}).$$

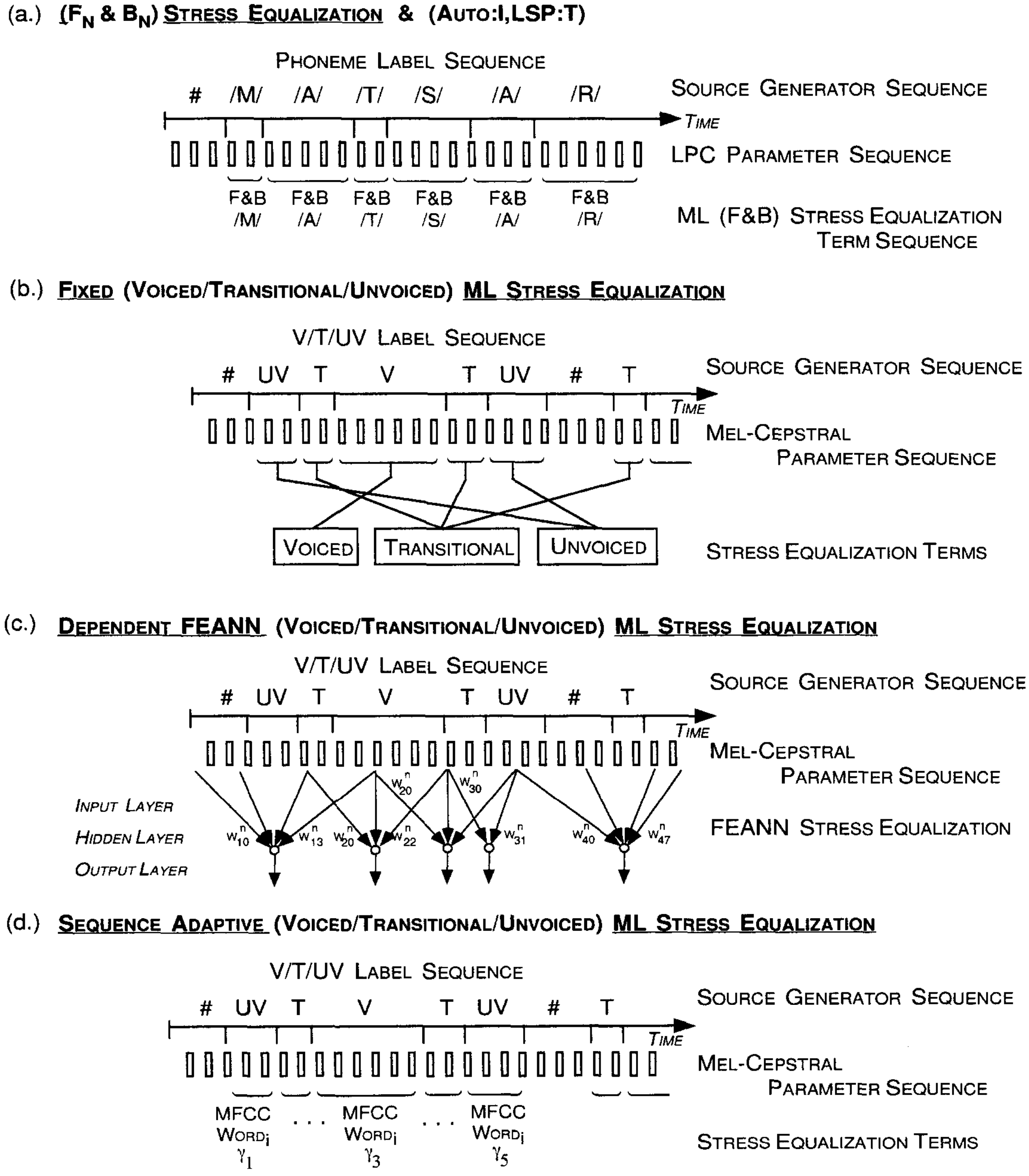

(a.) (F_N & B_N) Stress Equalization & (Auto:I,LSP:T)
Phoneme Label Sequence
# /M/ /A/ /T/ /S/ /A/ /R/
Source Generator Sequence
Time
LPC Parameter Sequence
F&B /M/
F&B /A/
F&B /T/
F&B /S/
F&B /A/
F&B /R/
ML (F&B) Stress Equalization Term Sequence
(b.) Fixed (Voiced/Transitional/Unvoiced) ML Stress Equalization
V/T/UV Label Sequence
# UV T V T UV # T
Source Generator Sequence
Time
Mel-Cepstral Parameter Sequence
Voiced
Transitional
Unvoiced
Stress Equalization Terms
(c.) Dependent FEANN (Voiced/Transitional/Unvoiced) ML Stress Equalization
V/T/UV Label Sequence
# UV T V T UV # T
Source Generator Sequence
Time
Mel-Cepstral Parameter Sequence
Input Layer
Hidden Layer
Output Layer
w^n_10
w^n_13
w^n_20
w^n_22
w^n_20
w^n_30
w^n_31
w^n_40
w^n_47
FEANN Stress Equalization
(d.) Sequence Adaptive (Voiced/Transitional/Unvoiced) ML Stress Equalization
V/T/UV Label Sequence
# UV T V T UV # T
Source Generator Sequence
Time
Mel-Cepstral Parameter Sequence
MFCC Word_i γ_1
. . .
MFCC Word_i γ_3
. . .
MFCC Word_i γ_5
Stress Equalization Terms

Fig. 8. A comparison of how stress equalization is applied to extracted speech features. (a) Stress equalization of formant location and bandwidth across a source generator phoneme sequence. (b) Stress equalization of Mel-cepstral features using fixed compensation terms across a V/T/UV source generator sequence. (c) Stress equalization of Mel-cepstral features using a word dependent feature enhancing artificial neural network (FEANN) across a V/T/UV source generator sequence. (d) Stress equalization of Mel-cepstral features using word dependent Maximum-Likelihood compensation across a V/T/UV source generator sequence.

Here, $\Psi_{\mathrm{SPECTRAL}(\vec{k}),c}[\gamma_j]$ is an eight-dimensional spectral transformation which is unique for each generator γ_j. The spectral dimensions SPECTRAL($\vec{k}$) are defined as the first four formant locations and bandwidths $(\vec{F},\vec{B})$. Stress equalization of the speech feature set is achieved using a unique maximum likelihood transformation term $\mu_{\mathrm{SPECTRAL}}(\Psi[\lambda,d_i,\gamma_j])$, which is estimated for each feature dimension d_i, stressed condition $c=\lambda$, and source generator γ_j as

$$\mu_{\mathrm{SPECTRAL}}\left(\Psi\left[\lambda,d_i,\gamma_j\right]\right) = \frac{(1/N_j)\sum_{t_n=t_1}^{t_{N_j}}\psi_{i,j}(t_n)}{(1/N_{j,\lambda})\sum_{t_n=t_{1,\lambda}}^{t_{Nj,\lambda}}\psi_{i,j}^{(\lambda)}(t_n)}. \tag{8}$$

Here, if d_1 is the first formant location F_1, then the stress equalization term for the first formant location $\mu_{\mathrm{SPECTRAL}}(\Psi[\lambda,d_1,\gamma_j])$ is found by finding the sample mean formant location under neutral conditions $(1/N_j)\sum_{t_n=t_1}^{t_{N_j}}\psi_{1,j}(t_n)$, and dividing it by the sample mean under stress condition λ, as $(1/N_{j,\lambda})\sum_{t_n=t_{1,\lambda}}^{t_{Nj,\lambda}}\psi_{1,j}^{(\lambda)}(t_n)$. This is repeated for each stress condition c. This establishes the set of stress equalization terms for the sequence of source generators for each input word in the recognition dictionary.

Next, using an HMM recognition framework, baseline scores are obtained in Table 2 for speech under neutral, stressful, noisy neutral and ten noisy stressful speaking conditions (e.g., loud, angry, computer task conditions, Lombard effect, etc). Combined stress equalization with constrained feature enhancement is shown to reduce the average word error rate for recognition of noisy stressful speech by −38.7% (mean recognition for noisy stressful speech increased from 30.3% to 57.3%). Significant improvement occurred for noisy speech under loud, angry and Lombard effect stress conditions. The tandem recognition algorithm is also shown to be more consistent across noisy stressful conditions as measured by a decrease in the standard deviation of recognition rate (from 9.1 to 5.7). Further details can be found in previous studies (Hansen, 1988; Hansen and Clements, 1989, 1995). The results suggest that the combination of a flexible source generator framework to address stressed speaking conditions, and a feature enhancement algorithm which adapts based on speech specific constraints, can be effective in reducing the consequences of stress and noise for robust automatic recognition.

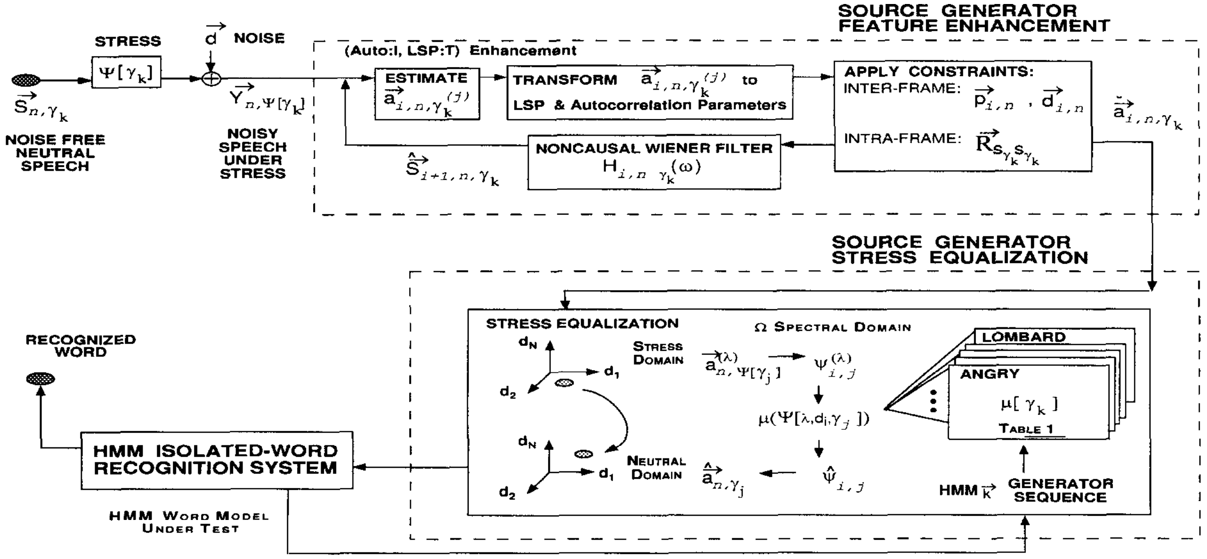

SOURCE GENERATOR
FEATURE ENHANCEMENT
STRESS
$\Psi[\gamma_k]$
$\vec{d}$ NOISE
$\vec{S}_{n,\gamma_k}$
NOISE FREE NEUTRAL SPEECH
$\vec{Y}_{n,\Psi[\gamma_k]}$
NOISY SPEECH UNDER STRESS
(Auto:I, LSP:T) Enhancement
ESTIMATE $\vec{a}_{i,n,\gamma_k}^{(j)}$
TRANSFORM $\vec{a}_{i,n,\gamma_k}^{(j)}$ to LSP & Autocorrelation Parameters
APPLY CONSTRAINTS:
INTER-FRAME: $\vec{p}_{i,n}$, $\vec{d}_{i,n}$
INTRA-FRAME: $\vec{R}_{s_{\gamma_k}s_{\gamma_k}}$
$\vec{a}_{i,n,\gamma_k}$
NONCAUSAL WIENER FILTER $H_{i,n\ \gamma_k}(\omega)$
$\hat{\vec{S}}_{i+1,n,\gamma_k}$
SOURCE GENERATOR
STRESS EQUALIZATION
STRESS EQUALIZATION
Ω SPECTRAL DOMAIN
d_N
d_1
d_2
STRESS DOMAIN
NEUTRAL DOMAIN
$\vec{a}^{(\lambda)}_{n,\Psi[\gamma_j]}$
$\psi^{(\lambda)}_{i,j}$
$\mu(\Psi[\lambda,d_i,\gamma_j])$
$\hat{\vec{a}}_{n,\gamma_j}$
$\hat{\psi}_{i,j}$
LOMBARD
ANGRY
$\mu[\gamma_k]$
TABLE 1
HMM $\vec{k}$
GENERATOR SEQUENCE
RECOGNIZED WORD
HMM ISOLATED-WORD RECOGNITION SYSTEM
HMM WORD MODEL UNDER TEST

Fig. 9. Flow diagram of (i) (Auto:I,LSP:T) constrained iterative feature enhancement, (ii) stress equalization and HMM recognition algorithm.

Table 2
Recognition performance of noisy stressful speech with combined generator enhancement and stress equalization

	Noisy stressful speech recognition results [a]											
Condition	N	Sl	F	So	L	A	C	Q	C50	C70	Lom	Avg10	StDev10
Stressful, noisy	49%	45%	28%	33%	18%	15%	40%	28%	35%	33%	28%	30.3%	9.12
with (Auto:I,LSP:T)	83%	57%	53%	43%	35%	28%	58%	55%	58%	53%	38%	47.8%	10.92
plus ($\vec{F}$), ($\vec{B}$), ($\vec{F}$ & $\vec{B}$) equalization		61%	53%	53%	61%	50%	58%	56%	55%	55%	70%	57.3%	5.69

[a] Additive white Gaussian noise, +30 dB SNR.

Stressed speech style key: N: neutral; Sl: slow; F: fast; So: soft; L: loud; A: angry; C: clearly spoken; Q: question; C50: moderate computer workload task condition; C70: high computer workload task condition; Lom: Lombard effect noise condition.

4.2. Fixed ML and FEANN stress normalization

We have seen that front-end equalization of speech under stress can improve the performance of neutral trained speech recognition algorithms. While this has been useful, the requirement of a phoneme level sequence partition prevents recognition usage in actual speech under stress environments. In order to remove this requirement, a maximum likelihood stress equalization method was formulated which normalizes input speech feature sequences using a set of fixed equalization terms (see Fig. 8(b)) (Hansen and Bria, 1990). This method assumes that input speech is parsed into a sequence of voiced/transitional/unvoiced (V/T/UV) labeled sections (Hansen, 1991), and that one of three maximum likelihood stress equalization terms are used to compensate for the effects of stress. Results show that stress compensation using three fixed V/T/UV stress equalization terms improves Lombard speech recognition performance by +10%. This method was later adapted for real-time implementation and evaluated for ten noisy stressful conditions with a +17% improvement in recognition (Hansen and Cairns, 1995).

While performing stress compensation using fixed V/T/UV equalization terms was successful, it had been observed that the impact of stress will depend on the lexical stress placed on syllables in the phoneme sequence (Hansen, 1988; Hansen and Cairns, 1995). For example, in a multi-syllable isolated word such as *degree*, the stress variations due to Lombard effect will be less for the vowel portion of the first syllable than for the second syllable. Therefore, it is desirable to perform stress equalization across the source generator sequence on a word dependent basis.

In the next approach, a feature enhancing artificial neural network (FEANN) is developed which reduces stress effects during parameterization (Clary and Hansen, 1992). Fig. 8(c) illustrates the basic approach. Here, a unique FEANN is formed for each

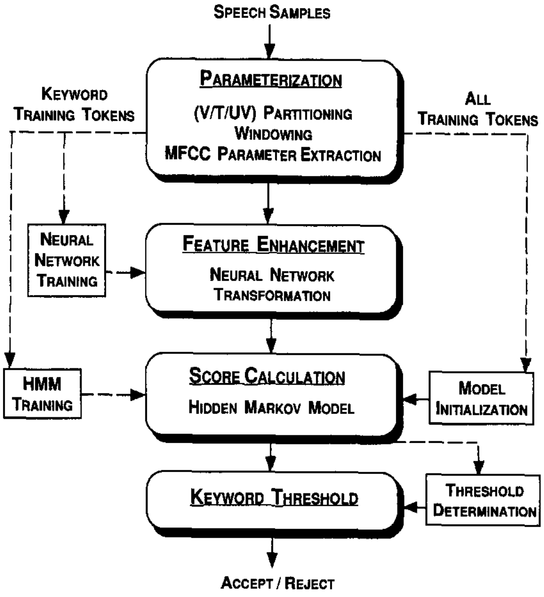


Fig. 10. A flow diagram of a stress equalization method using a feature enhancing artificial neural network (FEANN) with application to keyword recognition.

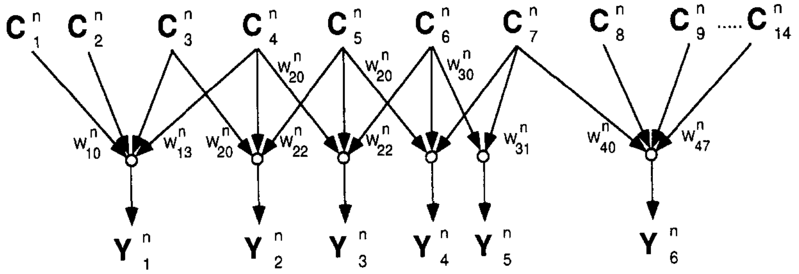


Fig. 11. A snapshot of the C^n subnetwork applied to an instance of "six". Four out of five unique subnetwork weights are pictured, W_{jk}^n, $k = 0, \ldots, M_j - 1$.

keyword model and evaluated using a semi-continuous HMM recognizer followed by a likelihood ratio test for keyword detection. The following subsections present details of each of the score-producing steps and the rationale for the likelihood ratio test (see Fig. 10).

4.2.1. Parameterization

During parameterization, the input speech is partitioned into a source generator sequence across time using a previously formulated voiced/transitional/unvoiced (v/t/uv) detector (Hansen, 1991). Endpoints are identified and v/t/uv classifications made using characteristics of frame energy curvature and noise adaptive thresholds. Nine Mel-cepstral coefficients C^n, n = 1,...,9, are extracted using Hamming window analysis on a 16 msec frame-by-frame basis for speech sampled at 8 kHz.

4.2.2. Feature enhancing artificial neural network

The design criteria of the linear feature enhancing artificial neural network is that it should have class-dependent weights, preserve information and take advantage of application-specific knowledge of the input signal. To provide class-dependent weights, the weights are determined using training tokens of the modeled class. In addition, the Karhunen–Loeve transform is used to insure that the neural network is information-preserving based on a minimum mean square error between the actual network input and the input reconstructed from the network output and weights. Finally, the width of the input layer of the neural network adapts as characteristics of the speech signal vary and new segment types are encountered. Segment types are classified in the parameterization step as v/t/uv.

A time sequence of vectors, each consisting of nine mel-cepstral coefficients, provides the input to the neural network and is linearly transformed by sets of weights. Each NT-MFCC [6] time series is transformed by a subnetwork, which "slides" across the input frames (Fig. 8(c)). The size of a subnetwork's input layer depends on segment type. At a particular instant in time, all subnetworks have the same input layer width, but different weights. As long as the segment type remains the same, the input layer width remains the same. The input layer widths are chosen based on how fast the Mel-cepstral coefficients change in a given segment type and how often the type occurs. Parameters change more slowly in a voiced section; thus, the largest input layer width is chosen for voiced segments. Fig. 11 shows how the input layer width of the neural network changes as new segments are encountered for a single MFCC. Note that all 9 MFCCs undergo a transformation, but only coefficient n, C_i^n are pictured here, where i is the frame number. The resulting transform coefficient at time t is denoted by Y_t^n. The network output is (assuming a mapping from i onto t)

$$Y_t^n = \sum_{k=0}^{M_j - 1} W_{jk}^n * C_{i+k}^n, \tag{9}$$

where the segment at time t is of type j and M_j is the corresponding input layer width.

To determine the sets of weights for the neural

[6] The notation NT-MFCC refers to non-transformed vectors of Mel-cepstral coefficients. The notation T-MFCC is used to represent Mel-cepstral coefficients which have been transformed by a FEANN.

network, sample correlation matrices are formed from training data for each coefficient n and segment type j, and the principal eigenvector is found for each matrix. Sample correlation matrices are formed as follows:

$$\overline{Q}_j^n = \frac{1}{N_j} \sum_{i \in j} \left(\left\{ C_i^n, \ldots, C_{i+M_j-1}^n \right\}^{\mathrm{T}} \left\{ C_i^n, \ldots, C_{i+M_j-1}^n \right\} \right), \tag{10}$$

where N_j is the number of training samples for the jth segment type. For the subnetwork in Fig. 11, the following error quantity is therefore minimized:

$$E_j^n = \sum_{i,t \in j} E_{jit}^n = \sum_{i,t \in j} \left(\left(W_{jG(i,t)}^n * Y_t^n \right) - C_i^n \right)^2, \tag{11}$$

where $G(i,t)$ maps frame i and time t onto the weight index corresponding to i and t. Although this work was motivated in part by iterative algorithms which implement the Karhunen–Loeve transform, the Jacobi method is used here to find the principal eigenvector.

4.2.3. Semi-continuous HMM recognition

As shown in Fig. 10, five state semi-continuous hidden Markov models (SCHMM) are used to model each keyword (Huang et al., 1990). The mixture weighting coefficients $\mathscr{C}$ are unique for each state and each mixture density. The probability density function for state i is

$$b_i(x) = \sum_{j=1}^{64} f_j(x) \mathscr{C}_{ij}, \tag{12}$$

where $f_j(x) = N(\{x, \mu_j, \Sigma_j)$ is a multivariate Gaussian density with mean vector μ_j and covariance matrix Σ_j, from a codebook of $J = 64$ Gaussian densities. Model parameter re-estimation for training the SCHMM is accomplished via the Baum–Welch forward–backward algorithm. Finally, the word score is calculated for an observation sequence by finding the mean natural log of the forward variable over all frames and all states:

$$\text{score} = \frac{1}{T} \sum_{t=1}^{T} \ln\left(\sum_i \alpha_t(i) \right). \tag{13}$$

4.2.4. Likelihood ratio test

For a recognizer to "detect" a keyword, the score produced must be greater than a pre-determined threshold. Therefore, performance which depends on the threshold is measured in terms of probability of detection p_{d} and probability of false alarm p_{f} (Whalen, 1971).

One goal is to determine the impact of stress equalization using a FEANN for keyword recognition. Initial recognizers are evaluated by setting their thresholds to the minimum score produced by keywords for training data. Although this is not an automatic method for selecting thresholds, it serves to demonstrate the rejection benefits of the feature enhancing neural network. In the discussion below, this threshold is used to determine "theoretically best" results.

The optimal detection scheme is based on a likelihood ratio test. Hypothesis one (H1) is that the submitted word is the desired keyword. Hypothesis zero (H0) is that the word submitted is not the keyword represented by the recognizer. A decision rule can be determined by minimizing the Bayes average cost. For this purpose, the a priori probabilities are assumed to be equal. The decision rule is, if

$$\frac{p_1(y)}{p_0(y)} \geq \frac{C_{10}}{C_{01}} \tag{14}$$

choose H1, where C_{10} is the cost of choosing H1 when the correct decision is H0, C_{01} is the cost of choosing H0 when the correct decision is H1, and $p_1(y)/p_0(y)$ is the likelihood ratio.

To find $\mathrm{p}_1(y)$ and $p_0(y)$, Maxwell probability density functions (pdfs) are fit to sample pdfs obtained from scores under each hypothesis. The Maxwell pdf is formed as follows:

$$f(x) = \frac{\sqrt{2}\,x^2}{\alpha^3 \sqrt{\pi}} \mathrm{e}^{-x^2/2\alpha^2} \quad \text{with mean } \mu = 2\alpha \sqrt{\frac{2}{\pi}}, \tag{15}$$

which yields the following probability density function:

$$f_x(x_o) = \frac{\mu x_o^2}{2\alpha^4} \mathrm{e}^{-\sqrt{2}\,x_o^2/\alpha\mu\sqrt{\pi}}. \tag{16}$$

4.2.5. FEANN evaluations

A series of keyword recognition evaluations were performed using speaker dependent and multi-speaker FEANNs for neutral and Lombard effect speech recognition.

"Theoretically Best" evaluations are used to show

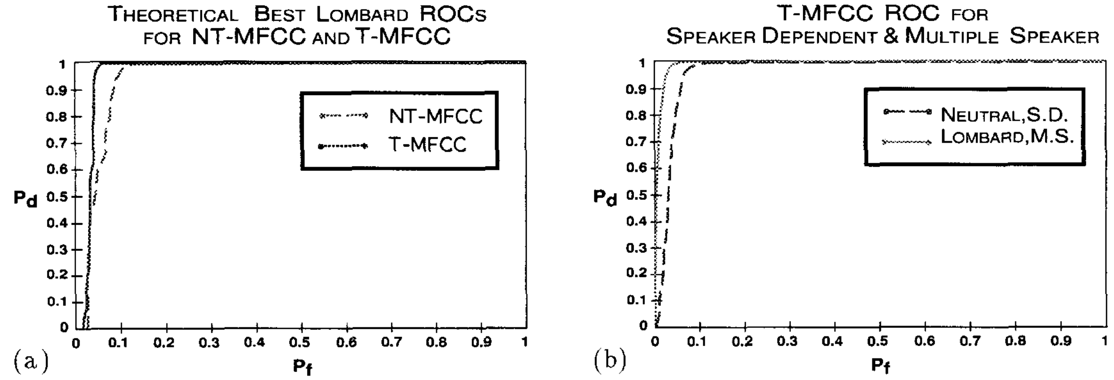


Fig. 12. (a) Theoretically best receiver operating characteristics for NT-MFCC and T-MFCC (FEANN) "break" keyword recognizers. (b) Receiver operating characteristic for the speaker dependent (S.D.) and multiple speaker (M.S.) Lombard effect T-MFCC (FEANN) "break" keyword recognizers.

that FEANN reduces the number of incorrectly accepted tokens for a recognizer for "help" for Lombard effect speech. Results show that FEANN reduced the number of incorrectly accepted tokens for "break" for the neutral case by $\frac{2}{3}$ for training data and nearly $\frac{1}{2}$ for test data. Fig. 12(a) shows receiver operating characteristics (ROCs) formed using Lombard test data for the keyword "break" for NT-MFCC (dotted line) and T-MFCC (solid line) recognizers. Detection and false alarm probabilities are also summarized in Table 3 for neutral trained keyword recognizers for both non-transform (NT-MFCC) and FEANN transformed (T-MFCC) input parameters. The results show that the recognizer which used FEANN stress equalization made no false acceptances.

Multiple speaker Lombard effect results for "break" are presented in Table 4. Both recognizers were trained using data from 9 speakers. The results show improved rejection versus the speaker dependent case, but that T-MFCC performance was lower than NT-MFCC. These results suggest two observations: first that additional training data does improve performance, and second that the intra-speaker variability under Lombard effect is significant and must either require speaker dependent stress equalization, or an adaptive FEANN across speakers.

In the last evaluation, a likelihood ratio test was added to both the speaker dependent T-MFCC "break" recognizer and the multiple speaker Lombard effect T-MFCC "break" recognizer. Sample Maxwell probability density functions (PDFs) were estimated by finding sample means and values of α corresponding to the optimal (in the least mean square sense) PDFs using a simulated annealing algorithm. Two sets of "best fit" PDFs are shown in Fig. 13 for neutral "break" recognizers. The FEANN has the effect of increasing the variance of scores under H0, causing the PDFs under each hypothesis to "separate" more for the T-MFCC recognizer. The increased separation yields improved performance.

A "semi-open" ROC for each T-MFCC recognizer was obtained by varying the threshold of the

Table 3
Detection and false alarm probabilities for two "break" keyword recognizers with thresholds obtained for theoretically best performance

Noise-free keyword detection evaluation results for "break"				
	Neutral data		Lombard data	
Recognizer type	p_d	p_f	p_d	p_f
NT-MFCC	1.0	0.0882	1.0	0.0149
T-MFCC	1.0	0.0294	1.0	0.0

Table 4
Detection and false alarm probabilities for two Lombard effect "break" keyword recognizers with thresholds set to show theoretically best possible performance

Multiple speaker Lombard effect evaluation results for "break"				
	Training data		Testing data	
Recognizer type	p_d	p_f	p_d	p_f
NT-MFCC	1.0	0.0	1.0	0.0133
T-MFCC	1.0	0.0	1.0	0.0167

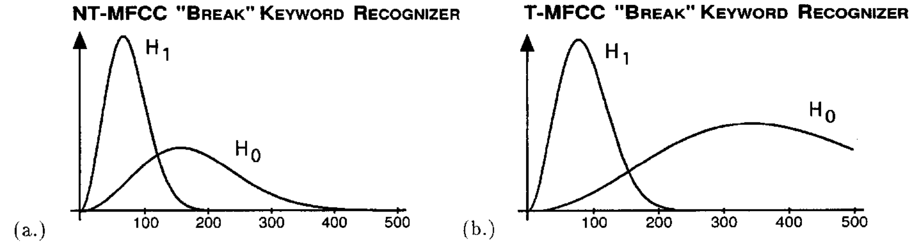


Fig. 13. (a) Probability density functions for the speaker dependent neutral NT-MCC "break" recognizer for both hypotheses. (b) Probability density functions for the speaker dependent neutral T-MCC "break" recognizer for both hypotheses.

likelihood ratio test (see Fig. 12(b)). The speaker dependent ROC follows closely the ROC pictured in Fig. 12(a) for the T-MFCC recognizer. The fact that the ROC obtained by applying a likelihood ratio test closely matches the "theoretically best" ROC verifies that reliable PDFs can be formed from training data. For the multiple speaker recognizer, performance is near the theoretically best possible, as is shown by the multiple speaker ROC.

In this section, we have shown that a keyword-dependent neural network is able to enhance MFCC speech parameters under stress and reduce the probability of false acceptances of non-keywords by adapting its weights and input layer width based on extracted speech characteristics. Keyword recognition evaluations show that FEANN reduces the number of false acceptances for neutral and Lombard stress by more than $\frac{1}{3}$.

4.3. MCE-ACC robust recognition

The two previous methods demonstrate that improved speech recognition can be achieved using a source generator framework with stress equalization on formant or MFCC spectral parameters. In this section, robust speech recognition is accomplished via morphological constrained feature enhancement (MCE) and stress compensation which is unique for each source generator across a stressed speaking class (see Fig. 8(d)) (Hansen, 1994). The algorithm uses a noise adaptive (v/t/uv) boundary detector (Hansen, 1991) to obtain a sequence of source generator classes, which is used to direct MCE parameter enhancement (Section 3.4.3) and stress compensation. This allows the parameter enhancement and stress compensation schemes to adapt to changing speech generator types. Fig. 14 illustrates a block diagram of the algorithm, entitled Morphological Constrained feature Enhancement with Adaptive mel-Cepstral Compensation based HMM recognition (MCE-ACC-HMM). The source generator sequence of MCE estimated spectral responses $\hat{S}_{\gamma_{b_j},\alpha,\beta,\mathscr{D}_g}(\omega_i)$, are then submitted for stress equalization. Stressed speaking conditions are addressed by the choice of a modified source generator for each phoneme-like section. Let the estimated speech vector under noisy neutral and Lombard stress condition be written as $\hat{\vec{s}}_{\gamma_{b_j}}(t_n)$ and $\hat{\vec{s}}_{\Psi[\gamma_j]_i}(t_n)$, respectively, where $\Psi[\cdot]_i$ represents a stressed based change in the source generator. The sequence of Mel-cepstral (MFCC) vectors for generator γ_{b_j} under Lombard effect stress is modeled as

$$\vec{C}_{\Psi[\gamma_{b_j}]_i}(t_n) = \vec{C}_{\gamma_{b_j}}(t_n) + \vec{C}_{\Psi_i(b_j)}: \quad t_n \epsilon \left[1, N_{b_j}\right], \tag{17}$$

where $\vec{C}_{\Psi_i(b_j)}$ represents an additive stress component which depends on the particular stress class Ψ_i and source generator b_j. Given an estimate of the MFCCs over time t_n, and stress component $C_{\Psi_i(b_j)}(k)$, the log-likelihood estimate of $\vec{C}_{\Psi[\gamma_{b_j}]_i}(t_n)$ can be found. The unknown model parameter $C_{\Psi_i(b_j)}(k)$ is estimated by maximizing the log-likelihood function, resulting in the ML estimate

$$\hat{C}_{\Psi_i(b_j)}(k) = \frac{1}{N_{b_j}} \sum_{t_n=t_1}^{t_{N_{b_j}}} C_{\Psi[\gamma_{b_j}]_i} - \frac{1}{N_{b_j}} \sum_{t_n=t_1}^{t_{N_{b_j}}} C_{\gamma_{b_j}}. \tag{18}$$

A compensation model vector $\hat{\vec{C}}_{\Psi_i(b_j)}$ is estimated for each detected source generator section during HMM training, and applied during recognition evaluation. An HMM system which includes a phonetic

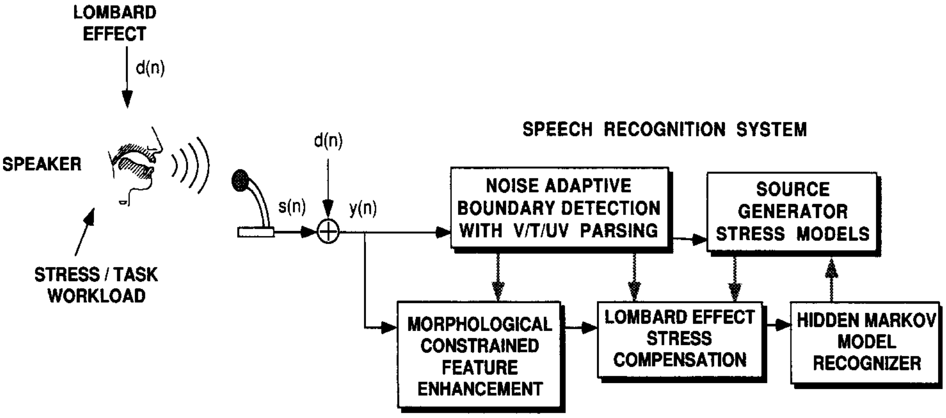

LOMBARD EFFECT
d(n)
SPEAKER
STRESS / TASK WORKLOAD
d(n)
s(n)
y(n)
SPEECH RECOGNITION SYSTEM
NOISE ADAPTIVE BOUNDARY DETECTION WITH V/T/UV PARSING
SOURCE GENERATOR STRESS MODELS
MORPHOLOGICAL CONSTRAINED FEATURE ENHANCEMENT
LOMBARD EFFECT STRESS COMPENSATION
HIDDEN MARKOV MODEL RECOGNIZER

Fig. 14. A general speech framework for noise and Lombard effect, and the resulting processing employed by the MCE-ACC-HMM speech recognition algorithm.

consistency rule is used for recognition. This rule is obtained from input (v/t/uv) generator duration models for each word, and partitions utterances into single and multi-syllabic classes prior to HMM recognition.

The algorithm was evaluated for noise free and nine noisy Lombard speech conditions which include additive white Gaussian, slowly varying computer fan, and aircraft cockpit noise (Hansen, 1994). System performance was compared to a traditional VQ-HMM recognizer with no embellishments (Table 5). Employing individual recognition scores for all 27 noisy Lombard effect stress conditions, the final mean recognition rate increased from 36.7% for VQ-HMM to 74.7% for MCE-ACC (+38% improvement). The MCE-ACC is also shown to be more consistent, as demonstrated by a decrease in standard deviation of recognition from 21.1 to 11.9, and a reduction in confusable word-pairs.

5. Summary and conclusions

In this paper, we have discussed the problem of analysis, modeling and recognition of speech under stress, noise and *Lombard* effect. A source generator framework was proposed in order to characterize speech production under stressed speaking conditions. Furthermore, we briefly discussed results from previous analysis of speech under simulated and actual stress (SUSAS). This study consisted of speech production parameters from five domains: pitch, du-

Table 5
Overall recognition results for the VQ-HMM recognizer and the new robust recognizer MCE-ACC-HMM for three types of noise. Noise-free, averages over all noisy conditions (10, 20, 30 dB SNR), and the standard deviation of noisy recognition rates are also shown

Overall noise-free and noisy Lombard effect recognition performance										
	Noise-free		Noisy Lombard conditions						Overall	
Speech & recognizer	$\bar{x}$	σ	WGN		Aircraft		PS-2 Fan		Noisy Lombard	
Neutral & VQ-HMM	96.0%	6.1	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}_{\mathrm{RECOG}}$	σ_{RECOG}
Lombard & VQ-HMM	65.7%	19.9	25.7%	19.0	46.2%	20.1	38.4%	20.9	36.7%	21.1
Lombard & MCE-ACC-HMM	86.7%	8.7	70.1%	11.6	76.3%	12.8	77.8%	11.1	74.7%	11.9

ration, intensity, glottal source and vocal tract spectrum. Stressed speaking styles included soft, loud, slow, fast, angry, clear, question, computer workload tasks, Lombard effect and actual motion-fear tasks. Next, several recently formulated enhancement algorithms were briefly reviewed for robust feature estimation. Three robust speech recognition techniques were then discussed which are based on source generator theory. These methods include (i) constrained feature enhancement with formant based stress equalization, (ii) feature enhancing artificial neural network based stress equalization for keyword recognition, and (iii) morphological constrained feature enhancement with adaptive cepstral compensation for recognition in noise and stress. Improvement was demonstrated over traditional HMM based methods. These results show that the use of a flexible source generator framework for robust front-end feature enhancement and stress equalization can contribute significantly to improved recognition performance in a variety of adverse conditions.

Acknowledgements

I wish to extend thanks to the NATO Research Study Group on Speech Processing (RSG.10), and especially Drs. R. Moore (DRA), I. Trancoso (INESC) and J. Cupples (RADC) for organizing the ESCA-NATO Tutorial and Research Workshop on Speech Under Stress, Lisbon, Portugal, 14–15 September 1995. I also wish to acknowledge a number of my graduate students who have contributed in different ways to our efforts over the years in speech recognition in noisy stressful conditions: S. Bou-Ghazale, O. Bria, D. Cairns, G. Clary, B. Womack and L. Yao.

Appendix A

A brief audio demonstration of speech data from SUSAS is available at http://www.elsevier.nl/locate/specom. The demonstration consists of two parts.

Part 1. Simulated Speech Under Stress. Male speaker speaking the word "nav" (short for "navigation") and "help" under the following stressed speech styles: Neutral, Fast, Slow, Loud, Soft, Angry, Question, Clear, Moderate Computer Response Workload Task, Heavy Computer Response Workload Task, and Lombard Effect (85 dB SPL Pink Noise).

(1a) Word: "Nav"
File: nav-NFSLWAQC57LM.au
(1b) Word: "Help"
File: help-NFSLWAQC57LM.au

Part 2. Actual Speech Under Stress. Male and female speakers producing speech under a G-force Motion/Fear Task (i.e., speakers on amusement park roller-coaster ride).

(2a) Female Speaker. In Vocabulary examples (from 35 word vocabulary)
Words: "degree eighty"
Neutral, Stressed
File: degree_eighty_F_NeuAct.au
(2b) Male Speaker. In Vocabulary examples (from 35 word vocabulary)
Words: "degree histogram"
Neutral, Stressed
File: degree_histogram_M_NeuAct.au
(2c) Out of Vocabulary examples (words speakers produced outside 35 word vocabulary)
Words: "pilot helpme"
Neutral, Stressed (male speaker)
File: pilot_helpme_M_NeuAct.au
Words: "mayday"
Neutral, Stressed (male and female speakers)
File: mayday_MF_NeuAct.au

References

P. Alexandre and P. Lockwood (1993), "Root cepstral analysis: A unified view. Application to speech processing in car noise environments", *Speech Communication*, Vol. 12, No. 3, pp. 277–288.

Z.S. Bond and T.J. Moore (1990), "A note on loud and Lombard speech", *ICSLP-90: Internat. Conf. on Spoken Language Process.*, pp. 969–972.

S.E. Bou-Ghazale and J.H.L. Hansen (1994), "Duration and spectral based stress token generation for HMM speech recognition under stress", *IEEE Internat. Conf. Acoust. Speech Signal Process.-94*, pp. 413–416.

S.E. Bou-Ghazale and J.H.L. Hansen (1995), "Improving recognition and synthesis of stressed speech via feature perturbation in a source generator framework", *ECSA-NATO Proc. Speech*

Under Stress Workshop, Lisbon, Portugal, September 1995, pp. 45–48.

D.A. Cairns and J.H.L. Hansen (1994), "Nonlinear analysis and detection of speech under stressed conditions", *J. Acoust. Soc. Amer.*, Vol. 96, No. 6, pp. 3392–3400.

B. Carlson and M. Clements (1992), "Speech recognition in noise using a projection-based likelihood measure for mixture density HMM's", *IEEE Internat. Conf. Acoust. Speech Signal Process.-92*, pp. 237–240.

Y. Chen (1988), "Cepstral domain talker stress compensation for robust speech recognition", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 36, pp. 433–439.

G.J. Clary and J.H.L. Hansen (1992), "A novel speech recognizer for keyword spotting", *ICSLP-92: Internat. Conf. on Spoken Language Process.*, pp. 13–16.

K.E. Cummings and M.A. Clements (1990), "Analysis of glottal waveforms across stress styles", *IEEE Internat. Conf. Acoust. Speech Signal Process.-90*, pp. 369–372.

J.K. Darby (1981), *Speech Evaluation in Psychiatry* (Grune & Stratton, New York).

B.A. Dautrich, L.R. Rabiner and T.B. Martin (1983), "On the effects of varying filter bank parameters on isolated word recognition", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 31, pp. 793–806.

Y. Ephraim (1992), "Statistical-model based speech enhancement systems", *Proc. IEEE*, Vol. 80, pp. 1526–1555.

M. Flack (1918), Flying stress, Medical Research Committee, London.

M. Gales and S. Young (1995), "Robust speech recognition in additive and convolutive noise using parallel model combination", *Computer Speech and Language*, Vol. 9, pp. 289–307.

M.B. Gardner (1966), "Effect of noise system gain, and assigned task on talking levels in loudspeaker communication", *J. Acoust. Soc. Amer.*, Vol. 40, pp. 955–965.

L. Goldberger and S. Breznitz (1982), *Handbook of Stress: Theoretical and Clinical Aspects* (Free Press/Macmillan, New York).

C.N. Hanley and D.G. Harvey (1965), "Quantifying the Lombard effect", *J. Hearing and Speech Disorders*, Vol. 30, pp. 274–277.

J.H.L. Hansen (1988), Analysis and compensation of stressed and noisy speech with application to robust automatic recognition, Ph.D. Thesis, Georgia Inst. of Technology, Atlanta, GA, 428 pp.

J.H.L. Hansen (1989), "Evaluation of acoustic correlates of speech under stress for robust speech recognition", *IEEE Proc. 15th Northeast Bioengineering Conf.*, Boston, MA, pp. 31–32.

J.H.L. Hansen (1991), "A new speech enhancement algorithm employing acoustic endpoint detection and morphological based spectral constraints", *IEEE Internat. Conf. Acoust. Speech Signal Process.-91*, pp. 901–904.

J.H.L. Hansen (1993), "Adaptive source generator compensation and enhancement for speech recognition in noisy stressful environments", *IEEE Internat. Conf. Acoust. Speech Signal Process.-93*, pp. 95–98.

J.H.L. Hansen (1994), "Morphological constrained enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect", *IEEE Trans. Speech Audio Process.*, Special Issue: Robust Speech Recognition, Vol. 2, No. 4, pp. 598–614.

J.H.L. Hansen (1995), "A source generator framework for analysis of acoustic correlates of speech under stress. Part I: Pitch, duration, and intensity effects", submitted to *J. Acoust. Soc. Amer.*, 44 pp. (also: Robust Speech Proc. Lab. Report RSPL-95-31, Dept. of Electrical Engineering, Duke Univ., 1995).

J.H.L. Hansen and L.M. Arslan (1995a), "Robust feature-estimation and objective quality assessment for noisy speech recognition using the credit card (CCDATA) corpus", *IEEE Trans. Speech Audio Process.*, Vol. 3, No. 3, pp. 169–184.

J.H.L. Hansen and L.M. Arslan (1995b), "Markov model based phoneme class partitioning for improved constrained iterative speech enhancement", *IEEE Trans. Speech Audio Process.*, Vol. 3, No. 1, pp. 98–104.

J.H.L. Hansen and O.N. Bria (1990), "Lombard effect compensation for robust automatic speech recognition in noise", *ICSLP-90: Internat. Conf. on Spoken Language Process.*, Kobe, Japan, pp. 1125–1128.

J.H.L. Hansen and O.N. Bria (1992), "Improved automatic speech recognition in noise and Lombard effect", in: J. Vandewalle et al., Eds., *Signal Processing VI: Theories and Applications* (Elsevier, Amsterdam), pp. 403–406.

J.H.L. Hansen and D.A. Cairns (1995), "ICARUS: Source generator based real-time recognition of speech in noisy stressful and Lombard effect environments", *Speech Communication*, Vol. 16, No. 4, pp. 391–422.

J.H.L. Hansen and M.A. Clements (1987), "Evaluation of speech under stress and emotional conditions", *Proc. Acoust. Soc. Amer.*, H15, Vol. 82 (Fall Sup.), S17.

J.H.L. Hansen and M.A. Clements (1989), "Stress compensation and noise reduction algorithms for robust speech recognition", *IEEE Internat. Conf. Acoust. Speech Signal Process.-89*, pp. 266–269.

J.H.L. Hansen and M.A. Clements (1991), "Constrained iterative speech enhancement with application to speech recognition", *IEEE Trans. Signal Process.*, Vol. 39, pp. 795–805.

J.H.L. Hansen and M.A. Clements (1995), "Source generator equalization and enhancement of spectral properties for robust speech recognition in noise and stress", *IEEE Trans. Speech Audio Process.*, Vol. 3, No. 5, pp. 407–415.

J.H.L. Hansen and S. Nandkumar (1995), "Robust estimation of speech in noisy backgrounds based on aspects of the auditory process", *J. Acoust. Soc. Amer.*, Vol. 97, No. 6, pp. 3833–3849.

J.H.L. Hansen and B.D. Womack (1996), "Feature analysis and neural network based classification of speech under stress", *IEEE Trans. Speech Audio Process.*, Vol. 4, No. 4, pp. 307–313.

B.A. Hanson and T. Applebaum (1990), "Robust speaker-independent word recognition using instantaneous, dynamic and acceleration features: Experiments with Lombard and noisy speech", *IEEE Internat. Conf. Acoust. Speech Signal Process.-90*, pp. 857–860.

B.A. Hanson and T. Applebaum (1993), "Subband or cepstral domain filtering for recognition of Lombard and channel-distorted speech", *IEEE Internat. Conf. Acoust. Speech Signal Process.-93*, Vol. 2, pp. 79–82.

M.H.L. Hecker, K.N., Stevens, G. von Bismarck and C.E. Williams (1968), "Manifestations of task-induced stress in the acoustic speech signal", *J. Acoust. Soc. Amer.*, Vol. 44, No. 4, pp. 993–1001.

H. Hermansky and N. Morgan (1994), "RASTA processing of speech", *IEEE Trans. Speech Audio Process.*, Special Issue: Robust Speech Recognition, Vol. 2, No. 4, pp. 578–689.

H. Hermansky, N. Morgan and H.G. Hirsch (1993), "Recognition of speech in additive and convolutional noise based on RASTA spectral processing", *IEEE Internat. Conf. Acoust. Speech Signal Process.-93*, pp. 83–86.

J.W. Hicks and H. Hollien (1981), "The reflection of stress in voice – 1: Understanding the basic correlates", *1981 Carnahan Conf. on Crime Countermeasures*, pp. 189–195.

X.D. Huang, Y. Ariki and M.A. Jack (1990), *Hidden Markov Models for Speech Recognition* (Edinburgh Univ. Press, Edinburgh).

M.J. Hunt and C. Lefebvre (1989), "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech", *IEEE Internat. Conf. Acoust. Speech Signal Process.-89*, pp. 262–265.

B.H. Juang (1991), "Speech recognition in adverse environments", *Computer, Speech and Language*, pp. 275–294.

J.C. Junqua (1993), "The Lombard reflex and its role on human listeners and automatic speech recognizers", *J. Acoust. Soc. Amer.*, Vol. 1, pp. 510–524.

I. Kuroda, O. Fujiwara, N. Okamura and N. Utsuki (1976), "Method for determining pilot stress through analysis of voice communication", *Aviation, Space, and Env. Med.*, Vol. 5, pp. 528–533.

P. Lieberman and S. Michaels (1962), "Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech", *J. Acoust. Soc. Amer.*, Vol. 34, No. 7, pp. 922–927.

R.P. Lippmann, E.A. Martin and D.B. Paul (1987), "Multi-style training for robust isolated-word speech recognition", *IEEE Internat. Conf. Acoust. Speech Signal Process.-87*, pp. 705–708.

S. Lively, D. Pisoni, W. van Summers and R. Bernacki (1993), "Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences", *J. Acoust. Soc. Amer.*, Vol. 93, No. 5, pp. 2962–2973.

F.H. Liu, A. Acero and R.M. Stern (1992), "Efficient joint compensation of speech for the effects of additive noise and linear filtering", *IEEE Internat. Conf. Acoust. Speech Signal Process.-92*, pp. 257–260.

P. Lockwood and J. Boudy (1992), "Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars", *Speech Communication*, Vol. 11, Nos. 2-3, pp. 215–228.

P. Lockwood, J. Boudy and M. Blanchet (1992), "Non-linear spectral subtraction (NSS) and hidden Markov models for robust speech recognition in car noise environments", *IEEE Internat. Conf. Acoust. Speech Signal Process.-92*, pp. 265–268.

E. Lombard (1911), "Le signe de l'élévation de la voix", *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, Vol. 37, pp. 101–119.

D. Mansour and B.H. Juang (1988), "A family of distortion measures based upon projection operation for robust speech recognition", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 37, pp. 1659–1671.

C.E. Mokbel and G.F. Chollet (1995), "Automatic word recognition in cars", *IEEE Trans. Speech Audio Process.*, Vol. 3, No. 5, pp. 346–356.

I.R. Murray (1993), "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion", *J. Acoust. Soc. Amer.*, Vol. 93, pp. 1097–1108.

S. Nandkumar and J.H.L. Hansen (1995), "Dual-channel iterative speech enhancement with constraints based on an auditory spectrum", *IEEE Trans. Speech Audio Process.*, Vol. 3, No. 1, pp. 22–34.

S. Nandkumar, J.H.L. Hansen and R. Stets (1992), "A new dual-channel speech enhancement technique with application to CELP coding in noise", *ICSLP-92, Internat. Conf. on Spoken Language Process.*, Alberta, Canada, pp. 527–530.

D.B. Paul (1987), "A speaker-stress resistant HMM isolated word recognizer", *IEEE Internat. Conf. Acoust. Speech Signal Process.-87*, pp. 713–716.

D.B. Pisoni, R.H. Bernacki, H.C. Nusbaum and M. Yuchtman (1985), "Some acoustic-phonetic correlates of speech produced in noise", *IEEE Internat. Conf. Acoust. Speech Signal Process.-85*, pp. 41.10.1–4.

P.K. Rajasekaran, G.R. Doddington and J.W. Picone (1986), "Recognition of speech under stress and in noise", *IEEE Internat. Conf. Acoust. Speech Signal Process.-86*, pp. 733–736.

M.J. Russell and R.K. Moore (1983), Recordings made for automatic speech recognition assessment and research, Royal Signals and Radar Est. Tech. Report, AD-A146 824.

P.V. Simonov and M.V. Frolov (1977), "Analysis of the human voice as a method of controlling emotional state: Achievements and goals", *Aviation, Space, and Env. Sci.*, Vol. 1, pp. 23–25.

B.J. Stanton, L.H. Jamieson and G.D. Allen (1988), "Acoustic-phonetic analysis of loud and Lombard speech in simulated cockpit conditions", *IEEE Internat. Conf. Acoust. Speech Signal Process.-88*, pp. 331–334.

B.J. Stanton, L.H. Jamieson and G.D. Allen (1989), "Robust recognition of loud and Lombard speech in the fighter cockpit environment", *IEEE Internat. Conf. Acoust. Speech Signal Process.-89*, pp. 675–678.

L.A. Streeter, N.H. Macdonald, W. Apple, R.M. Krauss and K.M. Galotti (1983), "Acoustic and perceptual indicators of emotional stress", *J. Acoust. Soc. Amer.*, Vol. 73, No. 4, pp. 1354–1360.

W.V. Summers, D.B. Pisoni, R.H. Bernacki, R.I. Pedlow and M.A. Stokes (1988), "Effects of noise on speech production: Acoustic and perceptual analyses", *J. Acoust. Soc. Amer.*, Vol. 84, pp. 917–928.

A.D. Whalen (1971), *Detection of Signals in Noise* (Academic Press, New York).

M. Wang and S. Young (1992), "Speech recognition using hidden Markov model decomposition and a general background speech model", *IEEE Internat. Conf. Acoust. Speech Signal Process.-92*, pp. 253–256.

C.E. Williams and K.N. Stevens (1969), "On determining the emotional state of pilots during flight: an exploratory study", *Aerospace Medicine*, Vol. 40, pp. 1369–1372.

C.E. Williams and K.N. Stevens (1972), "Emotions and speech: Some acoustic correlates", *J. Acoust. Soc. Amer.*, Vol. 52, No. 4, pp. 1238–1250.