ELSEVIER

Speech Communication 16 (1995) 165–173

SPEECH COMMUNICATION

Acoustic characteristics of speaker individuality: Control and conversion

Hisao Kuwabara [a,*], Yoshinori Sagisaka [b]

[a] *Department of Electronics and Information Science, The Nishi-Tokyo Univ., 2525 Yatsuzawa, Uenohara-machi Kitatsuru-gun, Yamanashi, 409-01, Japan*

[b] *ATR Interpreting Telecommunications Laboratories, Hikaridai 2-2, Seikacho, Soraku-gun, Kyoto 619-02, Japan*

Received 20 May 1994; revised 8 August 1994 and 19 October 1994

Abstract

This paper introduces some recent studies on voice quality control and conversion technologies. After briefly summarizing some basic scientific findings on the acoustic correlates of speech individuality, we review the latest developments in speech technologies related to voice control and speaker characteristic copying. The main focus is on a survey of non-parametric methods for spectral segmental characteristics mapping between speakers, introducing some different types of spectral mapping methods that have evolved in relation to the speaker adaptation techniques being developed in speech recognition research.

Zusammenfassung

Dieser Artikel beschreibt einige neuere Studien von Stimmqualitäts-kontroll und -umwandlungstechnologien. Nachdem wir kurz einige fundamentale wissenschaftliche Erkenntnisse der akustischen Korrelationen der Sprecherindividualität zusammengefassen, beschreiben wir die letzten sprachtechnologischen Entwicklungen in Verbindung mit Stimmkontrolle und sprechercharakteristischem Kopieren. Der Schwerpunkt liegt in einer Zusammenfassung der nicht-parametrischen Methoden für spektralsegmentelle Karachteristikenabbildungen zwischen Sprechern, wobei wir einige verschiedene Spektralabbildungen die sich in Verbindung mit Sprecheradaptionsmethoden in der Spracherkennung entwickelt haben vorstellen.

Résumé

Cet article expose de récentes recherches sur les techniques de contrôle de la qualité ainsi que de conversion de la voix. Après un bref rappel de quelques résultats scientifiques de base sur les corrélations acoustiques dans le caractère individuel de la voix, nous décrivons les dernières techniques de traitement de la parole en matière de contrôle de la voix et de reproduction des caractéristiques du locuteur. Nous nous concentrons plus particulièrement

* Corresponding author.

SSDI 0167-6393(94)00053-0

sur une description des méthodes non paramètriques d'identification spectrale segmentée des caractéristiques des différents locuteurs et nous introduisons á ce sujet plusieurs types d'identifications spectrales, qui ont évolué de pair avec les techniques d'adaptation au locuteur développées en reconnaissance de la parole.

Keywords: Speaker characteristics; Voice conversion; Spectral mapping; Speech synthesis; Voice quality control

1. Introduction

The speech waveform carries a variety of information; linguistic, segmental, supra-segmental, para-linguistic etc., of which perhaps the linguistic category is of greatest interest to most leading speech technologies today. At the same time, however, the non-linguistic information in speech, such as voice quality and voice individuality, plays an important role in the understanding and recognition of speech and in daily speech communication between people. Voice individuality, in particular, is important not only because it helps us identify the person to whom we are talking, but also because it enriches our daily life with variety. However, for most speaker-independent speech recognition tasks, voice individuality is simply an obstacle that must be overcome, and speaker normalization and adaptation are methods that have been developed for that purpose.

Recent developments in speech analysis and synthesis technology have made it possible to analyze many of the acoustic features that correlate with voice quality so that we may one day precisely control for voice individuality in computer speech (Fant, 1960; Gobl, 1989; Klatt and Klatt, 1990; Childers and Lee, 1991; Savic and Nam, 1991; Karlsson, 1992a, 1992b; Carlson, 1993; Murray and Arnott, 1993; Quatieri and McAulay, 1992; Takagi and Kuwabara, 1986). In this paper, we describe these voice quality control and conversion technologies after briefly summarizing the acoustic correlates of speech individuality.

2. What is voice individuality?

The factors that are relevant to voice individuality can be categorized in terms of socio/psycho-logical versus physiological dimensions. Speaking style, which an individual acquires as he or she is raised, by family and through schools and neighbors is socially conditioned. This usually depends on such factors as age, social status, dialect, and the community to which the speaker belongs. The "sound" or "timbre" of the voice comes mainly from the physiological or physical properties of the speech organs but is also conditioned by the speaker's emotional state (Kasuya et al., 1986a, 1986b; Klatt and Klatt, 1990; Fant, 1993; Murray and Arnott, 1993). Speaking style is acoustically realized in prosodic features such as the fundamental frequency contour, the duration of words, timing, rhythm, pause, power levels, and so on. Voice quality is reflected more in the glottal source frequency and spectrum, and in the power spectrum of the vocal tract (including nasalization of vowels) for which physiological or anatomical properties of the speech organs are primarily responsible (Kasuya et al., 1983, 1986a, 1986b; Muta et al., 1987).

These dimensions can be simplified by thinking of them in terms of software and hardware. The socio-linguistic and psychological factors of voice individuality come more from the control commands to the speech organs, and resemble software which can be programmed. Physiological factors, such as the 'static' nature of the organs, are closer to the hardware and less easily changed. When someone mimics another person's speech, he/she usually tries to copy the 'software' of the target speaker. Perhaps this 'software' may contain more important information for voice individuality than the 'hardware.' The present speech technologies, however, do not yet allow us to extract and manipulate this 'software' precisely. When we refer to 'voice individuality' in this paper we are less concerned with 'software' aspects such as speaking style and more concerned with 'hardware' aspects such as are produced from differences in vocal tract size and length.

3. Acoustic characteristics of voice individuality

Two types of acoustic characteristics, the voice source and the vocal tract resonance, act together to influence voice individuality. There are several acoustic parameters to be considered with respect to these characteristics (Fant, 1960; Malah, 1979; Eskenazi et al., 1990; Karlsson, 1990; Olive, 1992).

3.1. Voice source and resonance parameters

The following acoustic parameters are thought to have most influence on voice individuality (Karlsson, 1988; Lalwani and Childers, 1991a; Karlsson, 1992a, 1992b; Bavegard, 1993; Cook, 1993; Milenkovic, 1993). We will discuss the hardware aspects of the control and conversion in the following sections.

Voice source:

(1) the average pitch frequency,
(2) the time–frequency pattern of pitch (the pitch contour),
(3) the pitch frequency fluctuation,
(4) the glottal wave shape.

Vocal tract resonance:

(1) the shape of spectral envelope and spectral tilt,
(2) the absolute values of formant frequencies,
(3) the time–frequency pattern of formant frequencies (formant trajectories),
(4) the long-term average speech spectrum,
(5) the formant bandwidth.

3.2. Research on voice individuality

Research studies on voice individuality have a relatively long history but recent studies have been conducted more from a view point of speech technology and speaker recognition. Earlier studies from psychology and phonetics show the relationships between acoustic parameters and speaker's age, sex, height, weight and other physical properties (Schwartz and Rine, 1968; Hartman and Danhauer, 1976; Lass and Brown, 1978; Graddol and Swann, 1983; Childers and Lee, 1991; Childers and Wu, 1991; Wu and Childers, 1991). The literatures of this area is well summarized in a review article by J. Suzuki (Suzuki, 1985). Matsumoto et al. investigated contributions of pitch (F_0), formant frequencies, spectral envelope and other acoustic parameters for male vowel samples (Matsumoto et al., 1973). They concluded that F_0 was the most important factor on individuality, with formant frequency the next most important, followed by F_0 fluctuation and voice source spectral tilt. Sato found that global spectral shape gives perceptual cues to gender discrimination (Sato, 1974; see also Karlsson, 1986, 1991). Furui studied the relationship between psychological and physical distances among speakers (Furui, 1986), and reported that the long-term average spectrum smoothed by cepstrum coefficients showed the highest correlation, followed by averaged F_0. In particular, the 2.5–3.5 kHz frequency range was found to have the greatest contribution to individuality. Nakatsui et al. exchanged the source and the resonance characteristics from the vowels of three speakers, and reported that F_0 had a greater influence than the resonance characteristics of the vocal tract. Itoh and Saito, on the other hand, claimed a different result (Itoh and Saito, 1982). They showed that the spectral envelope had the greatest influence on individuality, followed by F_0 and temporal structure, as they investigated through resynthesized speech parameters for vowels, syllables and short sentences.

We do not assume that any single specific acoustic parameter alone carries the entire individuality information, but that voice quality is an amalgam of many parameters and the degree or order of importance among them can differ from speaker to speaker. The importance of particular acoustic parameters will very much depend on the nature of the speech materials under focus (Prosek et al., 1987; Gobl, 1989; Savic and Nam, 1991).

4. Speech technology, voice quality and individuality

Because there is no single acoustic parameter that plays a decisive role in voice individuality, it is difficult to extract the dominant acoustic features from speech to totally reflect a speaker's characteristics, and therefore not feasible to

model the individuality of a voice unless other methods of manipulating speech can be used. Since every acoustic parameter has something to do with the voice quality of a given speaker, many parametric attempts to control voice quality have been tried (Kuwabara and Takagi, 1991). Alternatively, more formal non-parametric methods attempt voice conversion without explicitly extracting and changing acoustic parameters (Abe et al., 1990; Abe and Sagayama, 1991; Valbret et al., 1992; Matsumoto et al., 1994).

4.1. A parametric approach to voice quality control

Recent developments in the analysis–synthesis of speech have made it possible to obtain more accurate acoustic parameters from running speech than before (Rodet et al., 1987; Nakajima and Suzuki, 1988). One of the great advantages of the analysis–synthesis method is that it separates the voice source information from vocal tract information, and allows reconstruction of the speech after independently manipulating the acoustic parameters, producing speech very close to the original in sound quality.

Kuwabara developed a pitch synchronous analysis–synthesis system that was capable of modifying pitch (F_0), formant frequencies and their bandwidths (Kuwabara, 1984). Four parameters, linear predictor coefficients, residual signals, amplitude and pitch durations, are obtained for each pitch period of voiced speech. Formant frequencies and their bandwidths are estimated from the predictor coefficients and spectral manipulation is performed by modifying the coefficients. The modified formant frequencies and bandwidths form the roots of a new polynomial equation which has the new predictor coefficients as its coefficients. Because the analysis is pitch-synchronous, pitch manipulation is quite simple; the length of residual signals for one pitch period is exactly that of the period. Pitch frequency can be changed by controlling the length of the residual signal. To raise pitch frequency, some data points at the end of the residual are eliminated, and to lower the frequency, zero signals are padded.

Parameter-controlled modified speech is produced from either raw or modified residual signals through a (raw or modified) vocal tract filter. The contributions of individual parameters to a certain voice quality and voice quality change can be determined experimentally by this method (Kuwabara, 1984, Kuwabara and Ohgushi, 1987).

4.2. Non-parametric methods of voice quality conversion

Rather than attempt to modify individual parameters separately, techniques for global optimization and control have been developed. These methods employ non-parametric conversion to map from one set of acoustic features (which may themselves be parametric) to another, thereby controling speaker individuality characteristics.

4.2.1. Voice quality conversion as acoustic characteristics mapping between speakers

Stimulated by the recent advances in speaker adaptation techniques for speaker-tolerant recognition, quite a few attempts have been made to convert voice quality using adaptation techniques (Abe et al., 1989, 1990; Abe and Sagayama, 1991; Iwahashi and Sagisaka, 1995; Matsumoto et al., 1994; Valbret et al., 1992). They attempt to map acoustic characteristics from a source speaker to a target speaker automatically. As they do not require parametric estimation of particular speech generation models for this conversion, they are referred to as non-parametric methods, in contrast to the traditional methods described in the previous section.

In non-parametric methods, voice conversion is considered as a simple mapping from a source speaker to a target speaker. To define this mapping, we must define (a) the domain of acoustic characteristics representation, (b) the mapping algorithms, and (c) a suitable speech corpus for training.

For spectral representation, LPC-related parameters have long been used (Abe et al., 1990; Abe and Sagayama, 1991), since they have been commonly used in speaker adaptation. Parameter interpolation plays an important role in most voice conversion schemes, and spectral parameters that have good interpolation characteristics (e.g. log area ratio, line spectrum pair) have

proved most useful. To get higher quality speech, further attention has also been paid to improving synthesis methods using LPC residual waveforms (Valbret et al., 1992), fine power spectral envelope (PSE) parameters (Matsumoto et al., 1994) and short time Fourier transform (Abe et al., 1989).

Most mapping algorithms originated from speaker adaptation techniques. In general, a mapping function is formed (trained) by taking the correspondence between a source speaker's spectral data and a target speaker's spectral characteristics according to phonetic labels time-aligned to the training speech data. However, since simple spectral correspondence can only give a sparse many-to-many mapping, inter/extra-polation is usually applied to get a smoother mapping. These procedures are described in detail in the following section.

It might be assumed that the size and the contents of the speech data determine the reliability of the mapping that can be achieved. However, to date, very few studies have been carried out on the relationship between the quality of the training database and the degree of the mapping that can be obtained. Iwahashi and Sagisaka (Iwahashi and Sagisaka, 1995) showed the feasibility of a conversion from one speaker to another using very limited training data (one word).

4.2.2. Spectral conversion methods

Reflecting various speaker adaptation techniques for speech recognition, several spectral conversion methods have been proposed.

(1) Code book mapping

Fig. 1 shows the outline of spectral conversion using code-book mapping. In this scheme, by applying speaker adaptation through vector quantization (Shikano et al., 1986) to spectral conversion, a source speaker's VQ spectrum $V_i(S)$ corresponding to the input spectrum $X(S)$ is converted to the target speaker's spectrum $X_i(T)$ by summing the corresponding target speaker's VQ spectrum $V_j(T)$ with the weights h_{ij} which represent the $V_i(S) \rightarrow V_j(T)$ correspondence observed in the training data. This procedure can be described by the following equation:

$$X_i(T) = \sum_{j=1}^{N_i} h_{ij} V_j(T) \Big/ \sum_{j=1}^{N_i} h_{ij}.$$

To cope with the spectral discontinuity resulting from vector quantization, two further improvements have been proposed. One employs fuzzy VQ (FVQ) to reduce the spectral distortion of hard VQ. In FVQ, a source speaker's spectrum $X(S)$ is not uniquely quantized at one $V_i(S)$

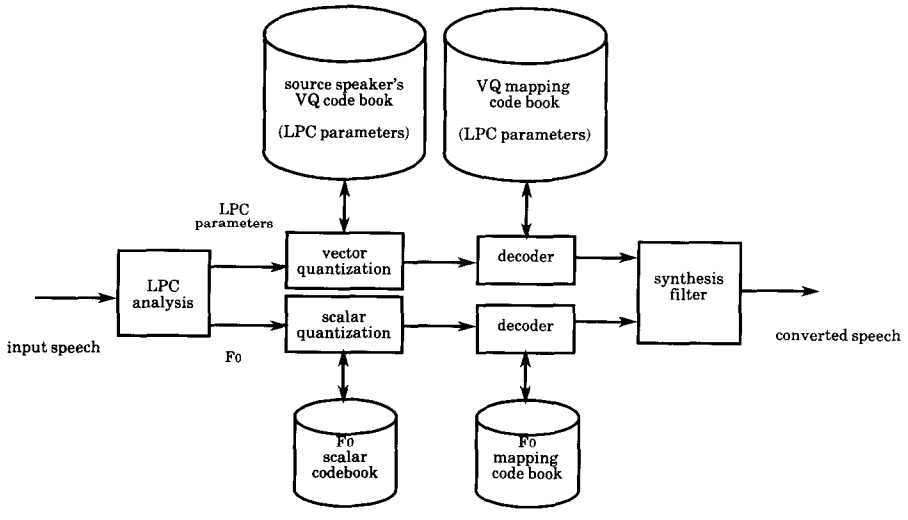

Fig. 1. Voice conversion using vector quantized code-book mapping.

but expressed as a combination of the neighboring $V_i(S)$ as follows:

$$\sum_{i=1}^{M_i} u_i V_i(T),$$

where u_i stands for a coefficient determined by the fuzzy membership function of $X(S)$ (Nakamura and Shikano, 1989). This FVQ has been used with speaker difference vector expression and effectively reduces VQ distortion.

The other attempted improvement employs segment VQ techniques. Instead of single-frame VQ coding, a phoneme-sized time interval is used, with the time-interval segmentation carried out using conventional HMM phone segmenters (Abe et al., 1990). Though this technique can achieve precise conversion, it also requires a large amount of speech training data to achieve accurate mapping performance.

(2) Speaker difference vector inter/extra-polation

A problem with the VQ mapping technique is that vectors are independent and no interpolation can be performed within an individual speaker's vector space. Because of this a large amount of training data is required to ensure adequate mapping between speakers. To counter this, relative positions between corresponding key points such as vowel centroids are employed as base points for interpolation (or extrapolation) for mapping between speakers. By generalizing speaker adaptation techniques proposed for coding (Shiraki and Honda, 1989) and for recognition (Niimi and Kobayashi, 1987), speaker difference vector inter/extra-polation has been applied to voice conversion (Matsumoto et al., 1994). In this method, as schematized in Fig. 2, an input spectrum $X(S)$ is converted to $X(T)$ by summing up the spectral differences $\{D_i\}$ of two speakers at typical points $\{V_i(S)\}$ as expressed in the following equation:

$$X(T) = X(S) + \sum_i W_i(S) D_i,$$

where $W_i(S)$ stands for a weighting coefficient of the difference vector D_i expressed as follows:

$$W_i(S) = \frac{\|X(S) - V_i(S)\|^{-p}}{\sum_j \|X(S) - V_j(S)\|^{-p}}.$$

Speaker difference vectors $\{V_i(S)\}$ are determined by minimizing the following function J which shows the difference in the training data between observed target speaker's spectra and the mapped spectra using this mapping function.

$$J(P,\{D_i\}) = \|Y(T) - X(T)\|^2,$$

where P denotes the corresponding function (DP-pass) between source speaker's data and target speaker's data and the sum on the right-hand side is taken over all training samples. By the reciprocal iterative minimization through P (by spectral DP-matching) and $\{D_i\}$ (by solving regular equations of regression analysis), J is minimized sub-optimally.

In this formulation, the parameter p controls the smoothness of interpolation, and experiments have shown that the closest approximation is given when $p = 1.0$ (Matsumoto et al., 1994). The original code-book mapping can be regarded as the least smooth case when p tends to infinity. On the other hand, under the most smooth case ($p = 0$), the weight $W_i(S)$ becomes constant and the difference between the two speakers is constant across all the source speaker's input speech.

(3) Frequency scaling

As often pointed out, the previous two methods suffer from spectral interpolation problems especially when the number of typical vectors $\{V_i(S)\}$ (VQ-code size) is small. To reduce the

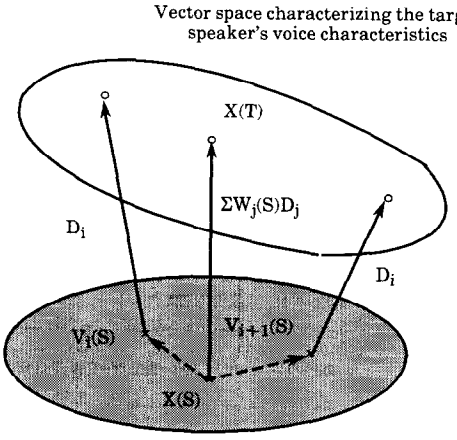

Fig. 2. Schematic view of speaker difference inter/extra-polation.

spectral distortion caused by interpolation, dynamic frequency warping (DFW) has been applied (Valbret et al., 1992). In this method, the power spectral envelope $E_i(f)$ is used to represent source speaker's spectral feature $X(S)$. The target speaker's spectrum is obtained by

$$X(T) = E_i(U_i(f)),$$

where U_i is a frequency warping function which is determined by minimizing source and target spectral difference in the training data. The calculation of a warping function is carried out after pre-emphasis to flatten spectral tilts for a better spectral match. As seen in these formulations, DFW can change spectral shape only in the frequency domain and therefore while it can adjust resonance frequencies (formants), and their bandwidth shift, it has little effect on their amplitudes.

(4) Multi-speaker inter/extra-polation

In the case when the target speaker's speech corpus is limited, precise data dependent mapping cannot be performed. What can be extracted from scarce data is limited to a speaker's global characteristics only. For limited training data, the speaker weighting method has been proposed for most effective use of intra-speaker consistencies using only a small number of weight parameters (Kosaka et al., 1993). Along with this idea, inter/extra-polation between speakers has also been studied (Iwahashi and Sagisaka, 1995).

In this method, a source speaker's input spectrum $X(S_i)$ is converted to a target speaker's spectrum $X(T)$ through the following equation:

$$X(T) = A_{S_i} X(S_i) + B.$$

For the calculation of transformation coefficients A_{Si} and B, traditional linear regressive analysis or neural network training is employed.

(5) Unsupervized conversion

As an application of unsupervised speaker adaptation techniques (Matsumoto and Yamashita, 1993) to voice conversion, FVQ mapping has been employed. In this conversion scheme, a target speaker's spectrum $X(T)$ is given by the following equation (Matsumoto et al., 1994):

$$X(T) = \sum_i \left[(a_i(T))^F X_i(T) \right] \Big/ \sum_i (a_i(T))^F,$$

where $\{X_i(T)\}$ are target speaker's code vectors given by minimizing fuzzy objective functions using the same technique as speaker difference vector inter/extra-polation (Matsumoto and Yamashita, 1993). $a_i(T)$ stands for FVQ class function with fuzziness F. The sum on the right-hand side is taken over vectors whose class function exceeds a certain threshold (i.e. $a_i(T) > a_{\text{threshold}}$).

4.2.3. Future prospects in non-parametric conversion methods

Non-parametric conversion methods do not assume any specific underlying model of speech production, but apply sets of weights that are derived by statistical methods to map from one set of features to another. The methods require three stages; parametric representation of acoustic characteristics, mapping algorithms, and speech corpus development to train the weights on the mappings. Each is a different field of research in itself. Good spectral interpolation characteristics are required, as is high quality speech synthesis for spectral conversion. Current non-parametric methods deal with only spectral envelope characteristics, and conversion of excitation source characteristics should be researched further for inclusion in the current scheme. From this respect, a good source-filter model is essential for both source characteristics and for prosody conversion to synthesize high quality speech with personal characteristics.

With regards to mapping algorithms, most to date have been applications of techniques already developed for speaker adaptation in speech recognition. Current work on the development of those algorithms for adaptation is aimed at finding a systematic mapping function adapting to the size and the contents of a target speaker's speech corpus. As seen in the current conversion techniques, inter/extra-polation according to speech data statistics under reasonable natural constraints that reflect speaker consistency and statistical phonotactics would help this integration.

Finally, as for speech corpora, almost no systematic research has been carried out on the design of an optimal (minimal) database for conversion. To characterize a speaker's spectral characteristics, the size and the contents (phonetic

contexts) should be considered. It is a logical consequence that the current mapping will be systematically improved through more consideration of the training corpus. Furthermore, it is also hoped that perceptual studies on personal characteristics for recognition and identification will soon start to give reasonable targets for these research efforts.

5. Conclusions

Recent studies on voice individuality and quality as well as on their control or conversion techniques have been reviewed. As far as speaker individuality is concerned, there is no single acoustic parameter that carries the entire information. Dominant acoustic features depend both on the speaker and on the speech material to be examined. This leads us to conclude that whereas it may not be feasible to take a simple parametric approach to changing voice individuality, developments in speech technology have made it possible to change the individuality from one speaker to another without explicit modeling of a speaker's voice characteristics by using extracted acoustic features directly. This technology seems to be promising as far as speaker conversion is concerned, though it still leaves much room for improvement with respect to the quality of the converted speech and the manipulation of its prosody.

Acknowledgements

The authors would like to express their thanks to Prof. Hiroshi Matsumoto for his information on research activities of speaker adaptation which correspond to the non-parametric voice conversion methods stated in Section 4.2.2.

References

M. Abe and S. Sagayama (1991), "Voice conversion based on segment mapping", *J. Acoust. Soc. Japan,* Vol. E-13, pp. 131–139.

M. Abe, S. Tamura and H. Kuwabara (1989), "A new speech modification method by signal reconstruction", *Proc. Internat. Conf. Acoust. Speech Signal Process.*, pp. 592–595.

M. Abe, S. Nakamura, K. Shikano and H. Kuwabara (1990), "Voice conversion through vector quantization", *J. Acoust. Soc. Japan,* Vol. E-11, pp. 71–76.

M. Bavegard et al. (1993), "Vocal tract sweeptone data and model simulations of vowels, laterals, and nasals", *STL-QPSR,* Vol. 4, pp. 43–76.

R. Carlson (1993), "Models of speech synthesis", *STL-QPSR,* Vol. 1, pp. 1–14.

D.G. Childers and C.K. Lee (1991), "Vocal quality factors; Analysis, synthesis, and perception", *J. Acoust. Soc. Amer.,* Vol. 90, pp. 2394–2410.

D.G. Childers and K. Wu (1991), "Gender recognition from speech, Part II; Fine analysis", *J. Acoust. Soc. Amer.,* Vol. 90, pp. 1841–1856.

P.R. Cook (1993), "SPASM, a real time vocal tract physical model controller, and singer, the companion software synthesis system", *Computer Music J.,* Vol. 17, pp. 30–44.

L. Eskenazi, D.G. Childers and D.M. Hicks (1990), "Acoustic correlates of vocal quality", *J. Speech and Hearing Research,* Vol. 33, pp. 298–306.

G. Fant (1960), *Acoustic Theory of Speech Production* (Mouton, The Hague).

G. Fant (1993), "Some problems in voice source analysis", *Speech Communication,* Vol. 13, Nos. 1–2, pp. 7–22.

S. Furui (1986), "Research on individuality features in speech waves and automatic speaker recognition techniques", *Speech Communication,* Vol. 5, No. 2, pp. 183–197.

C. Gobl (1989), "A preliminary study of acoustic voice quality correlates", *STL-QPSR,* Vol. 4, pp. 9–22.

D. Graddol and J. Swann (1983), "Speaking fundamental frequency: Some physical and social correlates", *Language and Speech,* Vol. 24, pp. 351–356.

D.E. Hartman and J.L. Danhauer (1976), "Perceptual features of speech for males in four perceived age decades", *J. Acoust. Soc. Amer.,* Vol. 59, pp. 713–715.

K. Itoh and S. Saito, (1982), "Effects of acoustical feature parameters of speech on perceptual identification of speaker", *IECE Trans.,* Vol. J65-A, pp. 101–108.

N. Iwahashi and Y. Sagisaka (1995), "Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks", *Speech Communication*, Vol. 16, No. 2, pp. 139–151.

I. Karlsson (1986), "Glottal wave forms for normal female speakers", *J. Phonetics,* Vol. 14, pp. 415–419.

I. Karlsson (1988), "Glottal waveform parameters for different speaker types", *Proc. Speech '88, 7th FASE Symposium*, Vol. 1, pp. 225–231.

I. Karlsson (1990), "Voice source dynamics for female speakers", *Proc. Internat. Conf. Spoken Language Proc.,* Vol. 1, pp. 69–72.

I. Karlsson (1991), "Female voices in speech synthesis", *J. Phonetics,* Vol. 19, pp. 111–120.

I. Karlsson (1992a), Analysis and synthesis of different voices with emphasis on female speech, Ph.D. dissertation, Royal Inst. Tech., Sweden.

I. Karlsson (1992b), "Modelling voice variations in female speech synthesis", *Speech Communication,* Vol. 11, Nos. 4–5, pp. 491–495.

H. Kasuya, Y. Kobayashi and T. Kobayashi (1983), "Characteristics of pitch period and amplitude perturbations in pathological voice", *Proc. Internat. Conf. Aoust. Speech Signal Process.,* pp. 1372–1375.

H. Kasuya, K. Masubuchi, S. Ebihara and H. Yoshida (1986a), "Preliminary experiments on voice screening", *J. of Phonetics,* Vol. 14, pp. 463–468.

H. Kasuya, S. Ogawa, Y. Kikuchi and S. Ebihara (1986b), "An acoustic analysis of pathological voice and its application to the evaluation of laryngeal pathology", *Speech Communication,* Vol. 5. No. 2, pp. 171–181.

D.H. Klatt and L.C. Klatt (1990), "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *J. Acoust. Soc. Amer.,* Vol. 87, No. 2, pp. 820–857.

T. Kosaka, J. Takami and S. Sagayama (1993), "Rapid speaker adaptation using speaker-mixture allophone models applied to speaker-independent speech recognition", *Proc. Internat. Conf. Acoust. Speech Signal Process.,* pp. 570–573.

H. Kuwabara (1984), "A pitch-synchronous analysis/synthesis system to independently modify formant frequencies and bandwidths for voice speech", *Speech Communication,* Vol. 3, No. 3, pp. 211–220.

H. Kuwabara and K. Ohgushi (1987), "Contributions of vocal tract resonant frequencies and bandwidths to the personal perception of speech", *Acustica,* Vol. 63, pp. 121–128.

H. Kuwabara and T. Takagi (1991), "Acoustic parameters of voice individuality and voice-quality control by analysis–synthesis method", *Speech Communication,* Vol. 10, Nos. 5–6, pp. 491–495.

A.L. Lalwani and D.G. Childers (1991a), "Modeling vocal disorders via formant synthesis", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.,* Vol. 1, pp. 505–508.

N.J. Lass and W.S. Brown (1978), "Correlational study of speaker's height, weight, body surface areas and speaking fundamental frequencies", *J. Acoust. Soc. Amer.,* Vol. 63, pp. 1218–1220.

D. Malah (1979), "Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals", *IEEE Trans. Acoust. Speech Signal Process.,* Vol. 3, pp. 121–133.

H. Matsumoto and Y. Yamashita (1993), "Unsupervised speaker adaptation from short utterances based on a minimized fuzzy objective function", *J. Acoust. Soc. Japan,* Vol. E-14, pp. 353–361.

H. Matsumoto, S. Hiki, T. Sone and T. Nimura (1973), "Multidimensional representation of personal quality of vowels and its acoustical correlates", *IEEE Trans. AU,* Vol. AU-21, pp. 428–436.

H. Matsumoto, Y. Maruyama and H. Inoue (1994), "Voice quality conversion based on supervised/unsupervised spectral mapping", *J. Acoust. Soc. Japan,* Vol. 50, pp. 549–555.

P.H. Milenkovic (1993), "Voice source model for continuous control of pitch period", *J. Acoust. Soc. Amer.,* Vol. 93, pp. 1087–1096.

I.R. Murray and J.L. Arnott (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.

H. Muta, T. Muraoka, K. Wagatsuma, H. Fukuda, E. Takayama, T. Fujioka and S. Kanou (1987), "Analysis of hoarse voices using the LPC Method", *Laryngeal Function in Phonation and Respiration,* ed. by T. Bear, C. Sasaki and K. Harris (College-Hill Press), pp. 463–474.

T. Nakajima and T. Suzuki (1988), "Power spectrum envelope (PSE) speech analysis–synthesis system", *J. Acoust. Soc. Japan,* Vol. 44, pp. 824–832.

S. Nakamura and K. Shikano (1989), "Spectrogram normalization using fuzzy vector quantization", *J. Acoust. Soc. Japan,* Vol. 45, pp. 107–114.

Y. Niimi and Y. Kobayashi (1987), "Speaker adaptation of a code book of vector quantization", *Proc. European Conf. on Speech Technology,* Vol. 2, p. 430.

J.P. Olive (1992), "Mixed spectral representation-formants and linear predictive coding (LPC)", *J. Acoust. Soc. Amer.,* Vol. 92, pp. 1837–1840.

A.R. Prosek, B.E. Montgomery and D.B. Hawkins (1987), "An evaluation of residue features as correlates of voice disorders", *J. Communication Disorders,* Vol. 20, pp. 105–117.

T. Quatieri and R.J. McAulay (1992), "Shape invariant time-scale and pitch modification of speech", *IEEE Trans. Signal Process.,* Vol. 3, pp. 497–510.

X. Rodet, P. Depalle and G. Poirot (1987), "Speech analysis and synthesis methods based on spectral envelopes and voiced/unvoiced functions", *Proc. European Conf. on Speech Technology,* Vol. 1, pp. 155–158.

H. Sato (1974), "Acoustic cues of female quality", *IECE Trans.,* Vol. 57-A, pp. 23–30.

M. Savic and I-H Nam (1991), "Voice personality transformation", *Digital Signal Processing,* Vol. 1, pp. 107–110.

M.F. Schwartz and H.E. Rine (1968), "Identification of speaker sex from isolated, whispered vowels", *J. Acoust. Soc. Amer.,* Vol. 44, pp. 1736–1737.

K. Shikano K.F. Lee and R. Reddy (1986), "Speaker adaptation through vector quantization", *Proc. Internat. Conf. Acoust. Speech Signal Process.,* pp. 2642–2646.

Y. Shiraki and M. Honda (1989), "Speaker adaptation algorithms based on piecewise moving adaptive segment quantization", *IEICE Trans.,* Vol. J72-DII, pp. 1118–1124.

J. Suzuki (1985), "Correlation of speaker's physical features and speech", *J. Acoust. Soc. Japan,* Vol. 41, pp. 895–900.

T. Takagi and H. Kuwabara (1986), "Contributions of pitch formant frequency and bandwidth to the perception of voice-personality", *Proc. Internat. Conf. Acoust. Speech Signal Process.,* pp. 889–892.

H. Valbret, E. Mouline and J.P. Tubach (1992), "Voice transformation using PSOLA technique", *Speech Communication,* Vol. 11, Nos. 2–3, pp. 175–187.

K. Wu and D.G. Childers (1991), "Gender recognition from speech, Part I; Coarse analysis", *J. Acoust. Soc. Amer.,* Vol. 90, pp. 1828–1840.