APPLICATION OF KOREAN TEXT-TO-SPEECH FOR X.400 MHS SYSTEM

Hee Dong Kim [1] , Jun Mo Koo [2], Ho Joon Choi [3], Sang Taek Kim [3]

The University of Suwon. Dept. of Telematics Engineering [1]
DigiCom Institute of Telematics [2]
Korea Telecom Research Center [3]

ABSTRACT

This paper presents the Korean text-to-speech (TTS) algorithm with speed and intonation control capability, and describes the development of the Voice message delivery system employing this TTS algorithm. This system allows the Interpersonal Messaging (IPM) Service users of Message Handling System(MHS) to send his/her text messages to user via telephone line using synthetic voice. In the X.400 MHS recommendation, the protocols and service elements are not specified for the voice message delivery system. Thus, we defined access protocol and service elements for Voice Access Unit based on the application program interface for message transfers between X.400 Message Transfer Agent and Voice Access Unit. The system architecture and operations will be provided.

1. INTRODUCTION

Unlike other languages, Korean has complex and unique properties such that any TTS algorithm for foreign language may not be applied with minor modification. For last decade, the Korean TTS algorithm has been studied by many researchers at numerous instututes. However, practical application of TTS has not yet been reported. In this paper, we present the Korean TTS algorithm that does not impose any restrictions on the input text, and its application to the Message Handling System(MHS) environment.

As electronic mail systems are wide spread, the needs for a e-mail standard has been increasing. In 1984, ITU-T has announced MHS recommandation X.400 and F.400 series including system model, services and its related protocols[4][6]. Korea Telecom (KT) developed public MHS service, named KT-MAiL, began to provide various additional services such as bulletin board and chattering services. Furthermore, in order to give the benefit to the subscribers, KT intended to support fax and telephone delivery service in a single platform, thus, installed the VOICE/FAX-AU system which allows the senders to deliver the text messages to the receivers regardless the terminal type using TTS and Text-to-Fax (TTF) technology. The discussions of this paper is concentrated on the Korean TTS algorithm used for media conversion, and implementation of this

system. Hereafter, we will denote the voice part of VOICE-FAX/AU as VOICE-AU.
In section two, the Korean TTS algorithm will be discussed followed by an overview of the MHS in section three. System configuration will be presented in section 4, followed by a conclusion.

2. DESCRIPTION OF SPEECH SYNTHESIS ALGORITHM

Speech synthesis algorithm with speed and tone control capability is designed to convert an unrestricted Korean text into synthetic speech waveforms. The algorithm consists of linguistic processing and acoustic processing phase. Each of these processes is composed of a series of modules as shown in Fig.1. In this section, we discuss the functional content of each module.

2.1 Text Analysis

2.1.1 Text Preprocessing

A normal Korean text consists of Korean alphabet called "Hangeul"; however, unrestricted text may contain symbols, abbreviations, and even foreign words (e.g., English). A text preprocessor converts input text into a data suitable for linguistic analysis. For example, "Kg", a unit symbol, following a number will be converted to the Korean words corresponding to "Kilogram". For the precise preprocessing, it is ideal to refer all the possible contextual informations. However, it is not easy to extract contextual informations completely in effective manner. So, in order to speed up the process, we use only the surrounding words for the conversion and refer to the look up dictionary for frequently used expressions. For English words, at first, they are broken into pre-defined subword clusters consisting of consonants and vowels. Afterwards, English letter to Korean sound rules are applied to the clusters. In addition, pronunciation dictionary is used for some exceptional English words for which the letter-to-sound rules does not apply.

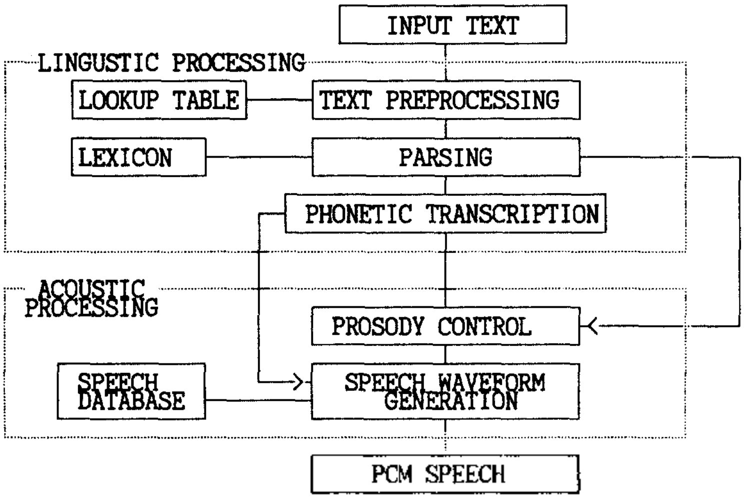

Fig. 1. Block Diagram of Korean Text-To-Speech System

2.1.2 Parsing

The TTS parser supplies a surface structure parse informations which are used for producing prosodic effects in the output speech. Also, clause and phrase boundaries are determined as a result of the parsing process. These phrase- and clause-level structures provide much of the syntactic information needed by the present prosodic algorithms. To minimize processing burdens, the parser employs only small amount dictionaries including the dictionary for the endings of words, auxiliary word dictionary, and adverb dictionary. But it generates indispensable informations such as phrase and clause boundaries, and the presence of the endings. The parsing is accomplished via ATN(Augmented Transition Network)-type interpreter and grammars for noun, adverb, adjective, and verb groups.

2.1.3 Phonetic transcription

The phonetic transcription is a process producing real pronunciations from phonetic notations by phonological rules. Every Korean alphabet has its own sound. A Korean alphabet, unless interfered with other alphabets, always produces same sound. However, if two consonants occur consecutively, they affect each other. There are several coarticulation rules for converting pronunciation of consonants[1]. The phonetic transcription algorithms are applied first to convert input text into pronunciations. Since there are some pronunciations which can not be described by the rules, an exceptions dictionary is also implemented for those words.

2.2 Acoustic Processing

2.2.1 Speech Database Generation

In order to implement TTS system based on the synthesis unit concatenation method, one should determine synthesis units and its concatenation method. The important factors in determining synthesis unit may be the quality of synthetic speech and complexity. In Korean, a syllable is composed of series of an initial consonant, a vowel, and a final consonant, where one or both of the consonants may be omitted. Phonological variations are observed between the final consonant and initial consonant of next syllable. One of the typical examples is that voiceless consonant between voiced sounds becomes voiced. These phonological variations can be included in diphone units. Thus, we selected diphones as synthesis units, where size is moderate and coarticulation effects are inherently contained in them. Since diphones are usually concatenated with other diphones which are not met at diphone extraction phase, discontinuity occurs between diphones thus degrading the speech quality. In order to alleviate this problem, we classified and extended diphones according to their context. In other words, diphone boundaries are adjusted to be located at voiced sounds or silence so that discontinuity can be minimized and additional compensation process may not be required at concatenation process. Using 1150 synthesis units that we extended, we can achieve more natural speech without employing complex concatenation rules.

2.3.2 Prosodic component

The quality of a synthetic speech sound can be evaluated by intelligibility and naturalness. The factors affecting the naturalness are accent and intonation. The accent of Korean words is a complex phenomena of duration lengthening, increase of fundamental frequency, and increase of the amplitude of a word. Unlike English, the accent of Korean word can not differenciate the meaning of the word. It represents only speaker's emotion and intention. Hence, this Korean accent can be modeled through the syntactic structure analysis producing the number of syllables in a word, the stem and the ending of a word. Duration of each synthetic segment is determined by a set of rules and parameters which are constructed from observations of a single speaker's reading of typical paragraphs. We employed the duration model proposed by Klatt[2]. The model is summarized by formula.

$$\mathrm{DUR} = ((\mathrm{INHDUR} - \mathrm{MINDUR}) * \mathrm{PRCNT}/100 + \mathrm{MINDUR}) * \mathrm{SPRATE}) \qquad (1)$$

where INHDUR is the inherent duration of segment, MINDUR is the minimum duration of a segment, SPRATE is the speaking rate, and PRCNT is the percentage shortening determined by applying rules.

Another important component in the generation of natural-sounding of speech is the fundamental frequency(F0) of the voiced sound. F0 algorithm utilizes the phrase structure of each sentence analized by the parser, and the accent position of each word. The F0 contour is determined by the baseline, the micro-melody, and the macro-melody component[3]. The baseline shows the monotonous decreasing pattern. The micro-melody is related to the accent location in words and the macro-melody is to the phrase structure of a sentence. The fundamental frequency at time t is obtained by

$$F0(t)= B(t) + Acc(t) + Phr(t), \qquad (2)$$

where B(t) is the F0 value in the baseline at time t, Acc(t) is for the micro-melody, and Phr(t) is for the macro-melody at time t.

2.3.3 Generation of Speech Signal

In order to generate synthetic speech signal, we used modified linear predictive coding method, where spectral coefficients are interpolated to improve discontinuity at the boundaries of synthesis units. For this, we used log-area ratio(LAR) coefficients for interpolation and reflection coefficients for signal generation. It was found that interpolation of LAR parameters works well because linear movements of formant frequencies remain essentially linear after interpolation. Formant frequency movements are thus reproduced fairly well after interpolation. Most of the nonlinear effects were found to be confined to amplitude and bandwidth of formant. But, from the perceptual points of view, errors in formant amplitudes and bandwidths are not so important as the errors in formant frequencies.

3. MHS OVERVIEW

The X.400 MHS recommendations describe a standard way for transporting electonic messages between different end systems. The functionality of X.400 is structured according to the Functional model as shown in Fig 2. In this model, the Message Transfer System(MTS) comprises a number of Message Transfer Agents(MTAs). Operating together, in a store and forward manner, the MTAs transfer messages and deliver them to the intended recipients.

Interpersoanl Messaging(IPM) Serivce provides a user with features to assist in communicating with other IPM Service users. The IPM Service uses the capabilities of the MT Services for sending and receving interpersonal messages. The IPM system comprises of the IPMS-UA, IPMS-Message Store(MS), and the optional Access Units(AU). User Agent(UA) supports one indivisual user for preparing, submitting, receiving and formatting messages. Through MS the user can submit messages, and retrieve messages that have been delivered to the MS. The AU permits the interchange of message between a MTA and non-X.400 systems, performing the necessary conversions to make them fit into its particular characteristics. The optional Access Units may allow for Teletex and Telex users to intercommunicate with the IPM Service and the optional Phisysical Delivery Access Unit allows IPM users to send messages to users outside the IPM Service who have no access to MHS.

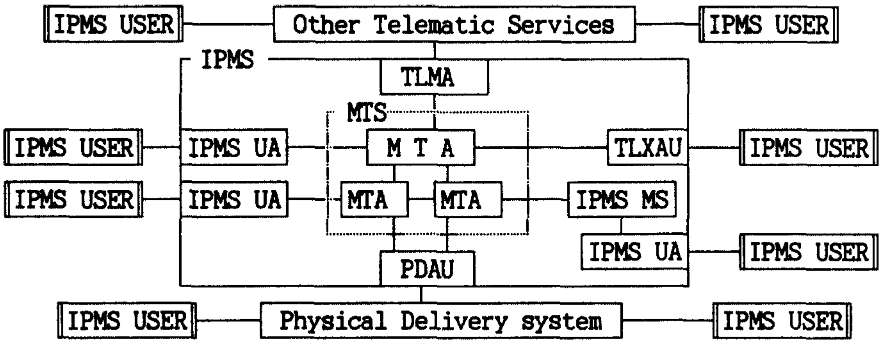

Fig.2. The Model of the Interpersonal Message Handling System

4. VOICE-AU SYSTEM

4.1 Operation and Access Protocol

The VOICE-AU is an additional fuctionality provided within the MHS to accomodate telephony users. It performs the necessary conversions to make them fit into its particular characteristics. And the VOICE-AU emulates the services for the the intended recipients to provide them with X.400 services. The e-mail users can transperently access the VOICE-AU. Now it supports both the mail addresses of the type RFC-822 for the Internet users and the Originator/Recipient address of CCITT.

Although the concept of MHS is intended to support multi-media message format in an application independent manner, the user interface for telematic terminal and its media conversion is not yet fixed. In the X.400 recommandation, Telex Access Unit(TLX-AU) is defined for permitting the existing telex terminals to communicate with MHS systems. However, the access protocol and service elements are not specified for the VOICE-AU system. Thus, we have to define our proprietary protocol based on Teletex/IPM service and the T.330 Telematic Access Protocol[5].

The operations of VOICE-AU related to the message delivery is shown in Fig.3. It consists of IP message reception, message conversion, message delivery, and notification. Message is converted in the TTS and TTF module. Telephone line interface module handles a lot of functions related to the telephone interface, call processing, and T.30 procedure for fax delivery. System dials the receipient and checks the status of line. If the receiving party answers the phone, it gives the voice prompt saying," Hello, You have the message from Korea Telecom.", and goes on providing the synthetic speech. If the message was not delivered by any reason, the system should retry in a predetermined interval. The result of message delivery will be sent back to the sender in the form of notification message.

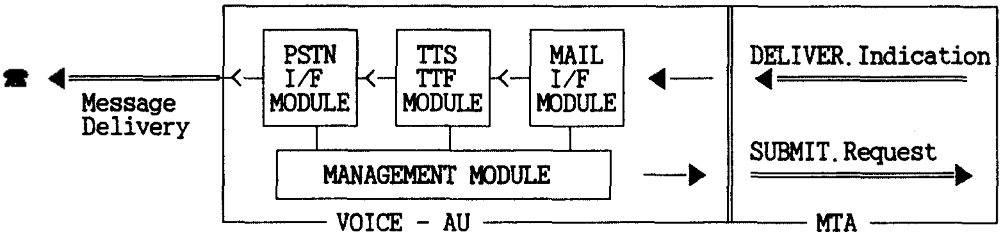

Fig.3. Message Delivery Operation of VOICE-AU

VOICE-AU access protocol comprises two actions, Message Delivery action and Notification action. The Message Delivery action is initiated by the VOICE-AU upon receipt of *DELIVERY.Indication* for an IP-message. In this actions, the VOICE-AU sends to the telephone user the guide information related to IPM service elements and body messages. The elements of Service related to the guide information of the VOICE-AU are listed in Table 1.

Table 1. Service Element for guide information

F.400 Elements of Service	Message Delivery telephony user	Guide Info. by VOICE-AU
Authorizing User Indication	X	X
IP-message Indicaiton	X	
Importance Indication	X	X
Expiry Date Indication	X	X
Forwaded IP-message Indication	X	X
Multi-part Body	X	
Originator Indication	X	X
Reply Request Indication	X	X
Sensitivity Indication	X	X
Subject Indication	X	X
Submission Time Stamp Indication	X	X

The Notification action is initiated by VOICE-AU. In this action, the AU automatically generates auto-receipt status report after successful delivery of a message, and submits reports to the MTA. When attempts to deliver a message fails, AU responses to MTA by sending the non-delivery reasons reflecting local problems between AU and telephone terminal.

4.2 Network Configuration

Korea Telecom has the nation-wide public switched telephone network(PSTN) and the public switched data network(PSDN) for information distribution purpose. For the public MHS service, Korea Telecom installed KT-MAiL host systems to the PSDN as an information provider as well as the Packet Assembler/Disassembler (PAD) and VOICE-AU systems as gateways between PSTN and PSDN. Network configuration is shown in Fig.4.

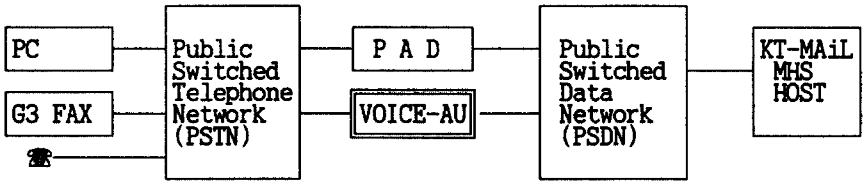
PC
G3 FAX
☎
Public Switched Telephone Network (PSTN)
P A D
VOICE-AU
Public Switched Data Network (PSDN)
KT-MAiL MHS HOST

Fig. 4. Network Configuration of KT-MAiL System

4. 3 System Configuration

The VOICE-FAX/AU system was designed to be used as either a stand alone voice/fax mail system or a media-conversion system. The hardware of the system has five major components: hot and stand-by CPU, telephone interface, X.25 interface, speech synthesis, and mass storage modules. These modules are interconnected to each other using VMEbus as shown in Fig.5. The two CPU modules working in active/stand-by mode to ensure high availablity control all the functions of system. The telephone interface module is equiped with telephone line interface circuit, 32kbps Adaptive Differential Pulse Code Modulation(ADPCM) coder, and fax modem chip. The communication module provides the standard X.25 interface to the public switched data network. Speech synthesis module(SSM) converts the text message into the ADPCM coded speech in

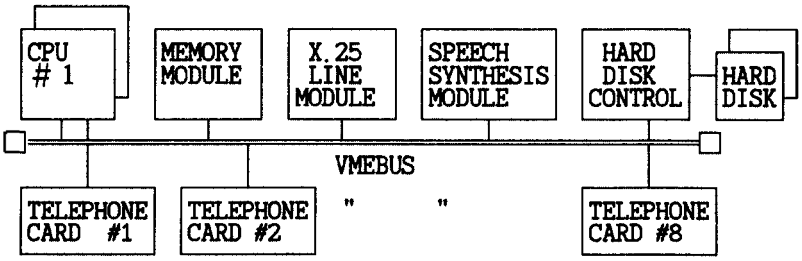

Fig.5. Hardware Block diagram of VOICE/FAX-AU System

real time. Finally, mass storage module are used for storing the voice prompt for guidance and fax cover sheet data as well as the user data.
The more detail description of SSM will be in order. The block diagram of SSM is shown in Fig.6. SSM is equiped with 8051 one chip microcontroller for interfacing with the CPU module, and executing the text preprocessing. The input to the 8051 is single sentence or commands which controls the speed and tone of the synthetic speech. Then, the text is passed to the acoustic processor, based on AMD 2105 Digital Signal Processor(DSP). DSP synthesizes the speech using the LPC coded database, and generates the 12-bit PCM coded data followed by DA converter. The analog signal is fed to the ADPCM codec, and 32kbps speech data will be passed to the main CPU module.

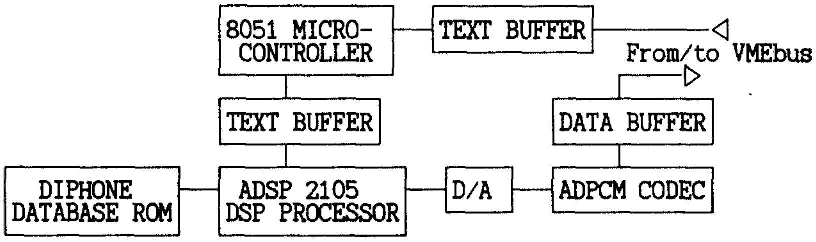

Fig.6. Block diagram of Speech Synthesis Module

5. CONCLUSION

We have discussed the Korean TTS algoritm and its application to the VOICE-AU system. The system has begun to nation-wide service since 1992. From this experince, the MHS service and its additional media conversion service twas proven to be a powerful means of delivering information to any telematic users. Further study include the improvement of speech quality, bidirectional message delivery, and the interconnection with mobile service.

REFERENCES

[1] W. Huh, Korean phonoloy(in Korean), Saem Munhwasa, 1985.

[2] J. Allen, et al., From text to speech, Cambridge University Press, 1987.

[3] H. Fujisaki, H. Kawaii., "Realization of linguistic information in the voice fundamental frequecy contour of the spoken Japanese," in Proc. of ICASSP 88, pp.663-666, 1988.

[4] CCITT Recommendation X.400-X.430 - Data communication Networks Message Handling Systems, CCITT, Geneva,1988.

[5] CCITT Recommendation T.300-T.330 - General Principles of Telematic Interworking, CCITT, Geneva,1988.

[6] CCITT Recommendation F.400-F.430 - "Message Handling and Directory Service - Operations and Definition of Service," CCITT, Geneva,1988.