# 무선랜 환경에서의 분산 음성 인식을 이용한 음성 다이얼링 시스템

박성준(KT), 구명완(KT)

<차 례>



<Abstract>

## A Voice-Activated Dialing System with Distributed Speech Recognition in WiFi Environments

Sung-Joon Park, Myoung-Wan Koo

In this paper, a WiFi phone system with distributed speech recognition is implemented. The WiFi phone with voice-activated dialing and its functions are explained. Features of the input speech are extracted and are sent to the interactive voice response (IVR) server according to the real-time transport protocol (RTP). Feature extraction is based on the European Telecommunication Standards Institute (ETSI) standard front-end, but is modified to reduce the processing time. The time for front-end processing on a WiFi phone is compared with that in a PC.

* Keywords : Distributed speech recognition, WiFi phone, Front-end.

# 1. 서 론

초기에는 데이터 전송 위주였던 인터넷이 급속한 통신 기술의 발전에 따라 VoIP(voice over internet protocol) 기술을 이용한 인터넷 전화도 등장하였다. VoIP란 기존의 회선교환 방식의 PSTN(public switched telephone network)으로 음성을 전송하던 것과 달리 음성을 데이터화한 후 인터넷을 통하여 전송하는 기술을 의미하는 것으로서, IP(internet protocol) 네트워크상에서 음성 신호를 패킷 형태로 전송하는 음성 서비스에 사용될 수 있다.

VoIP의 장점은 일반 텔레포니 환경과는 달리 각각의 단말도 어느 정도 지능적인 기능을 갖추어 클라이언트, 서버 각각이 발전해 나가면서 기존의 집중형 모델의 발전 한계를 빠른 시간 내에 뛰어넘을 수 있다는 것이며, DSR의 경우에도 마찬가지로 지능적인 클라이언트인 VoIP 단말기와 서버로 구성된 분산형으로 구성되었다.

인터넷에 접속하는 기기 중에는 화면과 키패드가 작고 무선으로 작동하는 무선 인터넷 단말기들도 있는데, 이들 기기에 음성 인식 기능이 제공된다면 사용자들이 편리하게 서비스를 이용할 수 있다. 그런데, 이러한 기기들은 계산 속도, 메모리, 배터리 에너지 등에서 제약을 많이 받기 때문에 음성 인식과 같이 계산량이 많은 작업을 수행하기에는 무리가 있다.

음성 인식은 전형적으로 신호 처리를 하는 특징 추출 단계와 추출된 특징 데이터를 이용하여 탐색을 하는 단계로 나눌 수 있는데, 신호 처리를 하는 전처리 단계는 전체 계산량과 메모리 사용량에서 적은 부분만 차지하는 반면, 탐색 단계는 상당한 양의 계산량과 메모리가 필요하기 때문에 모든 작업을 사용자 단말기에서 수행한다면 실시간적인 처리가 어렵게 된다. 따라서 사용자의 단말기에서는 음성의 특징 데이터만 추출하고 탐색은 서버에서 이루어지는 분산 음성 인식이 효과적인 방법이 될 수 있다. 또한 단말기에서의 특징 데이터 추출은 MFCC(mel-freqeuncy cepstral coefficients)와 같이 기계의 인지에 맞는 방법을 사용함으로써 기존의 코덱을 통해 서버로 전달되는 음성에서 특징 데이터를 추출할 때보다 나은 인식 성능을 가질 수 있다.

본 논문에서는 무선 VoIP에 분산 음성 인식을 적용한 시스템을 소개하고 처리 시간에 초점을 맞추어 실험한 결과를 제시한다. 그리고 처리 시간을 단축하기 위해 사용된 방법에 대하여 기술한다. 단말기는 무선 환경에서 동작하는 WiFi 전화기를 사용하며, IP-PBX(internet protocol - private branch exchange)와 IVR 서버를 구현한다. WiFi 전화기는 802.11 프로토콜을 따르며, G.711 음성 코덱을 포함하고 있다. WiFi 전화기는 ETSI DSR(distributed speech recognition) 표준에 따른 전처리 모듈이 포함하며, 음성으로부터 생성된 특징 추출 데이터를 IP-PBX를 거쳐 IVR 서버에 전송하고, IVR 서버에서는 전송된 데이터를 이용하여 음성 인식 작업을 수행

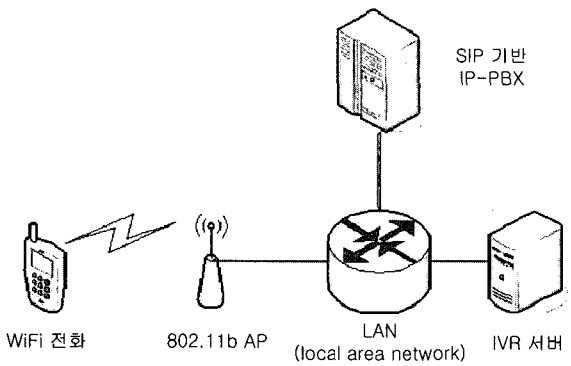
하고 그 결과를 WiFi 전화기로 전송한다.

논문의 구성은 다음과 같다. 2장에서는 본 논문에서 시스템의 구조와 분산 음성 인식을, 3장에서는 음성 다이얼링이 이루어지는 과정을 설명한다. 4장에서는 WiFi 전화기와 PC에서의 특징 추출에 소요되는 시간을 비교하며, WiFi 전화기에서 전처리에 사용되는 시간을 단축하기 위하여 수행된 작업을 설명하고, 5장에서 결론을 맺는다.

## 2. 시스템 구조

### 2.1 음성 다이얼링 시스템의 구조

음성 다이얼링 시스템은 크게 세 부분으로 나눌 수 있는데, WiFi 전화기, IP-PBX, IVR 서버 등이다. WiFi 전화기에서는 음성의 특징을 추출하고 이를 IVR 서버에 전달하며, IVR 서버에서는 내부의 음성 인식 모듈을 이용하여 음성 인식을 수행한다. SIP(session initiation protocol) 기반의 IP-PBX는 WiFi 전화기와 IVR 서버를 연결해 주는 역할을 한다. 시스템의 전체 구조는 <그림 1>과 같다.




<그림 1> 무선랜 환경에서의 음성 인식 자동 다이얼링 시스템의 구조도

무선 환경에서 WiFi 전화기를 IP 망에 연결해 주기 위해서 802.11b AP(access point)가 사용되며, WiFi 전화기 내에 802.11을 위한 모듈이 탑재되어 있다.

IP-PBX는 SIP 기반의 소프트스위치로서 기존의 텔레포니 환경의 PBX의 기능을 IP 환경에서 구현한 것으로 볼 수 있다. 본 논문에서 구현하는 분산 음성 인식 기능과 관련하여 IP-PBX는 변경할 사항이 없으며, 기존의 IP-PBX를 사용할 수 있다.

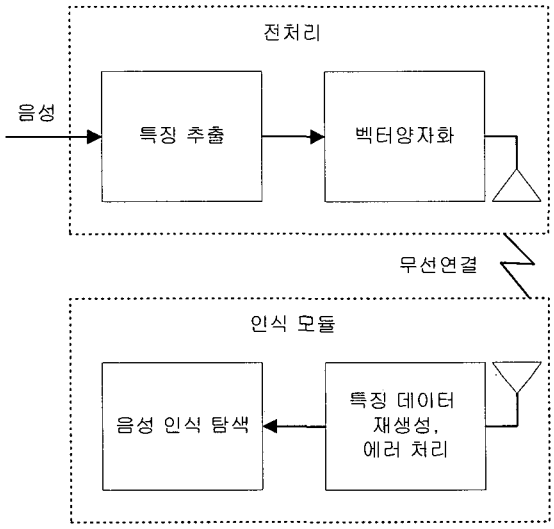
IVR 서버는 음성 인식을 수행하고 해당되는 전화 번호를 가진 단말기로 호를

전환하도록 IP-PBX로 요청하는 기능을 가진다.

분산 음성 인식을 위하여 WiFi 전화기와 IVR 서버 내에 각각 필요한 모듈이 구현되었으며, 다음 절에서 좀 더 자세히 설명한다.

## 2.2 분산 음성 인식

무선 환경에서의 분산 음성 인식 시스템은 전형적으로 <그림 2>와 같은 구조를 가진다.




<그림 2> 분산 음성 인식 시스템 구조의 블록도

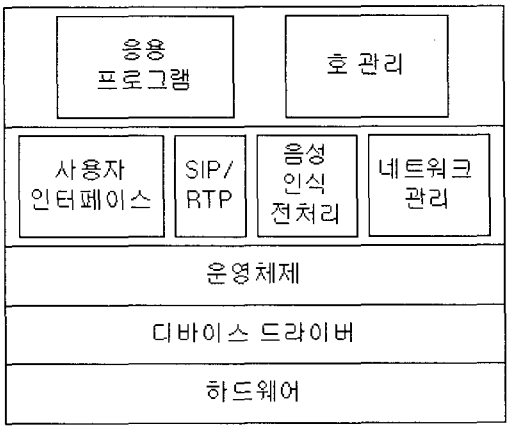
ETSI ES 202 212 표준에는 분산 음성 인식을 위한 특징 추출과 벡터양자화 등에 대한 틀이 제공되며[1], 본 논문에서는 ETSI DSR 표준을 따라 전처리 모듈을 WiFi 전화기에 구현하였다. 특징 추출 과정에는 잡음 제거, 피치 계산, 기타 연산이 포함되고, 벡터양자화 과정에는 에러 처리를 위한 CRC 생성이 포함되어 있다.

본 논문에서 사용한 WiFi 전화기는 기존에 이미 나와 있는 모델 중에서 선택하고, 일부 프로그램 모듈을 수정하거나 추가하였다.

전화기는 음성 인식을 위하여 사용자가 특정 버튼을 눌렀을 때 바로 음성 인식을 수행할 수 있는 단계로 넘어가도록 구현되어 있으며, 또한 단말기는 DSR 기능 이외에 SIP 기반 전화기로서의 기능을 갖춘다. 실제 제품으로의 기능성을 갖추기 위해서는 휴대성을 고려한 최소한의 인터페이스가 필요하며, 고정된 장소에서의 사용보다는 이동이 잦은 환경에서 기능을 발휘할 수 있도록 802.11 무선망 환

경에서의 사용을 지원한다. 802.11은 기존의 802.x 프로토콜과 마찬가지로 완전한 분산형 모델을 지향한다. 또한 통제된 상태로 동작하는 것이 아니라 각각의 네트워크 단말들이 상호 중재를 하면서 사용하는 환경으로 시스템 구축에 많은 고려가 필요하지 않고 소규모부터 대규모까지 쉽게 구축할 수 있는 장점이 있다.

SIP/RTP를 지원하고 분산 음성 인식을 위한 전처리 기능이 포함된, 본 논문의 실험에 사용된 무선 VoIP 단말기의 구조는 <그림 3>과 같다.




<그림 3> 전처리 모듈을 포함한 VoIP 단말기의 구조도

분산 음성 인식 전처리 모듈은 특징 추출과 양자화를 수행하며, 호 관리는 음성 인식에 따른 사용자 인터페이스 및 호의 흐름을 제어한다. 네트워크 관리는 단말기가 무선 환경에서 사용자가 이동하면서 망에 접속할 수 있게 해 준다.

SIP/RTP는 단말기가 인식 서버나 다른 단말기와 데이터를 주고 받을 때 사용하는 프로토콜로서 SIP는 음성 인식 서버 또는 다른 단말기와의 연결을 위한 프로토콜이며, RTP는 실제 데이터를 주고 받을 때 사용되는 프로토콜이다[2,3].

입력된 음성으로부터 추출하는 특징 데이터는 10ms 단위로 이루어지며, 각각 14차의 특징 벡터가 생성되며 이들을 벡터양자화한 후 인코딩한다. 이들 데이터를 ETSI ES 202 212 코덱으로 전송할 때는 240ms에 해당하는 12 프레임 쌍을 하나의 패킷으로 만들어 RTP를 통해 전송한다. 패킹하는 방식은 RFC 3557 201 108 DSR RTP Paylod Format과 같으며[3], <표 1>은 두 개의 프레임에 해당되는 데이터를 나타내었다. idx(n,n+1)은 14차의 특징 벡터에서 n과 n+1 차수에 해당되는 데이터의 인덱스를 의미한다. Pidx와 Cidx는 피치 정보와 Voice Class의 인덱스를 나타내는 것으로서 음성 인식에는 사용되지 않고 특징 데이터로부터 입력된 음성을 재생할 때에 사용된다.

인식 모듈은 IVR 서버 내에 포함되어 있으며, 단말기로부터 전달받은 특징 추

출 데이터를 디코딩하고 이를 비터비 탐색에 사용한다. IVR 서버는 음성 인식을 하여 사용자가 원하는 곳으로 전화를 연결해 주는 시스템이다.

<표 1> Payload 형식

| 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |
|---|---|---|---|---|---|---|---|
| idx (2,3) | | idx (0,1) | | | | | |
| idx (4,5) | | | | idx (2,3) (cont.) | | | |
| idx (6,7) | | | | | | idx (4,5) (cont.) | |
| idx (10,11) | | idx (8,9) | | | | | |
| idx (12,13) | | | | idx (10,11) (cont.) | | | |
| idx (0,1) | | | | idx (12,13) (cont.) | | | |
| idx (2,3) | | | | | | idx (0,1) (cont.) | |
| idx (6,7) | | idx (4,5) | | | | | |
| idx (8,9) | | | | idx (6,7) (cont.) | | | |
| idx (10,11) | | | | | | idx (8,9) (cont.) | |
| idx (12,13) | | | | | | | |
| Pidx1 | | | | CRC | | | |
| Pidx2 | | | | | Pidx1 (cont.) | | |
| 0 | 0 | 0 | 0 | PC-CRC | | Cidx2 | Cidx1 |

## 3. 음성 다이얼링 서비스

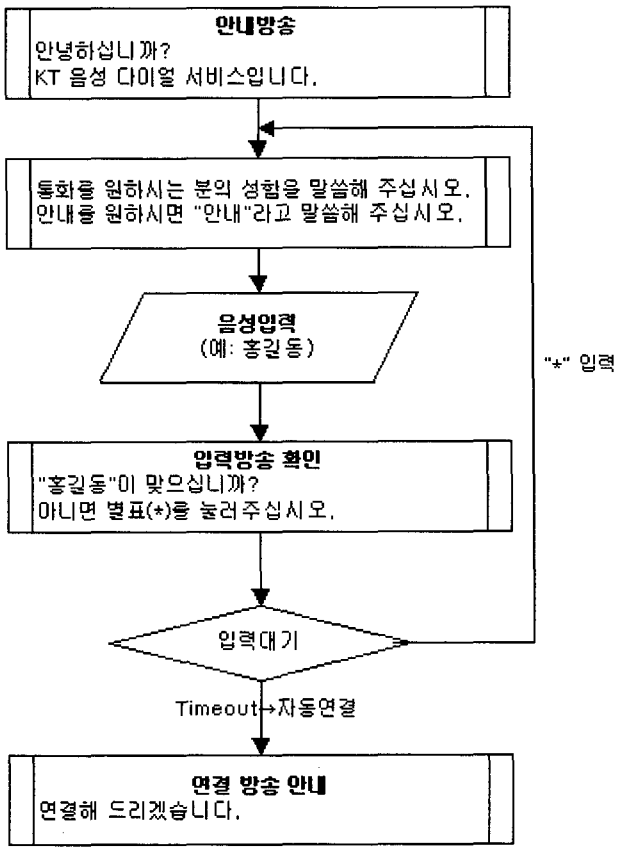
### 3.1 시나리오

WiFi 전화기를 이용하기 위해서는 우선 전화기를 IP-PBX에 등록하는 과정이 필요하다. VoIP 환경에서는 각 단말기의 식별을 IP 주소로 하기 때문에 IP-PBX 내에 각 단말기의 전화번호와 해당되는 IP 주소를 등록해 놓고 이들간의 연결을 가능하게 한다. 즉, 한 단말기에서 전화 번호를 입력하여 연결을 시도하면 IP-PBX에서는 그 전화번호에 해당되는 IP 주소를 가진 단말기와 연결을 하게 된다.

음성 인식을 이용하여 상대방과 전화 연결을 하기 위해서 WiFi 전화기에서는 특정 전화 번호를 누를 필요 없이 SEND 키만을 누른다. 즉, 기존의 PSTN을 이용

한 음성 다이얼링 시스템에서는 음성 인식 모듈이 있는 서버와 연결하기 위하여 그 서버에 해당되는 전화번호를 입력하는 과정이 필요했으나, VoIP 환경에서는 단지 연결하고자 하는 버튼만 누르면 된다.

본 논문에 제시된 시나리오는 <그림 4>에 나타내었으며, KT 내에서 운용되고 있는 유선 전화망에서의 음성 인식 자동 교환 서비스와 유사하다.



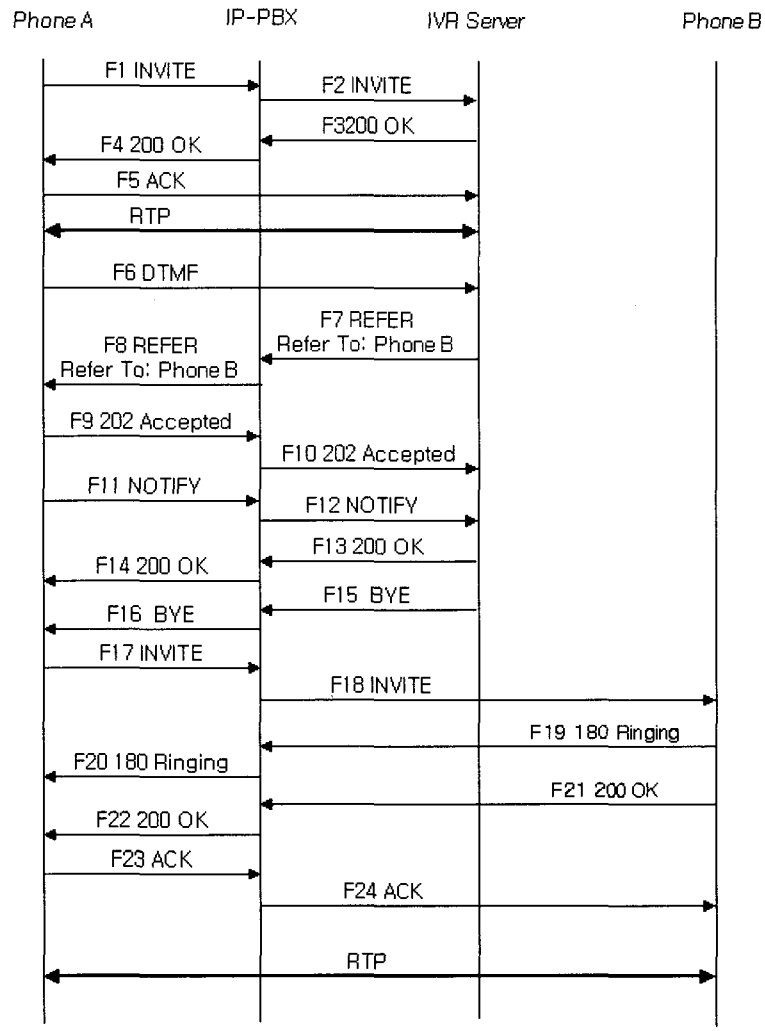

<그림 4> 음성 인식을 이용한 다이얼링 시나리오

IVR 서버와 연결되면 정해진 시나리오에 따라 통화하고자 하는 상대방의 이름을 말하고 IVR 서버는 이름을 확인한 후 해당되는 단말기로 호를 연결해 준다.

### 3.2 프로토콜

하나의 단말기에서 다른 단말기로 전화를 하기 위하여 본 논문에서 사용한 방

법은 호 전환 방식을 이용한 것으로서 일차적으로 단말기와 IVR 서버간에 연결이 되고, 그 다음에 IVR 서버가 단말기로부터 음성 입력을 받아들여 어느 단말기로 연결을 원하는지 확인한 다음, 해당되는 상대 단말기로 IVR 서버가 호를 전환하는 방식을 취한다. 이 과정을 <그림 5>에 나타내었다[4].




<그림 5> 기본 콜과 Refer를 이용하는 프로토콜

이 방식은 SIP의 기본 기능만을 이용하여 구현하므로 IP-PBX의 SIP 서버의 종류에 무관하게 상호 연동될 수 있는 장점이 있다. 각 단계를 살펴 보면 다음과 같다.

F1에서 F5까지는 Phone A와 IVR간의 세션을 연결하여 미디어 채널을 설정하는 단계로서 연결이 이뤄지면 IVR에서 Phone A로는 G.711 코덱을 사용하고 Phone A로부터 IVR까지는 DSR 전처리 모듈을 사용하여 데이터를 주고 받는다.

RTP를 통해 사용자의 음성이 전달되면 IVR 서버는 음성 인식을 수행하고 Phone A에서 DTMF 신호로 인식 결과가 맞음을 IVR 서버에게 알리면, F7에서 F24 과정을 거쳐 Phone B로 호를 전환한다. 그러면 RTP를 통해 Phone A와 Phone B간에 통화가 이루어진다.

## 4. 분산 음성 인식을 위한 특징 추출 실험

무선랜 환경에서 이동하면서 사용하는 WiFi 전화기의 경우, 크기에 제약을 받으며 계산 속도와 메모리가 제한되어 있다. 따라서 본 논문에서는 단말기에서 분산 음성 인식을 위한 전처리에 어느 정도의 시간이 소요되는지를 실험을 통해 알아 보았다.

단말기에 사용된 CPU는 모토롤라의 ColdFire V2 로서 클럭 주파수는 60Mhz이다. 음성의 샘플링 주파수는 8kHz이며 샘플링된 데이터는 16비트로 표현하였다. 고정 소수점 연산을 수행하며 32비트가 사용된다.

전처리 과정에서의 특징 추출은 25ms 의 프레임별로 이루어지며, 15ms씩 프레임을 중첩시킴으로써 매 10ms마다 특징 데이터가 생성된다. 특징 추출에 걸리는 시간을 줄이기 위하여 ETSI DSR 전처리 모듈에서 일부 연산의 결과값을 미리 테이블로 변환하여 사용하였다. 즉, 연산에서 나올 수 있는 값들을 테이블로 만들고, 실제 연산에서는 테이블의 값을 참조하여 사용하였다. 예를 들어 식 (1)과 같은 연산을 하나의 테이블로 생성하여 대신 사용하였다.

$$(\log(fPeriod) - LOG\_OF\_19) / DELTA\_WIDTH\_1 \quad (1)$$

식 (1)에서 LOG_OF_19, DELTA_WIDTH_1 은 상수이고 log(fPeriod)는 로그 함수이다. 일부 연산의 결과값을 테이블로 변환한 상태에서 한 프레임으로부터 특징 데이터를 추출하는 데 걸린 시간은 13.9ms이다. 한 프레임의 데이터를 처리하는 데 걸리는 시간이 10ms 이상 소요된다면 샘플링 시간보다 이를 처리하는 시간이 더 길기 때문에 샘플링되는 데이터가 분실되는 문제가 발생한다. 본 논문에서이 문제를 해결하기 위해 우선적으로 적용한 방법은 전처리 모듈에서의 작업을 백그라운드로 처리하는 것이다. 즉, 샘플링되는 데이터를 프레임별로 묶는 작업과 특징 추출하는 작업을 분리하여, 샘플 데이터를 버퍼에 모으는 작업은 끝점이 검출될 때까지 계속적으로 수행하고, 이 작업이 수행되지 않는 시간에는 특징을 추출

한다.

추가적으로 전처리 과정에서의 시간 단축은 전처리의 계산 일부를 제거함으로써 이루어진다. 전처리는 크게 잡음 제거, 피치 계산, 벡터양자화 및 기타 연산으로 나눌 수 있는데, 일부 연산의 결과값을 테이블로 변환한 상태에서 한 프레임으로부터 특징 데이터를 추출하는 데 각 부분별로 걸린 시간을 <표 2>에 나타내었다.

<표 2> WiFi 전화기에서 전처리에 소요되는 시간

| 분류 | 잡음제거 | 피치 계산 | 벡터양자화 및 기타 연산 | 전체 |
|---|---|---|---|---|
| 시간 (%) | 8.15ms (58.6%) | 3.9ms (28.1%) | 1.85ms (13.3%) | 13.9ms (100%) |

피치 계산은 서버 측에서 음성의 재생성에 사용되는 정보로서 ETSI ES 202 212 표준에는 포함되어 있지만, 음성 인식에 사용되지 않는다. 따라서 본 논문에서는 피치 계산 부분을 제거함으로써 추가적으로 시간을 단축시켰다. 피치 계산을 하지 않을 때, <표 2>에서 알 수 있는 것처럼 한 프레임에 대한 전처리 시간을 10ms에 맞출 수 있게 된다.

## 5. 결론 및 개선 방향

본 논문에서는 음성 인식 다이얼링을 무선랜 환경에 적용한 결과를 제시하였다. 단말기에서의 처리 시간을 줄이기 위하여 분산 음성 인식 방식을 채택하였으며, ETSI DSR 표준을 따랐다.

단말기와 IVR 서버간의 연결 설정은 SIP를 통해서 이루어지며, 데이터의 전달은 RTP를 사용한다. 음성 인식 과정은 IVR 서버에서 수행되며, 두 단말기간의 연결은 IVR 서버의 호 전환에 의해 이루어진다.

WiFi 전화기에서 한 프레임당 전처리하는 데 소요되는 시간을 줄이기 위해 일부 연산은 미리 결과값을 테이블로 만들어 사용하였으며, 이에 따른 전처리의 세부 내용별 소요 시간을 분석하였다. 그리고 ETSI ES 202 212 표준에는 포함되어 있으나, 음성 인식 과정에서는 사용하지 않는 피치의 계산을 제거함으로써 추가적으로 전처리에 소요되는 시간을 단축시켰다.

앞으로 처리 시간을 좀 더 단축하기 위하여 ETSI DSR 전처리 과정을 수정하고[5], 패킷 손실에 대한 처리도 보완할 필요가 있다[6][7]. 그리고 구현된 시스템

을 이용하여 음성 인식 성능 측정을 위한 실험을 수행할 것이다.

## 참 고 문 헌

[1] ETSI standard document, "Speech Processing, Transmission and Quality aspects (STQ) Distributed speech recognition Extended advanced front-end feature extraction algorithm Compression algorithm Back-end speech reconstruction algorithm", ETSI ES 202 212 v1.1.1, Nov. 2003.

[2] J. Rosenberg, H. Schulzrinne et al., RFC 3261 "SIP: Session Initiation Protocol", June 2002.

[3] Q. Xie, Ed., RFC 3557 "RTP Payload Format for European Telecommunications Standards Institute (ETSI) European Standard ES 201 108 Distributed Speech Recognition Encoding", July 2003.

[4] A. Johnston, S. Donovan et al., RFC 3665 "Session Initiation Protocol (SIP) Basic Call Flow Examples", Dec. 2003.

[5] J.-Y. Li, B. Liu et al., "A complexity reduction of ETSI advanced front-end for DSR", in *Proc. of ICASSP*, Vol. 1, pp.61-64, 2004.

[6] B. Delaney, "Increased robustness against bit errors for distributed speech recognition in wireless environments", in *Proc. of ICASSP*, Vol. 1, pp.313-316, 2005.

[7] A. James and B. Milner, "Soft decoding of temporal derivatives for robust distributed speech recognition in packet loss", in *Proc. of ICASSP*, Vol. 1, pp.345-348, 2005.

접수일자 : 2005년 8월 15일
게재결정 : 2005년 10월 30일

▶ 박성준(Sung-Joon Park) : 교신저자
주소: 137-792 서울특별시 서초구 우면동 17
소속: KT 마케팅연구소
전화: 02) 526-6771
E-mail: sjpak@kt.co.kr

▶ 구명완(Myoung-Wan Koo)
주소: 137-792 서울특별시 서초구 우면동 17
소속: KT 마케팅연구소
전화: 02) 526-5090
E-mail: mwkoo@kt.co.kr