핵심어 추출 기반 음성 다이얼링 시스템 개발

박전규[O]*, 서상원**, 한문성*
*시스템공학연구소 자연어정보처리부
** 시스템공학연구소 감성공학부

Development of Voice Dialing System based on Keyword Spotting Technique

Jeon-Gue Park[O]*, Sangweon Suh**, Mun-Sung Han*
*Natural Language Information Processing Dept., SERI
**Sensitivity Engineering Dept., SERI
e-mail: {jgpark,sangweon,mshan}@seri.re.kr

요 약

본 논문은 연속 분포 HMM을 사용한 핵심어 추출 기법(Keyword Spotting)과 화자 인식에 기반한 음성 다이얼링 및 부서 안내에 관한 것이다. 개발된 시스템은 상대방의 이름, 직책, 존칭 등에 감탄사나 명령어 등이 혼합된 형태의 자연스런 음성 문장으로부터 다이얼링과 안내에 필요한 핵심어를 자동 추출하고 있다. 핵심 단어의 사용에는 자연성을 고려하여 문법적 제약을 최소한으로 두었으며, 각 단어 모델에 대해서는 음소의 갯수 더하기 3~4개의 상태 수와 3개 정도의 mixture component로써 좌우향 모델을, 묵음 모델에 대해서는 2개 상태의 ergodic형 모델을 구성하였다. 인식에 있어서는 프레임 동기 One-Pass 비터비 알고리즘과 beam pruning을 채택하였으며, 인식에 사용된 어휘는 36개의 성명, 8개의 직위 및 존칭, 5개 정도의 호출어, 부탁을 나타내는 동사 및 그 활용이 10개 정도이다. 약 3~6개 정도의 단어로 구성된 문장을 실시간(1~3초이내)에 인식하고, 약 98% 정도의 핵심어 인식 성능을 나타내고 있다.

1 서 론

음성 다이얼링은 전화선을 통한 음성 인식 시스템의 실현이라는 관점에서 그 동안 국내외에서 가장 전형적인 음성 인식 및 그 응용시스템으로서 연구되어 왔다. 이러한 음성 다이얼링 기술은 기술적으로 극복해야할 다양한 목표가 집약되어 있는 것으로 알려져 있는데, 특히 다양한 잡음 요인, 화자 변이 등이 대표적이다[1][5][7][8].

한편, 회사, 연구소, 관공서 등에 전화를 거는 경우 수신자의 전화번호를 확보하지 못하는 한, 교환원의 도움을 받아야 원하는 사람과의 통화를 할 수 있었다. 이러한 제약은 교환원의 부재시 통화가 불가능하게 되고, 부서의 이동 등으로 해당 수신자의 번호가 변경되는 경우에 화자를 당황하게 하며, 효율적인 전화 교환 기능의 수행을 어렵게 한다.

본 논문에서는 현재 연속 음성 인식의 상용화에 가장 근접한 기술인 핵심어 추출 기술[4][6]을 응용한 음성 다이얼링 시스템에 대해 기술하고자 한다. 개발된 시스템은 컴퓨터 음성 신호 처리를 통한 자동 음성 인식 기술을 응용하여 화자의 발화문 중 통화를 원하는 사람을 인식하고, 그 인식 결과의 수신자에 대해 자동으로 전화 연결시켜 주는 음성 다이얼링 시스템이다. 또한 전화를 건 사람을 인식하여 반응하는 화자 인식 기능을 첨가 구현 하였다. 이러한 시스템은 교환원의 기능을 대체해주며, 24시간 안내 기능을 수행할 수 있어 그 유용성이 크며, 나아가 음성 메시지 시스템 또는 음성 전자 사서함 등에 직접적으로 활용될 수 있다.

2 시스템의 개요

음성 다이얼링은 그 특성상 전화선을 대상으로 한 불특정화자 연속 음성 인식 태스크이기 때문에 실용화 또는 상용화를 위해서는 여러가지 고려할 점이 많

다. 화자 변이와 채널 변이에 대한 대처는 그 핵심적인 내용이며, 본 논문에서는 전자에 중점을 두어 사용자로 하여금 가급적 자연스런 형태의 다이얼링 문장을 발성토록 유도하고 이를 대상으로 원하는 상대의 성명을 핵심어 추출하는 것을 목표로 하고 있다.

본 시스템에서는 인식 대상 어휘를 하나의 부서 단위로 제한하고 있는데, 즉 인식 대상이 되는 성명의 수를 40개 정도로 제한하고, 부서의 여러 수신자들에 대한 자연스런 형식의 전화 호출문을 인식하여 그 인식된 수신자와 자동 전화 연결을 수행할 수 있는 부서 안내를 겸한 음성 다이얼링 시스템을 제공함을 목적으로 한다.

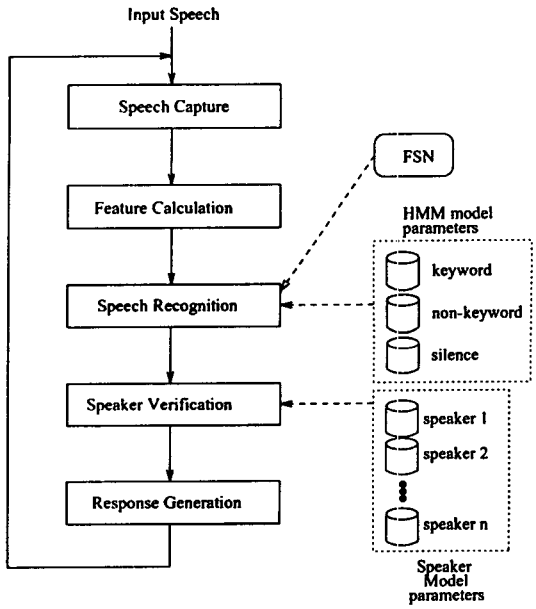

그림 1: 시스템의 구조

시스템의 주요한 부분은 그림 1에서와 같이, 음성 입력 및 특징 추출부, 음성 인식부, 화자 인식부, 및 결과 제시부 등의 네 개로 구성되어 있다. 음성 입력 및 특징 추출부에서는 마이크로폰에 입력되어 디지탈 신호로 변환된 음성 신호로부터 인식에 중요한 음성 특징들을 추출하는데, 이를 위해 선형 예측 분석법을 통해 캡스트럼, 동적인 특징을 나타내는 델타 캡스트럼을 구하고 음성 프레임의 에너지를 산출하여 인식 알고리즘의 입력이 되는 특징 벡터를 구성한다. 현재의 시스템은 사용자가 어느 순간에라도 서비스를 받을 수 있도록 마이크로폰의 입력 레벨을 체크하여 음성 입력이 가능하도록 하는 상시 대기(busy waiting) 기능을 구현하여 운영 중에 있다.

인식 시스템은 단어별 연속형 은닉 마코프 모델(HMM)을 기반으로 하였으며, 전화 호출문을 위한 문법 구성과 구성 어휘에 대한 학습용 음성 데이타를 학습 화자로부터 수집하여 HMM 학습을 수행했다. HMM 학습 모듈은 연속형 HMM의 Baum-Welch 알고리즘에 의한 파라미터 추정 절차이며[2], 핵심어 추출에 의한 음성 인식 부분은 프레임 동기 원패스(One-Pass) 빔(beam) 탐색[3][8]을 구현하며, 가비지 단어 모델링을 통하여 핵심 단어를 추출하는 융통성있는 문장 인식 기능을 수행하는 모듈이다.

문서 독립 HMM 학습을 통한 화자 인식 기능을 구현하여 인식기가 전화자의 신원을 파악하여 대응할 수 있게 하였는데, 화자 인식 모듈은 입력된 음성의 일부 또는 전부에 대해 비터비 알고리즘에 의한 스코어링을 통하여 표준적인 화자 분류를 실행한다.

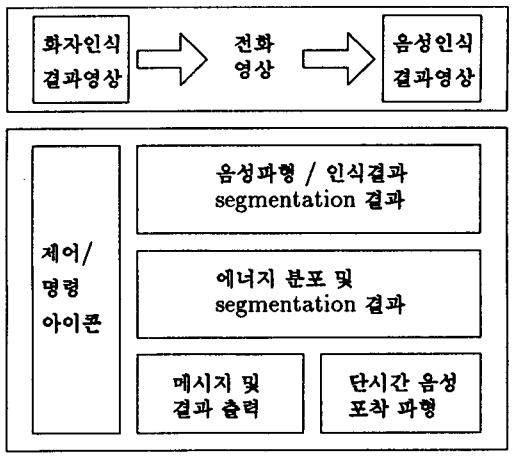

그림 2: 사용자 인터페이스

또한, 그림 2에서와 같이 인식 결과에 대한 그래픽 인터페이스를 통해 입력된 음성 신호의 파형과 에너지 및 검출된 음성 구간을 디스플레이하고, 인식 결과를

호출어	아	에	저	···
이름	박창호	박종만	김경옥	김재우
	황두성	전병태	주중원	이종현
	서상원	양영규	배창석	최경호
	김만희	강태호	하영렬	정한민
	이종훈	오원근	유병문	윤인숙
	강병호	김진서	김풍민	한문성
	정인숙	지수영	민병우	소정
	왕민	강석주	문경애	김현진
	박전규	임주희	윤호섭	조맹섭
직위	씨	님	연구원	선임
	책임	박사	실장	부장
부탁 어휘	요	부탁 합니다	좀 부탁 합니다	바꿔 주세요
	좀 바꿔 주세요	바꿔주 십시오	좀 바꿔주 주십시요	

표 1: 음성 다이얼링 시스템의 채택 어휘

문자 열로써 출력하도록 하였다. 아울러 인식된 화자와 화자가 원하는 수신자의 사진 영상을 디스플레이하여 전화 내용을 시각적으로 표현하도록 하였다.

3 문법 네트워크의 구성

음성 다이얼링용 문장을 위해 가정한 사용 단어들의 어휘는 표 1에 나타나 있으며 이들의 대체적 어순 또는 문법은 다음과 같이 가정하였다.

$$(\text{호출어}) \Longrightarrow \text{이름} \Longrightarrow (\text{직위}) \Longrightarrow (\text{부탁어휘})$$

괄호로 묶은 호출어, 직위, 부탁어휘 등은 임의로 생략할 수 있고 중간중간의 임의적인 무음(silence)의 삽입을 가정하였다.

그림 3은 채택된 문법의 세부 사항을 유한 상태 네트워크(Finite State Network)로 나타낸다. 채택된 어휘는 표 1에서와 같이 부서 성원의 이름들, 직위에 님 등이 붙은 자연스런 직위 표현 단어들, 전화 부탁 어휘와 그 활용어들, 그리고 다양한 감탄사 또는 간투사들로 구성되어있다. 그림에서 'B'로 레이블링된 노드는 문법 상의 시작 점들, 'E'는 문법 상의 끝

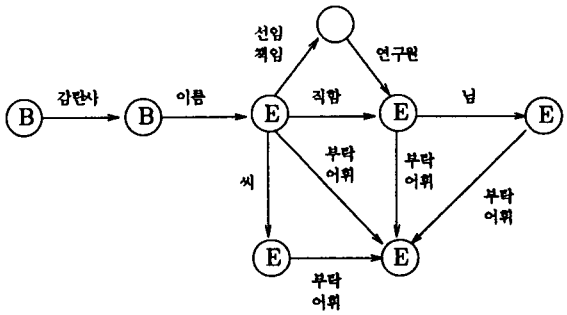

그림 3: 유한 상태 문법 네트워크 (FSN)

점(terminal)들을 나타낸다.

화자가 말한 문장 중 문법 상의 핵심 단어만을 추출하여 인식하는 프레임웍으로서, 화자의 발음 도중 삽입되는 무음(silence)과 배경 잡음 그리고 비핵심어들을 단어 모델로 처리, 연결단어 인식을 하여 출력된 단어열 중 핵심어를 취하도록 하고 있다. 인식을 위해 사용한 기본적인 탐색 방법론은 동적 프로그래밍(Dynamic Programming)의 응용인, 프레임 동기 빔 탐색 연결 단어 인식 알고리즘이며 단어 레벨과 단어 내 상태 레벨에서의 빔 프루닝 기능을 결합하여 구현하였다.

4 HMM 모델 구축과 학습

본 음성 다이얼링 시스템은 단어 모델을 기반으로 한 HMM의 구성과 학습으로 구축되었다. 표본화된 음성 신호 파형을 수동으로 segmentation 및 레이블링하여 단어별 학습 데이타를 구성하여 학습을 수행하였다. 입력 음성에 대해 8kHz의 sampling rate을 사용해서, 30ms의 프레임 크기에 대해 15ms씩 중첩시켜 12차 켑스트럼, 12차 차분 켑스트럼, 에너지, 차분 에너지로 구성되는 총 26 차원의 특징 벡터를 구성하여 사용하였다.

핵심어에 대해서는 left-right형, '부탁합니다'로부터 어미 변형된 호출어들에 대해서는 한데 묶어 ergodic형 모델링을 하였고, 각 단어별 HMM 상태 수는 음소 갯수 더하기 3~4 개로 정하였으며 상태 확률

밀도 함수의 mixture component 수는 대체로 3개 정도로 학습하였다. silence(무음) 모델은 2개 상태 ergodic형 HMM으로 하였다.

문서 독립적인 화자 모델의 학습은 ergodic형 HMM에 기반을 두어 수행한다. 상태의 갯수와 mixture component 수는 임의로 줄 수 있고, 각 상태에 해당하는 프레임들의 집합에 k-means clustering 알고리즘을 적용 클러스터링 한 뒤 각 클러스터의 평균 벡터와 공분산 행열, 각 클러스터에 속한 프레임들의 갯수의 비로써, 각 상태 혼합 가우시언 밀도함수의 평균 벡터, 공분산 행열, 혼합 확률 계수들의 초기 추정치로서 정하는 방식을 사용 하였다. 화자 인식시는 비터비 디코딩에 의한 점수 계산을 하여 최고 점수의 모델을 채택하는 표준적인 분류 방식을 사용하였으나, 우도 점수에 있어 최고치와 두번째 높은 점수의 차를 프레임 길이로 정규화하여 그 값이 일정치 이상 크지 않을 경우 화자 인식을 거부하는 기능도 첨가하였다.

5 실험 및 결과

본 시스템은 SUN SPARC 20, 내장 Ariel S-32C DSP 보드, ProPort-656 AD/DA 변환기, 및 SONY ECM-221 마이크로폰을 통해 실험되었다.

다음에 시스템의 온라인 실험결과를 예시한다.

- ...(5) 황두성(40) 선임(20) 연구원(34) 요(26) ...(1)
- 감탄사(35) ...(2) 양영규(34) 부장(22) 님(13) 부탁합니다(54) ...(1)
- ...(1) 박전규(35) 씨(14) 요(22) ...(1)
- 서상원(41)

위의 예에서 '...'은 무음 구간을, 괄호 안의 숫자는 비터비 디코딩을 거쳐 나타난 해당 단어의 지속 길이를 나타낸다. 여기서 보여지듯 문장의 시작과 끝 부분에 fading 효과가 나타나는 부분은 한두 프레임 길이의 무음 구간으로 인식되는 결과를 종종 보여주었다. 특히 중간 중간의 무음 삽입에 대처하는 인식은 만족스러우며, 단어들을 각각 격리적으로 띄워서 발음하는 식의 연속 음성 입력에 대해서도 무리없는 인식을 보여주고 있다.

탐색의 인식 성능은 핵심어인 이름을 디코딩하는데 있어 학습 화자의 정상적인 발음에 대해 98%의 인식율을 달성하고 있고, 화자 독립적으로도 이에 비견되는 높은 인식율을 보이고 있다. 문제점으로는 문장 중의 '님', '씨' 등의 단음절 어휘가 분명히 발음되지 않는 경우 생략되는 현상, 감탄사 발음 후 단어와 단어 사이에 수십 프레임 길이 이상의 긴 무음 구간을 두는 경우 인식율이 다소 떨어지는 현상 등이 있으나, 전체적으로 95% 이상의 전체 단어 인식율을 보이고 있다.

빔 탐색의 인식 시간은 같은 길이의 음성 입력에 대해서도 프루닝의 혼든 정도에 의존해서 약간씩 다를 수 있지만, 대체로 이름만 발음하는 등의 25~50 프레임 길이의 입력에 대해 1초 정도, 80~140 프레임 길이의 '0 0 0 연구원님 부탁합니다.' 등의 기본 문장에 대해 2~3 초, 감탄사와 긴 무음 구간 등을 포함하는 200 프레임 이상 길이의 입력에 대해 3~4 초 정도의 인식 처리 속도를 보이고 있다.

6 결 론

본 논문에서는 음성 인식의 실용화 측면에서 제한 어휘, 화자 독립의 음성 다이얼링 시스템을 구현하였다. 현재의 시스템은 약 40개 정도의 성명, 8개 정도의 직위 및 존칭, 10개 정도의 부탁 어휘와 그 활용에 대한 핵심어 추출을 수행하고, 감탄사 및 간투사의 활용에 유연히 대처하며, 화자 독립으로 약 98% 정도의 핵심어 인식 성능을 보이며, 전체적으로 약 95% 정도의 단어 인식율을 달성했다.

음성 다이얼링의 실용화를 위해서는 수백 개 수준의 고유 명사와 전화선에 내재하는 다양한 유형의 잡음에 대처하는 기술이 필수적이다. 현재 개발된 시스템은 약 100~200개 정도의 성명에 대처할 수 있는 시스템이며, 앞으로의 관건은 다양한 잡음 요인에 대처하는 적응형 신호 처리 기법과 특징 추출 기술 개발에 집중적인 연구를 수행할 예정이다.

감사의 글

이 글은 과학기술처에서 지원하는 특정 연구 사업의 연구 결과입니다.

참고 문헌

[1] H. Aust, M. Oerder, and F. Seide, "The Philips Automatic Train Timetable Information System", *Speech Communications*, vol. 17, pp. 249–262, 1995.

[2] L. R. Rabiner and B-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1993

[3] L. R. Rabiner and C. H. Lee, "A Frame-Synchronous Network Search Algorithm for Connected Word Recognition", *IEEE Trans. on ASSP*, vol. 37, pp. 1649–1658, Nov. 1989

[4] Y. Takebayashi *et al.*, "Keyword-spotting in Noisy Continuous Speech Using Word Pattern Vector Subtraction and Noise Immunity Learning", *ICASSP 92*, pp. II-85–88, 1992.

[5] G. Vysotsky, "VoiceDialing - the First Speech Recognition based Service Delivered to Custom's Fome from the Telephone Network", *Speech Communications*, vol. 17, pp. 235–247, 1995.

[6] J. G. Wilpon and L. Rabiner, "Automatic Recognition of Keyword in Unconstrained Speech Using Hidden Markov Models", *ICASSP 90*, pp. 1870–1878, 1990.

[7] 김재인, 구명완, "음성인식 증권정보시스템의 개발 및 시험운용결과 분석", 제 *13* 회 음성통신 및 신호처리 워크샵, pp. 185–191, 1996.

[8] 서상원, 박전규, 이종현, 김도석, 한문성, "연속분포 HMM에 의한 실시간 Word Spotting에 관한 연구", 제 *12* 회 음성통신 및 신호처리 워크샵, pp. 92–95, 1995.