DAB 용 음성 인터페이스 기술연구

최정규, 김규홍, 김원철, 한민수
한국정보통신대학원대학교

A study on speech interface technology for DAB

Jung-Kyu Choi, Kyu-Hong Kim, Won-Chul Kim, Minsoo Hahn
Information and Communications University

요 약

본 논문에서는 수년 내에 실용화될 것으로 예상되는 DAB (Digital Audio Broadcasting)에 필요한 음성 인터페이스 기술에 대한 기본 연구 결과를 소개한다. 연구의 시작 단계이므로 적용 분야는 고속도로 상에서의 교통 정보 안내 시스템으로 제한하였다. 즉 목표 시스템은 고속도로 상의 출발지와 목적지를 고립단어로 입력하면 시스템이 이를 인식한 후 미리 저장되어 있는 교통 정보 안내 text 중 해당 구간에 대한 정보를 추출하여 음성어로 사용자에게 들려 주는 것이다.
현재의 연구 결과는 상기 시스템 중 음성 인식 기능은 구현이 완료되었으며 교통 정보 안내는 아직은 문장으로 보여주는 수준이다. 향후 이를 편집 합성기를 이용하여 음성어로 들려 주는 연구를 금년 말까지 개발하여 전체 시스템에 대한 초벌 구현을 완료할 예정이다.
논문에서 소개될 내용은 전체 시스템 개념, 고립단어 인식 기술, 표본화 주파수 및 양자화 bit 수에 따른 인식율 변화, 최종 시스템 구현을 위한 향후 계획 등이다.

1. 서론

음성인식 및 합성기술은 인간과 컴퓨터 사이의 가장 자연스런 의사 전달 형태인 음성언어를 이용한 man-machine interface 로서 기계 또는 컴퓨터에 음성으로 명령하여 필요한 정보를 검색하고 컴퓨터는 검색된 정보를 음성으로 사용자에게 알려주는 기술이다. 따라서 음성 명령을 컴퓨터가 얼마나 잘 알아 듣느냐를 말해 주는 음성 인식율과 검색된 정보를 읽어주는 합성음의 자연성 및 명료도로 나타내어지는 합성음 품질은 음성 인터페이스의 입출력 단의 품질을 결정하게 되며 이는 사용자가 음성 인터페이스를 이용할 것인지 아닌지를 결정하는 주요한 요소가 된다. 본 논문에서는 DAB 음성 인터페이스의 구현에 대하여 기술한다.

2. 고립단어 인식기술

일반적으로 음성인식 시스템에는 벡터 양자화(Vector Quantization)를 이용하는 방법과 동적 시간정합(DTW)을 이용하는 방법, 신경회로망(Neural Network)을 이용하는 방법 은닉 마르코프 모델(Hidden Markov Model, HMM)을 이용하는 방법 등이 사용되고 있다[1][2][3][4].

벡터 양자화는 차원이 큰 입력 패턴을 양자화함으로써 차원을 줄이는 방법으로 음성 부호화 등에 널리 이용되며, 유사한 패턴 간의 군집화 특성을 이용하여 인식에도 사용되고 있다.[5] 동적 시간정합을 이용한 음성인식 방법은 패턴들을 비선형 신축에 의해 보다 유연하게 정합 시킴으로써 길이가 서로 다른 두 개의 패턴에서 최적의 정합 경로를 찾아 두 패턴을 서로 비교할 수 있는 방법을 제공한다.[6] 또한, 은닉 마코프 모델은 길이가 일정치 않은 시계열 패턴들을 확률적으로 모델화하는 방법으로 음성에 포함된 다양한 변이나 시간 정보들을 효과적으로 나타낼 수 있기 때문에 현재 음성 모델링의 방법으로 주축을 이루고 있다.[7] 마지막으로 신경 회로망은 패턴 인식 분야에서 다양하게 적용되는 방법으로 인간의

뇌 작용과 유사한 특징들을 가지고 있어 음성 분야에 있어서도 다양한 가능성을 보이고 있다.[8]

현재 실험실 수준의 음성인식 연구는 주로 은닉 마코프 모델을 이용한 음성인식의 연구가 활발하게 진행 중이며, 또한 최근에는 은닉 마코프 모델과 신경 회로망을 결합하여 음성인식에 적용하는 연구 사례가 많이 발표되고 있다. 하지만, 실제 상용 분야에서는 여전히 동적 시간정합을 이용한 음성인식 시스템이 가장 널리 사용되고 있다. 이는 동적 시간정합 알고리즘이 간단하고, 구현이 쉬우면서도 고립단어 인식에서 좋은 성능을 보여주고 있기 때문이다.[9]

본 논문에서는 12 차 LPC 켑스트럼 계수를 특징 벡터로 사용 하였다. LPC 계수는 음성 신호 혹은 음성 스펙트럼이 가진 특성을 상대적으로 적은 파라미터만으로 정확히 표현할 수 있는 장점을 가지고 있고 켑스트럼은 음성이 갖는 정보에서 스펙트럼 포락정보와 세부구조를 분리해 낼 수 있다는 특성을 가지고 있다.[4][9][10][11] 본 논문의 DAB 용 음성 인터페이스를 위하여 사용된 음성인식 알고리즘은 동적 시간정합알고리즘이었으며 특징벡터는 20 msec 프레임 크기로 10 msec 씩 shift 하며 구한 12 차 켑스트럼 계수를 사용하였다. 동적시간정합은 Vinsyuk, Sakoe 와 Chiba 에 의해 제안된 알고리즘으로 서로 다른 두개의 자료에서 비선형의 최적 정합경로를 찾아 서로 다른 길이의 특징 벡터를 비교하는 방법이다.[6]

3. 실험 방법

본 논문의 내용을 시뮬레이션하기 위하여 16bit 로 양자화하고 8kHz 로 표본화하여 고립단어 음성 데이터베이스를 구성하였으며 인식대상어휘는 표 1 에 보인 것과 같이 경부고속도로와 호남고속도로 상의 지명이다. 전체단어 개수는 49 개 이며 서울지역에서 성장한 남성화자 3 명 여성화자 3 명이 각각 2 번씩 발성하여 총 294 단어를 인식실험에 사용하였다. 전체 시스템의 구성은 사용자가 고립단어로 출발지와 목적지의 지명을 입력하면 음성인식을 수행한 후 해당 구간의 교통정보를 텍스트 형태로 화면에 보여주는 것이다. 즉 사용자가 "기흥"과 "오산"을 입력하였다면 화면에 "기흥에서 오산 구간은 차량이 증가하여 지체되고 있습니다." 또는 "기흥에서 오산 구간은 소통이 원활합니다." 등의 정보가 화면에 표시되는 것이다.

표 1. 인식실험에 사용된 단어 목록

1. 서울	2. 양재	3. 판교	4. 신갈
5. 수원	6. 기흥	7. 오산	8. 안성
9. 천안	10. 독립기념관	11. 목천	
12. 청주	13. 남이	14. 청원	15. 신탄진
16. 회덕	17. 대전	18. 옥천	19. 금강
20. 영동	21. 왕간	22. 김천	23. 구미
24. 왜관	25. 금호	26. 북대구	27. 동대구
28. 경산	29. 영천	30. 경주	31. 언양
32. 통도사	33. 양산	34. 구서	35. 부산
36. 유성	37. 서대전	38. 논산	39. 익산
40. 삼례	41. 전주	42. 김제	43. 금산사
44. 태인	45. 정읍	46. 백양사	47. 장성
48. 광산	49. 서광주		

4.표본화 주파수 및 양자화 Bit 수에 따른 인식율 변화

표본화율과 양자화 레벨에 따른 인식율을 조사하기 위하여 49 개의 단어에 대하여 인식율을 조사하였다. 표 2 부터 표 6 까지는 양자화 레벨을 8bit 로 표현한 경우와 12bit 로 표현한 경우 그리고 16bit 로 표현한 경우에 sampling rate 을 4kHz 에서 8kHz 까지 1kHz 씩 변화 시켜가며 인식율을 나타내며, 그림 1, 2, 3 은 sampling rate 의 변화에 따른 평균 인식율의 변화를 나타낸다. 본 실험결과에서 알 수 있듯이 평균 인식율의 변화는 입력 음성의 양자화 레벨이 어느 순간까지는 민감하지 않다가 12bit 이후에는 급격히 감소하는 성질을 볼 수 있다. 또한 sampling rate 의 변화에는 상대적으로 민감하지 않음을 알 수 있다. 이는 대부분의 음성 특징정보가 저주파에 치우쳐 있기 때문이다. 인식율 조사결과 구현시에 가장 경제적이면서도 비교적 우수한 성능을 나타내는 경우는 4kHz 로 Sampling 하고 12bit 로 양자화했을 경우이다. 단, 이 경우는 잡음이 없는 실험실 환경에서 수집한 데이터를 이용하여 실험을 하였으므로, 잡음환경에서의 인식실험 결과와는 거리가 있다. 따라서 본 연구에서는 추후 잡음에 대한 처리를 첨가하여 인식율을 조사할 예정이고, 잡음환경을 극복하기 위한 방법을 연구 중이다.

표 2. 8kHz sampling rate 일 때 인식율

양자화비트 / 화자	8BIT	12BIT	16BIT
여성화자 1	87.8%	98.0%	98.0%
여성화자 2	89.8%	93.9%	93.9%
여성화자 3	87.8%	93.9%	95.9%
남성화자 1	85.7%	95.9%	95.9%
남성화자 2	79.6%	93.9%	89.8%
남성화자 3	73.5%	91.8%	91.8%
평균 인식율	84%	94.6%	94.2%

표 3. 7kHz sampling rate 일 때 인식율

양자화비트 / 화자	8 BIT	12BIT	16 BIT
여성화자 1	83.7%	98.0%	98.0%
여성화자 2	87.8%	95.9%	93.9%
여성화자 3	87.8%	93.9%	95.9%
남성화자 1	89.8%	95.9%	95.9%
남성화자 2	79.6%	93.9%	89.8%
남성화자 3	77.6%	91.8%	91.8%
평균 인식율	84.4%	94.9%	94.2%

표 4. 6kHz sampling rate 일 때 인식율

양자화비트 / 화자	8 BIT	12BIT	16 BIT
여성화자 1	81.6%	95.9%	98.0%
여성화자 2	81.6%	93.9%	93.9%
여성화자 3	85.7%	95.9%	95.9%
남성화자 1	79.6%	85.7%	87.8%
남성화자 2	59.2%	93.3%	87.8%
남성화자 3	75.5%	95.9%	93.9%
평균 인식율	77.2%	93.5%	92.8%

표 5. 5kHz Sampling rate 일 때 인식율

양자화비트 / 화자	8 BIT	12BIT	16 BIT
여성화자 1	73.5%	98.0%	98%
여성화자 2	75.5%	95.9%	95.9%
여성화자 3	81.6%	93.9%	95.9%
남성화자 1	69.4%	85.7%	83.7%
남성화자 2	42.9%	87.8%	85.7%
남성화자 3	65.3%	100%	98%
평균 인식율	68.0%	93.6%	92.8%

표 6. 4kHz sampling rate 일 때 인식율

양자화비트 / 화자	8 BIT	12Bit	16 BIT
여성화자 1	57.1%	93.9%	91.8%
여성화자 2	77.6%	98.0%	95.9%
여성화자 3	67.3%	83.7%	79.6%
남성화자 1	59.2%	87.8%	85.7%
남성화자 2	38.8%	89.8%	85.7%
남성화자 3	57.1%	98.0%	98.0%
평균 인식율	59.5%	91.9%	89.45%

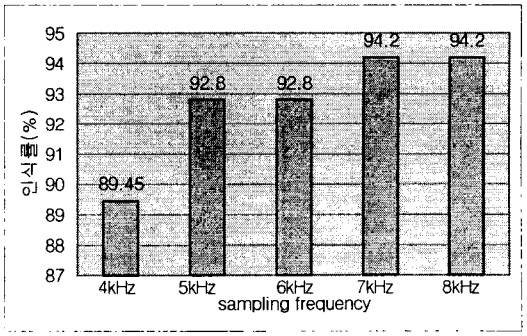

그림 1. 16bit 양자화시 인식율 변화

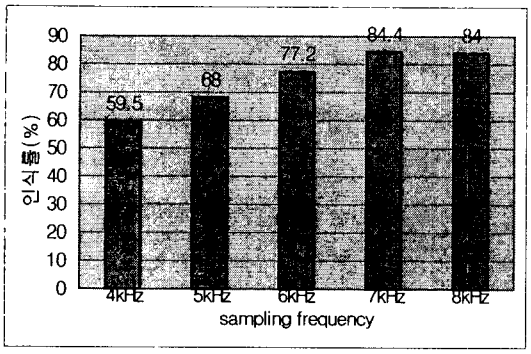

그림 2. 8bit 양자화시 인식률 변화

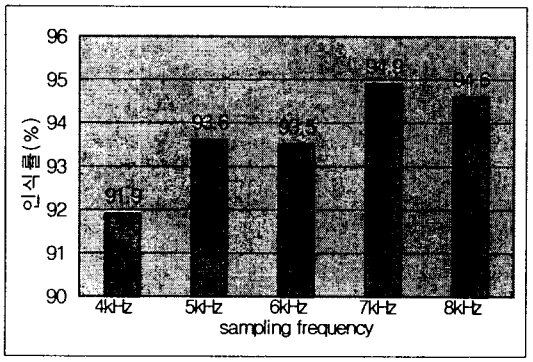

그림 3. 12bit 양자화시 인식율 변화

5. 결론 및 향후 계획

이상 기술한 바와 같이 본 논문에서는 DAB 용 교통정보 안내를 위한 음성 인터페이스 기술 개발에 대한 기본 연구 결과를 기술하였다. 즉 교통정보 서비스를 위한 구간 정보를 위하여 필요한 고속도로 상의 지명 단어음성 DB 를 구축하기 위하여 성인 남녀 각각 3 인에게 49 단어를 2 번씩 발성하게 하여 DAT 에 녹음하였고 고립단어 인식기술은 현재 잡음이 없는 실험실 환경에서의 테스트는 끝마친 상태로 단어 인식률은 약 94% 정도이며 구간에 해당하는 교통정보는 화면에 텍스트 형태로 표시되고 있다. 향후 수행 해야 할 연구 내용은 우선 구축된 DAB 용 응답문장 음성 DB 를 이용하여 편집합성기를 제작하여 텍스트 형태로 표시되는 구간별 교통정보를 음성정보로 사용자에게 들려주는 것이다. 한편 달리는 자동차 안에서와 같은 소음이 큰 환경에서의 음성 인식을 위하여 잡음제거 기술에 대한 연구도 지속적으로 수행되어야 할 것이다.

참고문헌

[1] W. Koenig, "A new frequency scale for acoustic measurements", Bell Telephone Lab. Record, Vol. 27, pp. 299-301, 1949.

[2] J.D. Markel and A.H. Gray, Jr., "Linear Prediction of Speech, Springer-Verlag", New York, 1976.

[3] A. M. Kondoz, "Digital Speech, Coding for Low Bit Rate Communications Systems", JOHN WILEY & SONS, 1994.

[4] S. Furui, "Digital Speech Processing, Synthesis, and Recognition", Marcel Dekker, 1992.

[5] R.M. Gray, "Vector quantization", IEEE ASSP Magazine, pp. 4-29. April, 1984.

[6] H. F. Silverman and D. P. Morgan, "The Application of Dynamic Programming to connected Speech Recognition", IEEE ASSP Magazine, pp 6-25, July 1990.

[7] L.R. Rabiner and B.H. Juang, "An introduction to hidden Markov models", IEEE ASSP Magazine, pp. 4-16, Jan. 1986.

[8] D.P. Morgan, C.L. Scofiled, "Neural networks and speech processing", Kruwer Academic Publishers, 1991.

[9] H. Sakoe and Chiba, "Dynamic programming optimization for spoken word recognition", IEEE Trans. Accoustics, Speech, Signal Proc., ASSP-26(1): 43-49, Feb. 1978.

[10] L.R. Rabiner, B.H. Juang, "Fundamentals of speech recognition", Prentice-Hall, 1993.

[11] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signal.", Trans. Committee on Speech Research, Acoust. Soc. Jap., S75-34, 1975.