내용기반 비디오 색인 및 검색을 위한 음성인식기술 이용에 관한 연구

A Study on the Use of Speech Recognition Technology for Content-based Video Indexing and Retrieval

손 종 목*, 배 건 성*, 강 경 옥**, 김 재 곤**
(Jong-Mok Son*, Keun-Sung Bae*, Kyeongok Kang**, Jae-Gon Kim**)

*경북대학교 전자전기공학부, **한국전자통신연구원 방송미디어연구부

(접수일자: 2000년 12월 7일; 채택일자: 2001년 1월 15일)

비디오 프로그램 색인 및 검색에 있어서 비디오 프로그램을 의미있는 부분으로 분할하는 것, 즉 내용기반 비디오 프로그램 분할은 중요하다. 본 논문에서는 내용기반 비디오 프로그램 분할을 위해 음성인식기술을 이용하는 새로운 방법을 제안한다. 제안한 방법은 음성신호와 캡션 (Closed Caption)의 정확한 동기를 위해 음성인식 기법을 사용한다. 실험을 통하여 내용기반 비디오 프로그램 분할을 위해 제안한 방법의 가능성을 확인하였다.

핵심용어: 비디오 정보 색인, 비디오 정보 검색, 폐쇄자막, 음성인식, HMM
투고분야: 음성처리 분야 (2.5)

An important aspect of video program indexing and retrieval is the ability to segment video program into meaningful segments, in other words, the ability of content-based video program segmentation. In this paper, a new approach using speech recognition technology has been proposed for content-based video program segmentation. This approach uses speech recognition technique to synchronize closed caption with speech signal. Experimental results demonstrate that the proposed scheme is very promising for content-based video program segmentation.

***Key words**: Video information indexing, Video information retrieval, Closed caption, Speech recognition, HMM*
***Subject classification**: Speech signal processing* (2.5)

1. 서 론

오늘날 디지털 장치 및 초고속 통신망의 발달과 더불어 디지털 멀티미디어 정보를 이용하려는 수요가 급속히 증가하고 있으므로, 방대한 멀티미디어 정보에서 사용자가 원하는 실질적인 내용을 빠르고 효율적으로 검색하는 것이 대단히 중요한 문제로 부각되고 있다. MPEG (Moving Picture Experts Group)에서는 이러한 문제들을 해결하기 위해 "Multimedia Content Description Interface" 표준화를 위한 연구에 착수했는데 이를 MPEG-7이라 한다. MPEG-7에서 정보의 기술은 크게 비디오/오디오 신호의 특징을 이용하는 저수준 (Low-level)에서의 기술과 의미론적 정보 (Semantic information)를 이용하는 고수준 (High-level)에서의 기술로 이루어진다. 점차 늘어가는 다양한 사용자의 요구를 수용하기 위해서 이와 같은 정보의 계층적 기술 기법과 내용에 기반한 효율적인 검색 및 색인 기법에 대한 중요성이 증대되고 있다.

내용에 기반한 멀티미디어 정보의 검색을 위해서 미리 비디오 정보를 요약하고 색인을 만드는 작업은 많은 시간과 비용을 소모하기 때문에 방대한 비디오 데이터의 효율적인 처리를 위해서 이를 자동화하는 것이 요구된다. 때문에 비디오/오디오 신호의 특징을 사용한 색인과 검색 기법들이 연구되고 있으며[1,2], 연속음성인식 기법이나 화자인식 기법등을 오디오 데이터의 색인과 검색에 이용하려는 시도도 활발히 이루어지고 있다[3,4,5,6].

최근에 드라마나 방송뉴스 등에 포함되는 경향이 증가하고 있는 캡션은 비디오 정보의 내용을 문자로 표현하고 있으므로 내용기반 비디오 검색 및 색인에 효율적으로 이용될 수 있다. 본 연구에서는 내용에 기반한 비디오 색인 및 검색을 위해 캡션정보와 음성인식기술을 이용하는 방법을 제안하고, 실제 방송된 뉴스 비디오 프로그램을 대상으로 한 실험결과를 제시한다. 본 논문의 구성은 다

책임저자: 배건성 (ksbae@ee.knu.ac.kr)
702-701 대구광역시 북구 산격동 1370
경북대학교 전자전기공학부
(전화: 053-950-5527; 팩스: 053-950-5505)

음과 같다. 2장에서는 내용기반 비디오 분할을 위해 캡션정보 및 음성인식기술을 이용하는 방법에 대해 설명하고, 3장에서 실제 방송된 뉴스 비디오 프로그램을 대상으로 한 실험결과를 제시하고 검토한다. 마지막으로 4장에서 결론을 내리고 향후 연구방향을 제시한다.

II. 내용기반 비디오 분할을 위한 캡션정보 및 음성인식 기술의 이용

2.1. 비디오 분할

드라마 및 방송뉴스 비디오 프로그램에 주어지는 캡션정보는 프로그램에 포함된 음성 정보 및 오디오 정보를 문자열로 표현해 준다. 때문에 캡션정보의 문자열을 이용할 경우 자연어 처리 기법을 적용하여 내용기반 비디오 프로그램의 분할을 용이하게 할 수 있으며, 이를 이용하여 비디오 프로그램의 색인 및 검색 작업을 효율적으로 수행할 수 있다. 그런데 비디오 프로그램에서 비디오 데이터와 음성 데이터는 동기가 이루어져 있지만 캡션정보는 그러하지 못하다. 특히, 실제 방송되고 있는 뉴스 비디오 프로그램의 경우에서는 음성신호에 비해 캡션정보가 상당한 시간이 경과한 후에 나타나는 경향을 보인다. 때문에 캡션정보를 내용기반 비디오 검색에 사용하기 위해서는 캡션정보에 해당하는 비디오/오디오 신호 구간을 정확하게 찾을 수 있어야 한다. 이를 위해 본 연구에서는 캡션정보에 대한 음성인식 과정을 수행하여 음성 신호구간을 찾고, 그에 대응되는 비디오/오디오 신호구간을 검출하였다. 캡션정보 및 음성인식 기술을 이용하여 본 연구에서 제안한 내용기반 비디오 색인 및 검색 시스템을 그림 1에 나타내었다. 우선 방송뉴스 비디오 프로그램으로부터 오디오 데이터와 캡션정보를 분리하고, 찾고자하는 키워드를 캡션정보에서 찾는다. 키워드에 대한 대략적인 시간정보와 문자정보를 이용한 음성인식 과정을 통해 정확한 비디오/오디오 신호구간을 검출하여 비디오 프로그램 분할에 이용한다. 그림 2는 캡션정보에서 찾고자 하는 키워드에 해당되는 음성구간을 검출하여 비디오 프로그램을 분할하는 예를 보인 것인데 그 과정은 다음과 같다.

STEP 1. 색인 또는 검색하고자 하는 내용의 키워드를 입력한다.

STEP 2. 캡션정보의 문자열에서 키워드를 검출하고, 키워드에 해당하는 캡션정보가 나타난 비디오 프레임의 시간을 찾는다.

STEP 3. 캡션정보가 오디오 데이터와 동기되어 있지 않으므로, 키워드에 해당하는 캡션정보가 나타난 비디오 프레임의 시간을 기준으로 오디오 데이터의 탐색 영역을 정한다. 일반적으로 방송뉴스 프로그램의 경우 오디오 신호보다 캡션이 2~7초 정도 늦게 나타난다.

STEP 4. 캡션정보에서 키워드가 나타난 주위의 문자정보를 이용하여 인식하고자 하는 음성모델을 구성한다.

STEP 5. (4)번 과정에서 구한 음성모델을 이용하여 (3)번 과정에서 정한 탐색영역에서 음성인식 기법을 적용하여 키워드에 해당하는 음성신호 구간의 시간정보를 검출한다.

STEP 6. (5)번 과정에서 구한 음성신호의 시간정보를 이용하여 비디오 데이터의 시작과 끝 프레임을 계산한다.

STEP 7. (6)번 과정에서 구한 시작과 끝 프레임 정보를 사용하여 비디오/오디오를 분할한다.

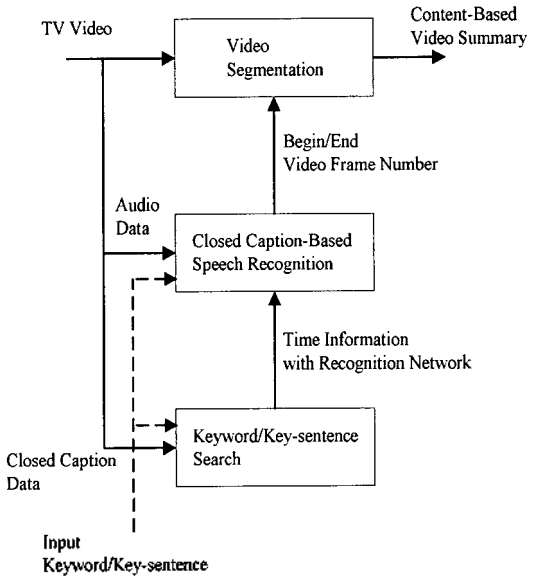


그림 1. 제안한 내용기반 비디오 분할 시스템
Fig. 1. The proposed content-based video segmentation system.

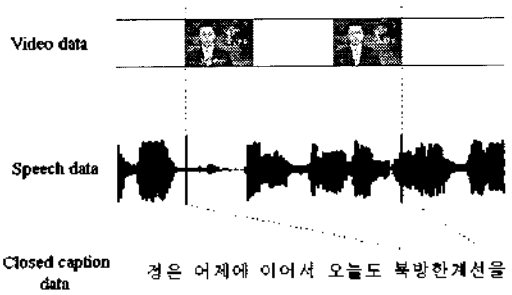


그림 2. 캡션정보를 이용한 비디오 분할 예
Fig. 2. The example of video segmentation using closed caption.

또한 입력된 키워드가 포함된 문장의 시작 어구와 끝 어구에 대응되는 음성신호를 검출함으로써 요약정보 생성에 보다 실용적으로 사용할 수 있는 문장단위의 검출을 할 수 있으며, 보다 효율적인 DB 생성을 위해서 위의 과정을 사용하여 전체 캡션정보를 비디오/오디오 데이터와 동기화할 수 있다. 본 연구에서는 실제 방송된 뉴스 비디오 프로그램을 대상으로 위의 기법을 적용하였다.

2.2. 음성인식 시스템 및 음성모델

오디오 트랙에서 얻어지는 음성 신호를 표본화율 16 kHz, 16bits/sample로 변환하여 preemphasis 계수 0.95로 전처리한 후, 20ms 길이의 해밍 윈도우를 10ms간격으로 오버랩하여 구간단위로 분석하였다. 각 구간에서 CMN (Cepstral Mean Normalization)과 MMSE (Minimum Mean-Square Error) STSA (Short-Time Spectral Amplitude Estimator)를 적용하여 1차의 에너지와 12차의 멜켑스트럼을 구하고[7,8], 현재 구간을 포함한 전후 각 3구간 (전체 7구간)의 정보를 이용하여 1차의 차분 에너지와 12차의 차분 멜켑스트럼을 구하였다. 음성모델은 음소모델을 연결시켜 구성하였다. 음소모델은 문맥종속 (Context Dependent) SCHMM (Semi-Continuous HMM)을 사용하였으며[9], 하나의 음소를 모델링하기 위하여 3상태 Bakis (Left-to-Right) 모델을 사용하였다. 기본 음소 모델로 44개를 설정하고 이를 확장하여 2000여 개의 문맥종속 모델을 인식기에서 사용하였다. 각 음소 모델은 K-Means 알고리듬과 Baum-Welch 재추정식을 사용하여 훈련하였으며[10,11,12], 훈련 데이터로 ETRI (Electronics and Telecommunications Research Institude)의 445DB (훈련용 남성 화자 16명, 여성 화자 15명)와 611DB (남성 화자 3명)를 사용하였다. 음성모델은 캡션정보로부터 키워드를 검출한 후 키워드 전후의 단어들을 연결하여 구성하였다.

주어지는 캡션 정보를 인식기에서 사용하기 위해서는 문자열을 발음 기호열로 표현해야 한다. 한글의 자소가 구체적인 음운현상을 반영한 것이 아니기 때문에 한글 자소를 발음기호로 나타내는 것이 필요한데 이를 위해서 자음동화, 된소리되기, 연음법칙, 음운축약, 구개음화, 끝소리 규칙 등의 한글 읽기 규칙을 적용하였다. 이에 voiced closure, voiceless closure, voice offset의 부가적 음운을 611DB의 레이블링에 나타난 통계적 특성을 사용하여 첨가함으로써 발음사전을 구성하였다. 발음사전에 등록된 단어들의 연결 순서를 나타내기 위해 인식망을 구성하는데, 기본적으로 캡션 정보의 문자열이 나타나는 시간순으로 연결순서를 정하였으며, 단어간 또는 문장간에 나타나는 묵음 구간을 표현하기 위해 묵음 모델을 첨가하였다. 또한 단어간에 묵음구간 없이 이어지는 부분이 많기 때문에 첨가된 묵음 모델의 생략도 허용하였다.

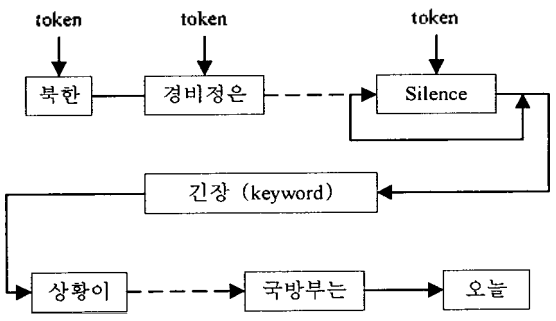


그림 3. 인식망 구성 예
Fig. 3. An example of the recognition network.

방송뉴스 프로그램의 /북한 경비정은 어제에 이어서 오늘도 북방한계선을 넘어오지 않았습니다. 긴장 상황이 이렇게 진정 국면에 접어들면서 국방부는 오늘/이라는 캡션정보에서 /긴장/이라는 키워드의 음성신호를 검출하기 위한 인식망 구성 예를 그림 3에 나타내었다.

음성모델을 이용하여 오디오 데이터의 탐색영역에서 키워드 부분을 검출하기 위하여 Viterbi 알고리듬을 사용하였으며, 계산량을 줄이기 위해 beam 탐색법과 pruning 기법을 적용하였다[12]. 오디오 신호에서 키워드 부분의 검출 과정은 아래와 같다.

STEP 1. 오디오 신호 탐색영역의 시작이 될 수 있는 인식망의 모든 어구에 초기화한 Token을 삽입시킨다. 일반적으로 Token은 인식망에서 키워드 이전의 모든 어구에 삽입된다. Token은 탐색영역에서의 시간정보와 확률적 유사도에 관한 정보, 그리고 과거 경로에 대한 정보를 가진다.

STEP 2. 오디오 데이터의 시간 t에서 Token이 존재하는 인식망의 각 상태 (State)에 연결된 모든 상태에 Token을 전파하고, Token정보를 갱신한다.

STEP 3. 각 상태에서 Token 정보를 확인해서 문턱값 이하의 우도를 가지는 Token을 제거한다.

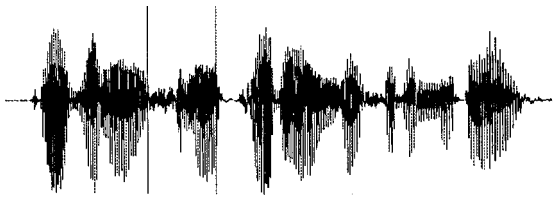
STEP 4. 시간 t가 오디오 신호 탐색영역의 끝에 도달할 때까지 (2), (3) 과정을 반복한다.

STEP 5. 오디오 신호에서 탐색영역의 끝이 될 수 있는 인식망의 모든 어구에서 도달한 Token의 우도를 확인하여 그 값이 가장 높은 것을 선택한다. 일반적으로 Token은 인식망에서 키워드 이후의 모든 어구에서 확인한다.

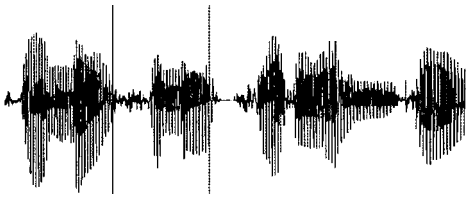
STEP 6. 선택된 Token의 경로 정보를 사용하여 오디오 신호의 키워드 부분을 검출한다.

III. 실험 및 검토

캡션정보 및 음성인식 기법을 이용하여 오디오 데이터에서 키워드 또는 키워드가 포함된 문장에 대응되는 음성신호 부분을 검출하는 실험을 수행하였다. 우선, 탐색영역의 변동에 따른 영향을 살펴보기 위해 /여기는 신호처리 연구실입니다./라는 음성신호에 대해 /신호/라는 키워드의 검출 실험을 하였다. 그림 4(a)는 전체 문장에서 키워드에 대응되는 음성구간을 검출한 예를 보인 것이며, 그림 4(b)는 탐색영역이 키워드는 포함하고 있지만 그 영역이 임의로 주어졌을 경우의 검출 예이다. 탐색영역이 다르지만 성공적으로 /신호/라는 키워드를 검출하고 있음을 볼 수 있다.



(a) /여기는 신호처리 연구실입니다/



(b) /기는 신호처리 연구/

그림 4. /신호/ 검출 예
Fig. 4. Examples of /신호/ detection.

실제 방송된 뉴스 프로그램에서 키워드 /북방한계선/에 해당하는 남성 앵커의 음성신호 검출 예를 그림 5에 나타내었으며, 여성기자의 음성신호에서 키워드 /김정일/의 검출 예를 그림 6에 나타내었다.

그림 5. 남성 앵커 방송뉴스에서 핵심어에 대응되는 음성구간 검출 예
Fig. 5. Examples of keyword detection from a male anchors voice in TV news.

그림 6. 여성 앵커 방송뉴스에서 핵심어 검출 예
Fig. 6. An example of keyword detection from an anchorwomans voice in TV news.

그림 5와 6의 검출 결과를 보면 비교적 정확하게 음성구간을 검출해 낸 것을 볼 수 있다.

실제 방송된 뉴스 프로그램의 데이터를 사용하여 수행한 검출 실험의 결과 57개의 핵심문장 검출 실험에서 56개의 핵심문장을 바르게 검출해 내었다. 핵심문장의 검출이 문장의 시작과 끝 어구를 검출함으로써 이루어지기 때문에, 이는 114개의 키워드에 대한 검출실험으로 볼 수 있다. 키워드 검출의 관점에서는 3개의 오검출이 발생하였으며, 이 오검출은 캡션정보에 실제 발생한 음성데이터가 포함되어 있지 않거나 배경잡음이 심한 기자 음성에서 발생하였다.

또한 전체 캡션정보를 비디오/오디오 데이터와 동기화하는 실험을 수행하였다. 그림 7은 캡션 디코더에서 출력되는 캡션정보와 시간정보 (Closed Caption with Time Code)의 예를 나타낸 것이고, 그림 8은 인식과정을 적용해 캡션정보의 실제 내용이 재생되는 비디오/오디오 시간 (Closed Caption with Audio-Synchronized Time Code)으로 시간 정보를 바꾼 예이다.

앵커: [0x 26E5] 한편 [0x 26F7] 당시 [0x 26FD] 수사 [0x 2703] 결과를 [0x 2717] 뒤집는 [0x 272B] 환전표가 [0x 2759] 발견되면서 [0x 2771] 당시 [0x 2783] 수사의 [0x 2797] 문제점을 [0x 27AD] 규명하기 [0x 27DB] 위한 [0x 27ED] 검찰의 [0x 2801] 행보가 [0x 281D] 빨라지고 [0x 2833] 있습니다.

그림 7. 캡션정보와 시간정보의 예
Fig. 7. An example of closed caption with time code.

한편 [4976, 4985] 당시 [4985, 4994] 수사 [4994, 5005] 결과를 [5005, 5021] 뒤집는 [5021, 5035] 환전표가 [5035, 5056] 발견되면서 [5056, 5076] 당시 [5076, 5085] 수사의 [5085, 5097] 문제점을 [5097, 5116] 규명하기 [5116, 5126] 위한 [5126, 5139] 검찰의 [5139, 5160] 행보가 [5160, 5175] 빨라지고 [5175, 5185] 있습니다. [5185, 5200]

그림 8. 캡션정보의 실제 내용이 재생되는 오디오/비디오 시간의 예
Fig. 8. An example of closed caption with audio-synchronized time code.

위의 실험 결과에서 제안한 방법의 타당성을 확인할 수 있다.

IV. 결 론

본 연구에서는 내용기반 비디오 색인 및 검색을 위해 캡션정보 및 음성인식기술을 이용하는 방법을 제안하고, 방송뉴스 프로그램에서의 키워드 검출 실험에서 97.37%의 검출율을 얻어 그 타당성을 확인하였다. 음성신호가 비디오 신호와 동기가 이루어져 있고 캡션이 비디오 프로그램의 내용을 충실히 담고 있기 때문에, 제안한 방법은 내용기반 비디오 프로그램 분할에 효율적으로 이용될 수 있다.

향후 다양한 비디오 프로그램에 적용할 수 있는 비디오 색인 및 검색 시스템 구현에 대한 연구가 이루어져야 한다.

감사의 글

본 연구는 한국전자통신연구원 방송미디어연구부의 지원으로 수행되었습니다. 지원에 감사드립니다

참 고 문 헌

1. John S. Boreczky and Lynn D. Wilcox, "A Hidden Markov Model Framework for Video Segmentation Using Audio and Image Features," *Int. Conf. on Acoustics, Speech and Signal Processing*, vol. VI, pp. 3741-3744, 1998.
2. Claude Montacie and Marie-Jose Caraty, "Sound Channel Video Indexing," *Proc. European Conf. on Speech Communication and Technology*, vol. 5, pp. 2359-2362, 1997.
3. Howard D. Wactlar, Alexander G. Hauptmann and Michael J. Witbrock, "IFORMEDIA™ : News-On-Demand Experiments In Speech Recognition", *Proc. of ARPA Speech Recognition Workshop*, pp. 18-21, 1996.
4. John Choi, Don Hindle, Julia Hirschberg, Ivan Magrin-Chagnolleau, Christine Nakatani, Fernando Pereira, Amit Singhal and Steve Whittaker, "An Overview of the AT&T Spoken Document Retrieval System," *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998.
5. Deb Roy and Carl Malamud, "Speaker Identification based Text to Audio Alignment for Audio Retrieval System," *Int. Conf. on Acoustics, Speech and Signal Processing*, vol. II, pp. 1099-1102, 1997.
6. Ivan Magrin-Chagnolleau, Aaron E. Rosenberg and S.Parthasarathy, "Detection of Target Speakers in Audio Databases," *Int. Conf. on Acoustics, Speech and Signal Processing*, vol. II, pp. 821-824, 1999.
7. Alejandro Acero, "Acoustical and Environmental Robustness in Automatic Speech Recognition," Ph. D. thesis, CMU, 1990.
8. Yariv Ephraim and David Malah, "Speech Enhancement Using a Minimum Mean-Square Erro Short-Time Spectral Amplitude Estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109-1121, Dec., 1984.
9. C.-H. Lee, L. R. Rabiner, R. Pieraccinit, and Jay G. Wilpon, "Acoustic Modeling of Subword Units for Speech Recognition," *Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 2, pp. 721-724, 1990.
10. B.-H. Juang, L. R. Rabiner, "The Segmental K-Means Algorithm for Estimation Parameters of Hidden Markov Models," *IEEE Trans. on Acoustics, Speech and Signal Processing IEEE Trans. On ASSP*, vol. 38, no. 9, pp. 1639-1641, Sep., 1990.
11. L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. of IEEE*, vol. 77, no. 2, pp. 257-286, Feb., 1989.
12. Mosur K. Ravishankar, "Efficient algorithms for Speech Recognition," Ph. D. thesis, CMU, 1996.
13. 손종목, 배건성, "HMM 인식기에서 상태별 다중 특징 파라미터 가중," 한국음향학회지, 제 18권, 제 4호, pp. 47-52, 1999.

▲ 손 종 목 (Jong-Mok Son)

1997년 2월 : 경북대학교 전자공학과 공학사)
1999년 2월 : 경북대학교 전자공학과 (공학석사)
1999년 3월 ~ 현재 : 경북대학교 전자공학과 박사과정 재학 중
※ 주관심분야 : 디지털 신호처리, 음성신호처리, 음성인식

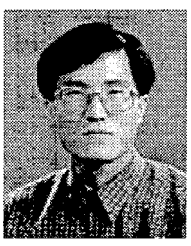
▲ 배 건 성 (Keun-Sung Bae)

1977년 2월 : 서울대학교 전자공학과 (공학사)
1979년 2월 : 한국과학기술원 전기 및 전자공학과 (공학석사)
1989년 5월 : University of Florida (공학박사)
1979년 3월 ~ 현재 : 경북대학교 전자공학과 교수
※ 주관심분야 : 음성분석 및 인식, 디지털 신호처리, 디지털 통신, 음성 부호화, 웨이브렛 분석 등

▲ 강 경 옥 (Kyeongok Kang)

1985년 2월 : 부산대학교 물리학과 (이학사)
1988년 2월 : 부산대학교 물리학과 (이학석사)
1991년 2월 ~ 현재 : 한국전자통신연구원 방송미디어연구부 선임연구원
※ 주관심분야 : 음향신호처리, 오디오 부호화 및 MPEG-7

▲ 김 재 곤 (Jae-Gon Kim)



1990년 2월 : 경북대학교 전자공학과 (공학사)
1992년 2월 : 한국과학기술원 전기 및 전자공학과 (공학석사)
1992년 3월 ~ 현재 : 한국전자통신연구원 선임연구원
※ 주관심분야 : 영상처리, 비디오 색인 및 검색