Inseon Jang, Panos Kudumakis,
Mark Sandler, and Kyeongok Kang

The MPEG Interactive Music Application Format Standard

The music industry is going through a transformation, and new interactive music services have emerged. It is envisaged that this new concept of digital music content will dominate the next generation of music services. A standardized file format is inevitably required to provide the interoperability between various interactive music players and interactive music albums. This issue is addressed in a new standard by the Moving Picture Experts Group (MPEG), known as the MPEG-A Interactive Music Application Format (IM AF). IM AF integrates multiple audio tracks with appropriate additional information, enabling users to experience various preset mixes and to make their own mixes complying with interactivity rules imposed by the music composers with the aim of fitting their artistic creation.

BACKGROUND

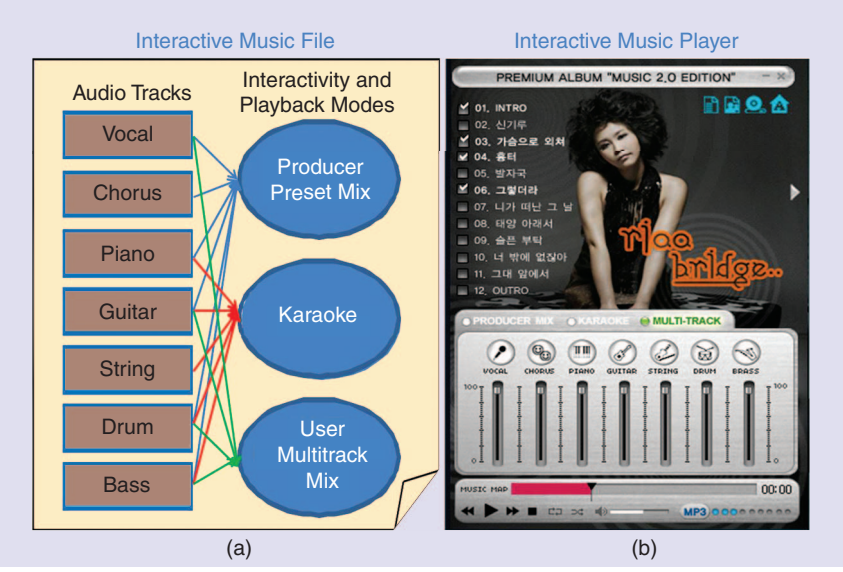
MOTIVATION

The exploding popularity of file sharing Web sites has challenged the music industry's traditional supply model that relied on the physical distribution of music recordings such as vinyl records, cassettes, and CDs [1], [2]. Music fans are disappointed by a lack of a rich user experience, while music creators are disappointed by lack of content governance [3]. In this direction, new interactive music services have emerged [4], [5]. This type of multitrack interactive music allows users not only to listen to music as usual but also to create a karaoke version and sing-along, isolate one or more musical instruments and practice their musical instrument skills for training purposes, or even create their own user-generated content, e.g., emphasizing the melody and harmony or rhythm depending upon their personal taste and legally share it with friends.

It should be noted that regarding the latter requirement on user-generated content, the IM AF is restrictive from the point of view that it imposes some interactivity rules; new mixes cannot exceed certain limits, that is, the initial song might remain recognizable, if the music composer so desires. On one hand, this is considered as a win-win case enabling music fans to experience rich interactive music services, while on the other hand, it respects the creators' wishes for music tracks governance and traceability (e.g., through filtering technologies) and their fair remuneration.

However, IM AF interactivity rules should not be confused with enforceable Digital Rights Management. The IM AF interactivity rules are motivated and defined by the music composers with the aim of fitting their artistic creation. For example, a composer might not wish his/her special guitar solo to be completely eliminated in a user-generated mix or his/her rock version to be mixed in a pop style. It should also be understood that the IM AF interactivity rules definition is optional and up to the music composer and is not imposed by the IM AF file format.

It is envisaged that this new concept of digital music content, by providing an environment that activates and stimulates passive listeners to become creative music producers, would not only attract music fans but also create new revenue opportunities for all parties involved in the music value chain. However, a standardized file



[FIG1] Conceptual model of an MPEG-A IM AF player. Part (a) shows the structure of the interactive music file format while part (b) shows a multritrack music player with volume sliders for each instrument/vocal track as well as playback modes.

Digital Object Identifier 10.1109/MSP.2010.939073
Date of publication: 17 December 2010

1053-5888/11/$26.00©2011IEEE

format is inevitably required to provide the interoperability between various interactive music players and interactive music albums. The new MPEG-A IM AF [6], [7], addresses this issue by taking into consideration both music fans' and creators' wishes for rich user experience and content governance.

APPLICATIONS

In particular, IM AF targets applications offering the following functionalities:

- *Enhanced listening user experience*: The interactive music player allows the user to select the number of instruments in the mix and to control their volume, in addition to preset mixes of tracks.
- *Learning and performing experience*: Practicing musical instrument (i.e., guitar) skills becomes possible by setting off the sound of the corresponding instrument and playing along. The karaoke version is also supported for practicing vocals.
- *Producer experience via user-generated content*: Add your own musical instrument/vocal track or substitute the existing ones and listen to the resulting mix.

Eventually, the multitrack interactive nature of IM AF combined with other emerging technologies, i.e., content-based search and retrieval and three-dimensional (3-D) audio/video, it is believed that would enable an unimaginable number of music applications for both fans and professionals.

Figure 1 shows the conceptual model of an IM AF player. In part (a), the structure of the interactive music file format is shown, which contains a number of musical instrument and vocal tracks enabling users to mix them to their own taste, along with preset mixes offered by the music producer. In part (b), a multitrack music player is shown with volume sliders for each instrument/vocal track as well as a number of playback modes: producer mix, karaoke, and multitrack-user mix.

ISSUING BODY AND SCHEDULE

Upon a number of documented market value-added propositions by the music industry, and after extensive collection of requirements and sufficient support in terms of resources from a numerous national bodies, International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC) MPEG decided to embark on a multitrack music file format and incorporate it in ISO/IEC 23000 (MPEG-A) family of MPEG standards, also known as Multimedia Application Formats (MAFs). MAFs specifications integrate elements from many different MPEG and non-MPEG standards into a single specification that is useful for specific (but very widely used) applications. ISO/IEC 23000-12 IM AF was developed under the auspices of both the Systems and Audio MPEG Groups. A working draft was issued in October 2008. The IM AF specification was finalized and published in July 2010, while its reference software and conformance files are planned to be finalized in early 2011.

[TABLE 1] SUPPORTED COMPONENTS IN IM AF.

TYPE	COMPONENT NAME	ABBREVIATION	SPECIFICATION
FILE FORMAT	ISO BASE MEDIA FILE FORMAT	ISO-BMFF	ISO/IEC 14496–12:2008
AUDIO	MPEG-4 AUDIO AAC PROFILE	AAC	ISO/IEC 14496–3:2005
	MPEG-D SAOC	SAOC	ISO/IEC 23003–2:2010
	MPEG-1 AUDIO LAYER III	MP3	ISO/IEC 11172–3:1993
	PCM	PCM	-
IMAGE	JPEG IMAGE	JPEG	ISO/IEC 10918–1:1994
TEXT	3GPP TIMED TEXT	3GPP TIMED TEXT	3GPP TS 26.245:2004
METADATA	MPEG-7 MULTIMEDIA DESCRIPTION SCENE	MDS	ISO/IEC 15938–5:2003

TECHNOLOGY

SUPPORTED COMPONENTS

IM AF involves formatting different types of media data, especially multiple audio tracks with interactivity data and storing them into an ISO-Base Media File Format. An IM AF file is composed of:

- multiple audio tracks: representing the music (e.g., instruments and/or voices)
- groups of audio tracks: a hierarchical structure of audio tracks (e.g., all guitars of a song can be gathered in the same group)
- preset data: predefined mixing information on multiple audio tracks (e.g., karaoke and rhythmic version)
- user mixing data and interactivity rules: information related to user interaction (e.g., track/group selection and volume control)
- additional media data: that can be used to enrich the user's interaction space (e.g., timed text synchronized with audio tracks which can represent the lyrics of a song, images related to the song, music album, and artist)
- metadata: data used to describe a song, music album, and artist.

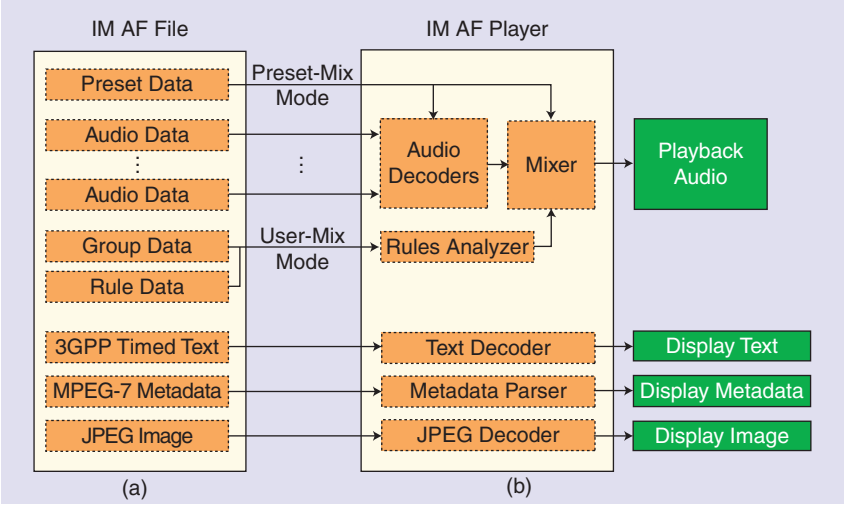
Table 1 lists the components supported by IM AF for formatting the aforementioned data types .

ARCHITECTURE

The IM AF player allows users to remix music tracks by enabling them to select the number of instruments to be listened and adjust their volume to their taste. Thus, IM AF enables users to publish and exchange this remixing data, enabling other users with IM AF players to experience their particular music taste creations (user-generated content). Preset mixes of tracks are also available. Figure 2 shows the architecture of an IM AF player. In part (a), the different types of media data (3rd Generation Partnership Project timed text, MPEG-7 metadata, and Joint Photographic Experts Group images) supported in the IM AF file format are shown. In part (b), the IM AF player is shown including the corresponding decoders/parsers of the media data in the IM AF file format. The operational modes and the interactivity rules are further described in the following section.

OPERATIONAL MODES AND INTERACTIVITY RULES

In particular IM AF supports the following two possible mix modes for interaction and playback:



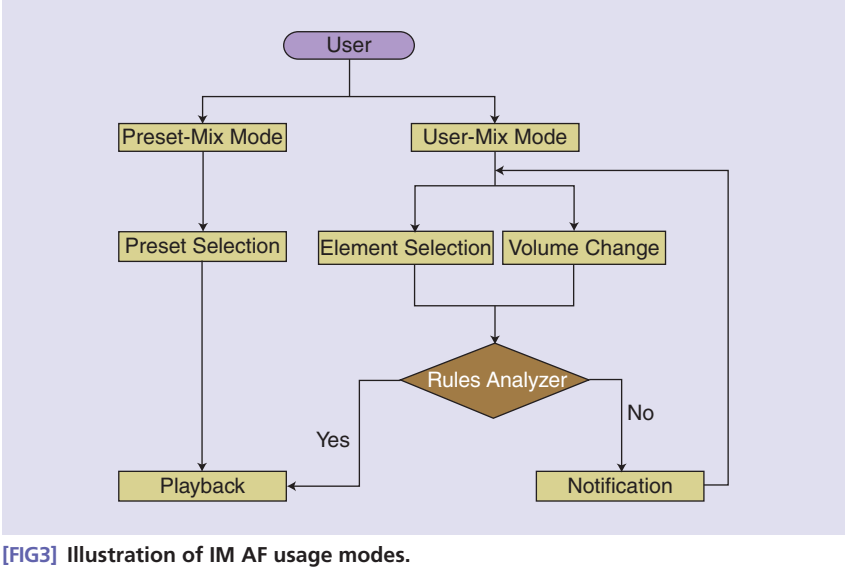
[FIG2] IM AF architecture: (a) The different types of media data supported in the IMAF file format are shown. (b) The IM AF player is shown including the corresponding decoders/parsers of the media data in the IM AF file format.

- preset-mix mode
- user-mix mode.

In the preset-mix mode, the user selects one preset among the presets stored in IM AF and then the audio tracks are mixed using the preset parameters associated with the selected preset and played. Both static and dynamic volume information for each track/group and their individual channels are supported. Some preset-mix mode examples are the following:

- general preset mix: composed of multiple audio tracks mixed by the music producer
- karaoke preset mix: composed of multiple audio tracks except vocal tracks
- a cappella preset mix: composed of vocal and chorus tracks.

In the user-mix mode, the user selects/deselects the audio tracks/groups and/or controls the volume of each of them. Thus, in the user-mix mode, audio tracks are mixed according to a user's control and taste; however, they should comply with the interactivity rules stored in the IM AF. Note that the interactivity rules allow the music producer to define the amount of freedom available in IM AF users' mixes. The interactivity rules analyzer in the player, shown in Figure 2, verifies whether the user interaction conforms to music producer's rules. Figure 3 depicts in a block diagram the logic for both the preset-mix and user-mix usage modes.



[FIG3] Illustration of IM AF usage modes.

User interaction should conform to certain rules defined by the music composers with the aim of fitting their artistic creation. However, the rules definition is optional and up to music composer and is not imposed by the IM AF format.

In general there are two categories of rules in IM AF: the selection and the mixing rules. The selection rules related to the selection of the audio tracks and groups at rendering time whereas the mixing rules related to the audio mixing.

IM AF supports four types of selection rules, as follows:

- *Min/max rule* specifying both minimum and maximum number of track/groups of a group that might be in active state for playback.
- *Exclusion rule* specifying that several track/groups of a song will never be in the active state at the same time.
- *Not mute rule* defining a track/group that should always be in active state.
- *Implication rule* specifying that the activation of a track/group implies the activation of another track/group.

IM AF also supports four types of mixing rules, as follows:

- *Limits rule* specifying the minimum and maximum limits of the relative volume of each track/group.
- *Equivalence rule* specifying an equivalence volume relationship between tracks/groups.
- *Upper rule* specifying a superiority volume relationship between tracks/groups.
- *Lower rule* specifying an inferiority volume relationship between tracks/groups.

BRANDS

IM AF also supports a number of brands according to application domain; these

[TABLE 2] BRANDS IN IM AF.

BRANDS	AUDIO				MAXIMUM NUMBER OF SIMULTANEOUSLY DECODED AUDIO TRACKS	MAXIMUM SAMPLING FREQUENCY/BITS	PROFILE/LEVEL	APPLICATION
	AAC	MP3	SAOC	PCM				
IM01	°	°			4	48 kHz/16 b	AAC/LEVEL 2	MOBILE
IM02	°	°			6			
IM03	°	°			8			
IM04	°	°	°		2		AAC/LEVEL 2 SAOC BASELINE/2	
IM11	°	°		°	16		AAC/LEVEL 2	NORMAL
IM12	°	°	°		2		AAC/LEVEL 2 SAOC BASELINE/3	
IM21	°			°	32	96 kHz/24 b	AAC/LEVEL 5	HIGH END

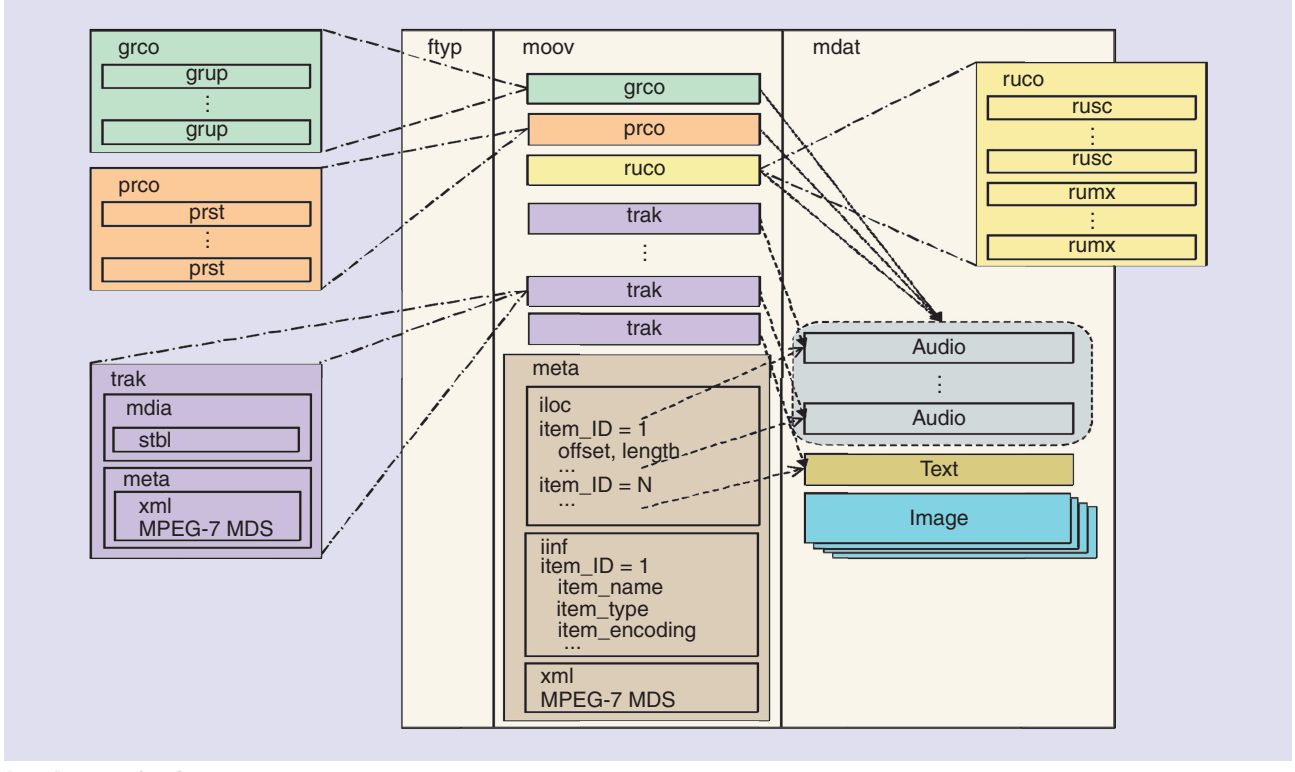
[Remark 1] The audio component data marked as "°" may exist in the file.
[Remark 2] For IM04 and IM12, simultaneously decoded audio tracks consist of tracks related to SAOC, which are a downmix signal and SAOC bit stream. The downmix signal shall be encoded using AAC or MP3.
[Remark 3] For all brands, the maximum channel number of each track is restricted to two (stereo).

are depending on the device processing power capabilities (mobile phone, laptop computer and high-fidelity devices) that consequently define the maximum number of audio tracks that can be decoded simultaneously in an IM AF player running on a particular device. IM AF brands are summarized in Table 2. In all IM AF brands, associated data and metadata are supported.

Backwards compatibility with legacy noninteractive music players is also supported by IM AF. For legacy music players or devices that are not capable of simultaneous decoding the multiple audio tracks, a conventional mixed/mastered audio track stored in IM AF file can still be played.

FILE STRUCTURE

The IM AF file format structure is derived from the ISO-Base Media File Format standard. As such facilitates interchange, management, editing, and presentation of different type media data and their associated metadata, in a flexible and extensible way. The object-oriented nature of ISO-Base Media File



[FIG4] IM AF file format structure.

Format, inherited in IM AF, enables simplicity in the file structure in terms of objects that have their own names, sizes, and defined specifications according to their purpose.

Figure 4 illustrates the IM AF file format structure. It mainly consists of "ftyp," "moov," and "mdat" type information objects/boxes. The "ftyp" box contains information on file type and compatibility. The "moov" box describes the presentation of the scene and usually includes more than one "trak" boxes. A "trak" box contains the presentation description for a specific media type. A media type in each "trak" box could be audio, image, or text. The "trak" box supports time information for the synchronization with other "trak" boxes described media. The "mdat" box contains the media data themselves described in the "trak" boxes. However, a "trak" box may also include a URL where from the media data could be imported. In this way the "mdat" box maintains a compact representation enabling consequently, efficient exchange and sharing of IM AF files.

Furthermore, in "moov" box some specific information is also included, such as the group container box "grco", the preset container box "prco" and the rules container box "ruco" for storing group, preset and rules information, respectively. The "grco" box contains, zero or more group boxes "grup" describing the group hierarchy structure of audio tracks and/or groups. The "prco" box contains one or more "prst" boxes that describe the predefined mixing information, in the absence of user interaction. The "ruco" box contains zero or more selection rules boxes "rusc" and/or mixing rules boxes "rumx" describing the interactivity rules related to selection and/or mixing of audio tracks.

FURTHER TECHNICAL DEVELOPMENTS

Due to the music industry's extreme interest in IM AF, there has already been a request for additional functionality to be incorporated in the IM AF standard, in terms of an amendment. This functionality is related to audio equalization (EQ) information aiming to offer users a complete music producer experience.

Audio EQ settings can be stored as presets in the preset-mix mode or specified by users in user-mix mode. Tweaking the EQ parameters directly in an IM AF player would be a similar process to the traditional producer-based music mixing environment. Alternatively, EQ settings could be applied automatically, based on a user's saved EQ profile, for masking reduction among different instruments and/or boosting or attenuating certain frequency bands of particular instruments, depending upon the user's preferences in music styles, personal taste, or mood.

AUTHORS

Inseon Jang (jinsn@etri.re.kr) is with the Realistic Acoustics Research Team at the Electronic and Telecommunications Research Institute (ETRI). She is a coeditor of the IM AF standard.

Panos Kudumakis (panos.kudumakis@eecs.qmul.ac.uk) is research manager at the Centre for Digital Music, Queen Mary University of London, United Kingdom. He has been an active member of the MPEG Committee since 1998.

Mark Sandler (mark.sandler@eecs.qmul.ac.uk) is a professor and director at the Centre for Digital Music, Queen Mary University of London, United Kingdom.

Kyeongok Kang (kokang@etri.re.kr) is a director at the Realistic Acoustics Research Team of the Electronic and Telecommunications Research Institute (ETRI).

RESOURCES

REFERENCES

[1] "Digital music report 2009: New business models for a changing environment," *Int. Federation of the Phonographic Industry (IFPI)*. London, Jan. 2009 [Online] Available: http://www.ifpi.org/content/library/dmr2009.pdf.
[2] S. Goel, P. Miesing, and U. Chandra, "The impact of illegal peer-to-peer file sharing on the media industry," *California Manage. Rev.*, vol. 52, no. 3, pp. 6–33, Spring 2010.
[3] L. Chiariglione, "The digital media manifesto-vision" [Online]. Available: http://manifesto.chiariglione.org/dmm.htm

SERVICES

[4] Korean Interactive Music Service [Online]. Available: http://www.audizen.com
[5] French Interactive Music Service [Online]. Available: http://www.iklaxmusic.com

STANDARD

[6] ISO/IEC 23000-12. (2010, July). *Information Technology—Multimedia Application Format (MPEG-A)—Part 12: Interactive Music Application Format* [Online]. Available: http://www.iso.org/iso/prods-services/ISOstore/store.html

CONFORMANCE AND REFERENCE SOFTWARE

[7] "Study of ISO/IEC 23000-12 FPDAM1 IM AF Conformance and Reference Software" [Online]. Available: http://phenix.it-sudparis.eu/mpeg/doc_end_user/current_document.php?id=29359, 93rd MPEG Meeting, Guangzhou, China, Oct. 2010, N11575.

[SP]

exploratory DSP continued from page 149

REFERENCES

[1] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.
[2] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. ICML*, 2008.
[3] T. Deselaers, S. Hasan, O. Bender, and H. Ney, "A deep learning approach to machine transliteration," in *Proc. 4th Workshop Statistical Machine Translation*, Athens, Greece, Mar. 2009, pp. 233–241.
[4] L. Deng, "Expanding the scope of signal processing," *IEEE Signal Processing Mag.*, vol. 25, no. 3, pp. 2–4, May 2008.
[5] L. Deng, M. Seltzer, D. Yu, A. Acero, A. Mohamed, and G. Hinton, "Binary coding of speech spectrograms using a deep auto-encoder," in *Proc. Interspeech*, 2010.
[6] G. Hinton, "A practical guide to training restricted Boltzmann machines," Univ. Toronto, Tech. Rep. 2010-003, Aug. 2010.
[7] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, pp. 1527–1554, 2006.
[8] A. Mohamed, D. Yu, and L. Deng, "Investigation of full-sequence training of deep belief networks for speech recognition," in *Proc. Interspeech*, Sept. 2010.
[9] A. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *Proc. NIPS Workshop Deep Learning for Speech Recognition*, 2009.
[10] V. Nair and G. Hinton, "3-D object recognition with deep belief nets," in *Proc. NIPS*, 2009.
[11] M. Ranzato, S. Chopra, Y. LeCun, and F.-J. Huang, "Energy-based models in document recognition and computer vision," in *Proc. Int. Conf. Document Analysis and Recognition (ICDAR)*, 2007.
[12] R. Salakhutdinov and G. Hinton, "Semantic hashing," in *Proc. SIGIR Workshop Information Retrieval and Applications of Graphical Models*, 2007.
[13] G. Taylor, G. E. Hinton, and S. Roweis, "Modeling human motion using binary latent variables," in *Proc. NIPS*, 2007.
[14] Y. Tang and C. Eliasmith, "Deep networks for robust visual recognition," in *Proc. ICML*, 2010.
[15] D. Yu, S. Wang, and L. Deng, "Sequential labeling using deep-structured conditional random fields," *J. Select. Topics Signal Processing* (Special Issue on Statistical Learning Methods for Speech and Language Processing), 2010.

[SP]