2003년도 한국정보통신설비학회 하계학술대회 논문집 (2003.8.29-30)

Digital Audio Adaptation in MPEG-21 Framework

Hyoung Joong Kim (Kangwon National University, Chunchon, 200-701, Korea)
Rin Chul Kim (University of Seoul, Seoul, 130-734, Korea)
Hae Kwang Kim (Sejong University, Seoul, 143-747, Korea)
Jeho Nam, Jinwoo Hong (ETRI, Daejeon, 305-350, Korea)

I. Introduction

Past media adaptation work has started from *Universal Multimedia Access* (*UMA*), which is an antecedent of the MPEG-21 digital item adaptation. Universal access technology allows users to access multimedia content anywhere and anytime without concern for specific encoding schemes, terminal capabilities, network conditions, or user preferences. To manage the problem of universal access, two different approaches have been commonly used: store variations of the same content and provide the most appropriate one, or store a single content and adapt it on-the-fly. Depending on the server capabilities (computing power, storage capacity, and number of variations of the same content it keeps) each of two approaches or both of them has been implemented. However, in this paper, universal access means to create different presentations of the same content.

MPEG-7 has laid a cornerstone for the UMA. MPEG-7 description schemes for annotating multimedia content facilitate universal access and personalized services. They shall support transcoding, translation, summarization and adaptation of multimedia according to the terminal capabilities or user and author preferences. Most of the past media adaptation approaches have focused on image, video, or Internet, speech, but audio adaptation is very few in number. Most of the previous have adapted content in terms of transcoding for better coding efficiency. However, adaptation for user preferences is quite primal. This papers implements optional adaptation for user preferences. This paper deals with audio descriptions only and has implemented them. Some of the implementations are so new from the conceptual perspective that it is difficult to find competing solutions. Moreover, those issues require further investigation for better solutions and improved performance. Audio compression and pitch scaling based on the audiogram, and audio effects for preset equalization, audio compression and equalization against ambient noise are those new topics to be studied further.

II. MEPG-21 and Digital Item Adaptation

The vision for MPEG-21 is to define a multimedia framework to enable transparent and augmented use of multimedia resources across a wide range of networks and terminals. The multimedia framework is required to support new type of multimedia usage never experienced before. It is believed that MPEG-21 will enable electronic creation, delivery, and trade of digital multimedia seamlessly. MPEG-21 identifies In MPEG-21 the most important keyword is digital item. A *Digital Item* (*DI*) is a structured digital object with a standard representation, identification and metadata within the MPEG-21 framework. This entity is also the fundamental unit of distribution and transaction in this framework. The digital item consists of multimedia resources and/or metadata.

One of the seven major key elements is "terminals and networks." The goal of the terminals and networks key element is to achieve interoperable transparent access to advanced multimedia content by shielding users from network and terminal installation, management and implementation issues. This will enable the provision of network and terminal resources on demand to form user communities where multimedia content can be created and shared, always with the agreed and/or contracted quality, reliability and flexibility, allowing the multimedia applications to connect diverse sets of users. From the user point of view the quality will be guaranteed. The adaptation of digital items is required to meet this goal. This concept is illustrated in Figure 1. As shown in this conceptual architecture, a digital item is subject to a resource adaptation (from R to $\underline{R}$) as well as a descriptor adaptation (from D to $\underline{D}$), which produce an adapted digital item. The adaptation engine enables the interoperable and transparent access to

content across network and terminal. Thus, adaptation allows user to access and render the multimedia content efficiently and effectively regardless of the contents, networks and terminals. Major goal of adaptation engine is to allow one source to be used by multiple terminals with different capabilities through different networks with different capabilities. DIA in MPEG-21 standardizes the usage environment description schemes of the user characteristics, terminal and network characteristics and natural environments. The adaptation engine adapts the original digital item to the usage environment description sent from the terminal and transmits the adapted digital item back to the terminal. Such adaptation is an essential element for interactive applications to maximize user satisfaction and to optimize network and terminal capabilities.

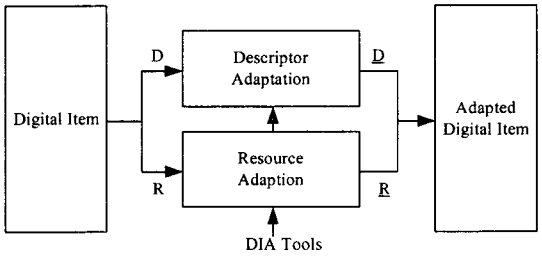


Figure 1. DIA System Architecture

The digital item adaptation tools are grouped into three major categories (see Figure 2), clustered according to their functionality and use for adaptation. The first category, which is the main scope of this paper, is the *Usage Environment Description Tools*. They include user characteristics, terminal capabilities, network characteristics and natural environment characteristics. These tools provide descriptive information about these various dimensions of the usage environment, which originate from users, to accommodate, for example, the adaptation of digital items for transmission, storage and consumption. The other tools include *Digital Item Resource Adaptation Tools* and *Digital Item Declaration Adaptation Tools*.

Audio descriptors for digital item adaptation include Audio Presentation Preferences. The proposed description scheme addresses the usage environment description for audio resource adaptation. The description scheme, adopted as MPEG-21 Committee Draft, consists of auditory impairment description scheme, audio presentation preference description scheme, and noise environment description scheme.

Auditory impairment description describes the characteristics of a particular user's auditory deficiency using audiogram. The audio presentation preference description represents the audio related preferences of the user including frequency equalizer, preset equalizer, mute, volume, audible frequency and level range descriptions. The noise environment description describes the natural audio environment of a particular user with noise frequency spectrum and noise intensity level. Each description scheme is explained in the following subsections in more detail.

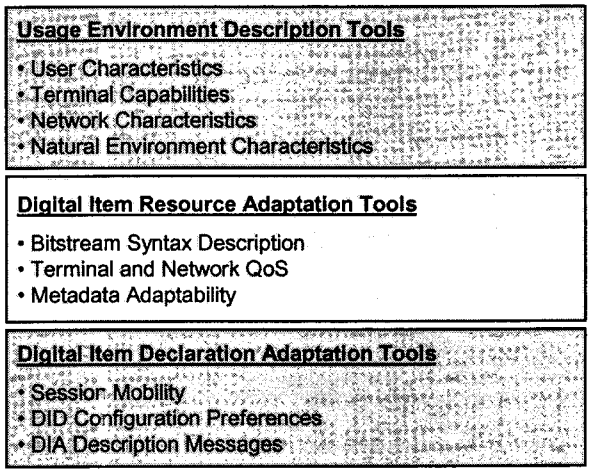


Figure 2. Overview/Organization of DIA Tools

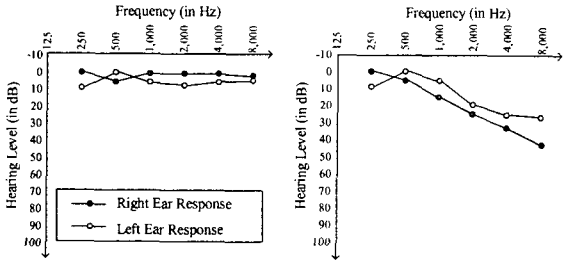
III. Auditory Impairment Descriptions

The `AuditoryImpariment` is used to describe the characteristics of a particular user's auditory deficiency. The adaptation engine uses the descriptions to deliver the best experience of audio contents for the user. Specifically, an XML document containing information to compensate for the user's auditory impairment such as hearing loss in specific frequencies will be provided.

(1) Auditory Impairment and Audiogram

Ear is a very sensitive perceptual organ and its sensitivity is usually degraded through aging process and may be greatly damaged due to long duration of exposure to strong sound pressure. An Audiogram is a chart that records the hearing response of each ear from 250 Hz to 8,000 Hz, which is the range most essential for speech perception. Thus, the descriptions in MPEG-21 are not sufficient for audio adaptation since the maximum frequency is 8 kHz and not up to 20 kHz. Hearing response is unique for each person. In the audiogram, the horizontal axis shows frequency in Hz. The vertical axis shows hearing loss in decibels (dB). Normal hearing is the 0 dB level. The degree of handicap is considered mild at 20 dB, moderate

at 40 dB, severe at 60 dB, and profound at 80 dB. Figure 3 (a) shows one example of the audiogram of a user who has normal hearing capability. Figure 3 (b) shows an audiogram having hearing loss at high frequencies. The thresholds for the audiogram are measured for both left and right ears.



(a) Audiogram of Normal Ears (b) Hearing Loss at High Frequencies

Figure 3. Audiogram of normal ears

Speech intelligibility is the most wanted application to those who have auditory impairment. We hear across a wide range of sound frequencies. Sounds from a bass drum, or a deep male voice, or the vowel sounds are low frequency sounds and register at the left side of the audiogram. Conversely, notes played on a flute, women's and children's voices and many consonant sounds such as "s," "t," "f," and "sh" are high frequency sounds and register at the right side of the audiogram. The most common type of hearing loss is sometimes called "sensorineural" or "nerve deafness." A common age-related sensorineural loss primarily affects high frequency sounds. Thus, this condition makes it difficult to understand the speech of women and children and leads to confusion of high frequency consonant sounds such as "sh," "f" and "s." The "s" sound in the word "cats" is high in pitch and fairly soft. In contrast, the "o" sound in "bow" is low in pitch and fairly loud.

It turns out that inside the ear are thousands of nerve cells that sense sounds. Some of those nerves sense soft sounds and different nerves sense loud sounds. Most people with a hearing loss are missing soft sound nerve cells; the loud sound nerves are working just fine. This means that most people with hearing loss cannot hear soft sounds but loud sounds are just as loud as they are to everyone else. Compression amplifiers adjust sound in varying amounts, depending on sound input level. As the input level of sound increases (or sound gets louder) the gain decreases. The amplifier can be set to amplify soft sounds by a greater amount than loud sounds. Note that in multichannel compression schemes, the sound or signal to be amplified is split into two or more channels, thus allowing different amounts of amplification and compression of the signal in different frequency regions. Two-channel amplification scheme separates signal into two channels: one channel for bass sounds or vowels and a separate channel for treble sounds or consonants.

A person having normal hearing capability can understand speech in a moderately noisy environment because speech is a highly redundant signal and thus even if part of the speech signal is masked by noise, other parts of the speech signal will convey sufficient information to make the speech intelligible. However, there is less redundancy in the speech signal for a person with hearing loss since part of the speech is either not audible or is severely distorted because of the hearing loss. Background noise that masks even a small portion of the remaining, impoverished speech signal will degrade intelligibility significantly because there is less redundancy available to compensate for the masking effects of the noise. As a consequence, people with hearing loss have much greater difficulty than normally hearing people in understanding speech in noise.

(2) Speech Intelligibility Enhancement

An adaptation engine may use this auditory impairment description of a particular user to adapt audio resources of a digital item destined to the user. Four possible cases according to the user's request are summarized as follows:

(Case 1) Muting specified bands: The adaptation engine does not send part of the audio signals that cannot be heard to save communication costs. If the sound on a specific frequency band is not necessary since it is not audible, the sound does not need to be sent. It can enhance compression gain (coding efficiency) or to reduce bit rates. In case of speech or alarm this approach may not be desirable since part of those bands conveys very important information. However, this approach may be effective over frequency range over 8 kHz.

(Case 2) Boosting specified bands: The adaptation engine can boost sounds at the frequency bands where the user has auditory deficiency. Equalization is a possible solution. It is usually performed with a number of band-pass filters all centered at different frequencies, and the band-pass filters have controllable gain.

This approach should consider three factors.

First, speech and other sounds should not be amplified equally. Ears with hearing loss need different amplification gains. The reason is stated above subsection. In other words, compression is needed. Second, background noise should be eliminated or reduced to enhance speech intelligibility. Third, sudden loud and transient sounds like doors slamming or street noises making striking sounds should be alleviated.

(Case 3) Pitch scaling: The adaptation engine can shift inaudible frequency up or down to the frequency the user can hear. Since many elderly people have more severe hearing loss at high frequencies than low frequencies, a popular approach is frequency lowering that lowers the frequency of received sound to the audible frequency band. Frequency lowering technology shifts the frequencies in a way that muddle sounds and does not maintain the frequency ratios within those sounds. Thus, sound distortions produced by frequency lowering algorithm limit their effectiveness.

More reliable algorithm is pitch scaling, which preserves the frequency ratios with a negligible change in the timing of speech. The pitch is equivalent of fundamental frequency, although they are not exactly synonymous. Assume that syllables uttered by a man show frequency peaks at 500, 1,500 and 2,500 Hz. Note that the frequencies are multiple of 500 Hz, which is a pitch. On the other hand, the same syllable uttered by a woman could peak at 700, 2,100 and 3,500 Hz. In this case, the pitch is 700 Hz. The pitch scaling algorithm will shift the woman's voice down about 70 percent so that it peaks at approximately the same points as the man's voice yet maintains its original frequency ratios. The speech cues are then perceived intelligibly. Of course, the adjusted female voice may sound unusual. The manipulation can also shift the male's voice to even lower frequencies.

Pitch scaling is not a cure at all. Those who have severe hearing loss at high frequency bands may still have mild or severe hearing deficiency at the low frequency bands. Thus, boosting bands at the low frequencies may be needed.

(Case 4) Changing modality: The adaptation engine replaces the audio signal with other modalities such as text resources or finger language. For example, user can ask the adaptation engine caption instead of audio or speech.

(3) Implementation

Adaptation engine implements four functionalities stated above. User can set audiogram threshold values. User having hearing loss can choose one or two options among four cases stated above.

For the case 1 (***muting bands***), user can set the minimum sound level in dB. The adaptation engine does not send audio data in these bands below that sound level to the user. For example, assume that user specifies the threshold of 50 dB at the 8 kHz band and audiogram threshold of the user is 60 dB. Then, the adaptation engine automatically does not send sounds in that band because the user's hearing loss is more severe than the specified minimum threshold by 10 dB. It is the easiest solution from the implementation point of view. Mute functionality is implemented with the descriptions.

For the case 2 (***boosting bands***), user can set the sound level in dB to boost up to that level. Default value is 0 dB. Equalizer filters in the adaptation engine then make up the bands to amplify sounds to that level. Amplification gain should be different band by band and person by person. Thus, the best way to adjusting adaptation gain is decided interactively and by trial-and-error manner. The main problem of this scheme is background noise. In general, definition of background noise itself is very ambiguous. In addition, separation of background noise is not easy task. The background noise reduction is beyond the scope of this paper, but should be considered for the better speech intelligibility. Frequency equalizer description in the following section can be used for the purpose of boosting bands.

For the case 3 (***pitch scaling***), the pitch scaling factor can be set by the user. The factor can be decided interactively and by trial-and-error manner. After the implementation of the pitch scale change, boosting bands at the remaining bands are necessary.

For the case 4 (***changing modality***), user can choose the most wanted modality. Cross-modal adaptation is a challenging process that encompasses speech-to-text conversion with synchronized manner to speech. Simplest way is to send additional text data for another modality. The modality change is not included in this implementation.

III. Concluding Remarks

Due to the limitation of pages, only auditory impairment description scheme is described. Other two schemes such as audio presentation preference description scheme and noise environment description scheme are not described. In this paper, the auditory impairment description scheme and its applications are presented. This scheme can be used for the people having hearing problems. This paper presents the adaptation engines based on this description.