# Towards Audio Enrichment through Images: A User Evaluation on Image Relevance with Spoken Content

Danish Nadeem
Human Media Interaction
University of Twente
Enschede, Netherlands
Email: d.nadeem@utwente.nl

Mariët Theune
Human Media Interaction
University of Twente
Enschede, Netherlands
Email: m.theune@utwente.nl

Roeland Ordelman
Human Media Interaction and
Netherlands Institute for Sound and Vision
Hilversum, Netherlands
Email: rordelman@beeldengeluid.nl

***Abstract*—In a visual radio scenario, where radio broadcast is consumed on mobile devices (such as phones and tablets), watching pictures as you listen, may improve information or entertainment value of the programme. We assume that audio enrichment through images can be useful to users when the selection of images is semantically associated to the spoken content. In this paper, we report about a user study to evaluate the relevance of images selected automatically based on the speech content of audio fragments (audio interviews in the Dutch language). A total of 43 participants took part in the study. They listened to a set of audio fragments and performed an image rating task. In addition to that, we conducted a small follow-up study with 3 participants to shed more light on the results of the first study. We observed that merely keyword similarity between image captions and speech fragments may not be a good predictor for image relevance from a user viewpoint, and therefore we speculate that taking topic of speech into account may improve image relevance. Furthermore, from a user perspective on the added value of audio enrichment with images, we learned that the images should strengthen the understanding of audio content rather than distracting the listeners. The insights gained in the study will open room for further investigation of audio enrichment through images and its effect on user experience.**

***Keywords–user evaluation study; linking audiovisual archives; multimedia semantics; audio augmentation.***

## I. INTRODUCTION

In recent times, as multimedia content has become increasingly easy to produce and process, technologies are being developed to automatically enrich media with links to additional information and create applications which access the multimedia content. The idea is to provide new added value services to consumers for information, education and entertainment [1]. Advances in audiovisual content enrichment techniques have generated interest in various domains like class lectures [2] and meetings [3] [4]. One of the applications - especially for audio enrichment - is the *visual radio* application, where the idea is to complement radio programmes with additional information in various modalities (e.g., text, images and videos) automatically. In this paper, we focus on enrichment of audio programmes like radio interviews, by presenting topically related images (i.e., images that are somewhat semantically associated with the speech), drawn automatically from an image collection (see an example application in Figure 1). The images presented in the figure are selected on the basis of the similarity between the keywords used to describe the image and keywords in the transcript of the audio fragment. For example, the topic of the audio fragment can be determined by a set of keywords like *baas* (Boss), *driehoeksverhouding* (love triangle) and *kwetsbaar* (vulnerable). Images containing these keywords in their description (caption, meta-data or annotation) are selected as relevant images.


Figure 1. A possible audio player interface showing images that are annotated with the terms also occurring in the spoken content. Interface source: Dutch Public Radio Broadcast platform (woord.nl).

Audio enrichment through images may deliver a richer experience for entertainment and provide additional visual information on spoken content for listeners. For the selection of images related to a speech fragment, typically a string-matching approach is taken by comparing the content of the speech transcript to the textual information of the image such as keywords, caption and meta-data, etc., used to describe the image. Generally, transcripts from speech are analyzed to extract knowledge from the speech in the form of named entities [5]. Then these named entities are used as search terms to find images from an image database collection. The relevance of the image is determined by measuring the similarity between the textual representation of the image and the search terms using a document retrieval-based approach [6]. However, such approaches developed for retrieval do not take into account the other aspects surrounding the speech such as time-frame, situations and topics. As a consequence they may not optimally represent what is said in the audio programme. Furthermore, context may play a crucial role to improve entertainment value by enriching audio content with related images, since each listener has different interests, values and preferences.

To gain some insight on the user perspective, and to

 ISBN: 978-1-61208-452-7 

understand image relevance for the presentation of images with audio, we seek answers to the following questions:

1) Do users perceive (automatically selected) images as an added value, when presented with audio content?
2) Is there any good predictor for the relevance of images for a given speech content?

We hypothesize that for the decision of presenting an image to a user, the content of the image should match with the topic of the speech content in order to add value to visualization. For example, if the speech is about "how-to" make *"Italian Tiramisu cake"*, whereas an image presented to a user shows an *"Italian Pizza"*, it will confuse the user because the topic is not fully matched. To test our hypothesis, we conducted evaluation studies, where we asked users to assess the relevance of images with the speech content of audio fragments. We consider this as a first step to develop our understanding of audio enrichment with images from a user perspective.

In the following sections, we will discuss related work on enrichment of various media (Section II), then we will describe our evaluation studies and discuss their respective results (Sections III and IV) and finally draw conclusions concerning our questions (Section V).

## II. Related work

Presenting images in audio programmes is an instantiation of what is generally called *semantic linking*, and which has gained a lot of attention recently in the audiovisual content retrieval and linking research community [7] [8] [9]. There have been investigations about hyperlinking from text sources [10] [11]. Among the latest work is semantic linking of Twitter posts with related news articles to contextualize Twitter activities [12]. Text enrichment through linking of images has found useful applications in multimodal question answering systems [13] and learning scenarios [14]. Recent research towards understanding the user perspective in image enrichment [15] and audio enrichment is also reported in our previous work [16]. Related work on video enrichment by linking additional resources through semi-automatic methods for a news show broadcast scenario is reported in the context of the LinkedTV project [17]. In the direction of audio enrichment, there have been studies on audiovisual chat conversation enrichment by linking Flickr (www.flickr.com) images to the topic of conversation [18]. Furthermore, with the emergence of visual radio applications, various techniques are deployed such as allowing users to tag contents (comments or images) on the audio timeline through an interactive audio player [19].

Towards audio enrichment, we intend to expand our general understanding of image relevance with speech content in audio programmes. Here, we focus on automatically linking speech content to related images - where a link connects an 'anchor' (information source) to a 'target' (information destination). In a speech to images linking situation, we consider an anchor as a spoken word or a phrase in an audio programme, such as the name of a person, a topic, a place, an event, location, etc., while a target can be topically related image drawn from an image database. In practice, multiple links may be created from an anchor to different target images.

## III. Image evaluation user study

We conducted an image evaluation user study where we asked participants to listen to audio fragments and provided them with a set of (automatically selected) images from a Dutch National Archive collection (www.gahetna.nl). Participants were asked to rate the suitability of the images according to the information they heard in the audio fragments. In the following section we describe the method of our study.

### A. Participants

Forty three native Dutch people who were all able to understand written and spoken Dutch participated in an online user study. Some of the participants were colleagues and common friends but most of the participants were from the general public, whom we found by visiting a public library in the town. Some of them said that they frequently listen (2-3 times a week) to radio programmes via public radio broadcast. We asked for their consent and emails to participate in the study. Later, we sent them an email with a survey link to participate in the study. Participants were between 25 and 67 years old (M = 44.7, SD = 16.2). Of the forty three participants, twenty four were female and nineteen were male.

### B. Materials

For the 43 participants who participated in the image evaluation study, we randomly selected four short-duration audio fragments (ranging from 2 to 5 minutes duration), from a collection of marathon audio interviews. The interviews in the audio were spoken in the Dutch language. The decision to use short-duration fragments was taken for practical reasons to keep study duration limited to 30 minutes. The audio interviews are publicly available at the Dutch Public Radio Broadcast platform. Furthermore, an image database collection of the Dutch National Archive, containing over 14 million historical images is used to find images relevant to Dutch audio fragments. A total of 40 images from the collection were used in the study. To draw images for a speech transcript, the images were indexed using Lucene plugin (Apache Lucene™), based on the keyword in captions and image descriptions. The search were performed using Elasticsearch® to retrieve images whose keyword meta-data match with the keyword in the transcript of an audio fragment. Because the audio fragments varied in speech content and duration (from 2 to 5 minutes), each fragment was presented with a different number of images.

### C. Rating task

All of the participants were asked to listen to each of the 4 audio fragments in the same order. We informed the participants that they could listen to the audio as many times as they liked before moving to the next audio fragment. After listening to each fragment, they had to rate on a 5-point likert scale how familiar they were with the spoken content of the audio fragment. Subsequently, for every fragment participants were presented with a varying number of images together with their captions. For fragment 1, they were presented with 14 images. For fragment 2, they were presented with 4 images. For fragment 3, they were presented with 8 images and for fragment 4, they were presented with 14 images, making a total of 40 images for all 4 audio fragments. Furthermore, the experiment interface was designed such that the participants could listen to the fragment by clicking a play button on top of the page, and then scroll down the page to see the images they were asked to rate. The images were presented statically (all at a time) according to the order in which speech was delivered

 ISBN: 978-1-61208-452-7 

TABLE I. RELEVANCE SCORES FOR ALL THE IMAGES ACCORDING TO USER RATINGS.

| Relevance score | % of user ratings |
|---|---|
| strongly disagree | 50% |
| disagree | 40% |
| neither agree nor disagree | 5% |
| agree | 2.9% |
| strongly agree | 2.1% |

in the audio fragment. For example, suppose if the speech fragment says: "Obama visits Paris and meets the president Hollande", the image of Obama is placed before the image of Hollande in the web interface.

For each image we asked the participants to provide a rating, indicating to which extent they agree that there is a match between the image and the speech fragment. The rating was distributed on a 5-point likert scale of agreement (1 = strongly disagree; 2 = disagree; 3 = neither agree nor disagree; 4 = agree; and 5 = strongly agree).

### D. Additional survey questionnaire

After the participants completed the image rating task for all of the four audio fragments, we asked them to respond to an additional survey questionnaire. The questionnaire consisted of two statements to be rated on the same 5-point likert scale for agreement, and two general questions on how frequently they listened to Internet radio and if they used mobile devices (phones, tablets, etc.,) for listening to Internet radio. The statements we asked them to rate were (translated from Dutch): (i) *The content of the image should match with the topic of the fragment.* (ii) *The images are still useful, even when the content of the images does not match with the content of the speech.* Finally, we also provided an option for general feedback to the participants about their perspective on combining audio programmes with images.

### E. Results of user responses

We analysed user rating responses from the likert-scale as a total percentage of participants' agreement with the relevance of images for the speech fragment. The result is presented in the Table I. From the result, we found that the overall image relevance scores were very low. This suggests that the selection algorithm did not perform well to select images that user would find relevant. Furthermore, to know the agreement among the participants, we computed inter-user agreements using Krippendorff's alpha [20], which was calculated using the SPSS software and a macro. We found the value of Krippendorff's alpha ($\alpha$) = 0.52, which is considered "fair agreement" level.

For the results on the additional survey questionnaire, 88% participants agreed that the content of the image should match with the topic of the fragment. Whereas, only 6% agreed that they may consider images useful even if the image content mismatches with the topic of speech fragment. Furthermore, concerning Internet radio: 23% listened several times per week, 6% listened 2-3 times per week, 18% listened once per week, 6% listened once per month, 12% listened less often and 35% never listened to Internet radio. Concerning the use of device for listening to audio, 50% responded that they used mobile devices such as phones or iPads to listen to the Internet radio. Furthermore, we received some feedback concerning the user perception on added value of images in audio programmes, as some of the participants gave general feedback in the survey questionnaire. Their responses suggest that generally the idea of combining audio and visual modalities is interesting. However, the images should strengthen the understanding of audio content rather than distracting the listeners.

### F. Analysis of user ratings

As the results obtained from the image evaluation user study showed low image relevance scores, we analysed the caption of each image and compared it with the transcript of the speech fragment where the image was presented. This gave us an indication of whether the selected images are relevant to the speech fragment.

To analyze all the images with their captions, we listed the main keywords (for, e.g., named-entities) from the caption of an image. Then we checked if the same keyword was present in the speech fragment. If it was indeed the case, and if the image appeared to be somewhat associated with the speech fragment, we considered that image to be relevant to the fragment. For example, there was a speech fragment mentioning the keywords *"Red Indian"* (native Americans). To enrich this fragment, the algorithm selected two images with captions that had the keyword *"Indian"*. However, one of the images showed a native American (caption containing the words *"Red Indian"*), whereas the other image showed *soldiers from India during World War 2*. Looking only at the captions, both of the images may be considered relevant, however, looking at the speech topic (aboutness of the speech fragment), only the first image can be considered relevant to the speech fragment.

We observed that the images matching the topic of a speech fragment in our subjective analysis, were also the images that were given higher relevance score in the user response study. To shed light on our analysis, we conducted a small follow-up study with three participants to assess the influence of image caption. We assumed that an image without a caption may or may not convey its relevance to the content of a speech fragment. Whereas images together with their caption will provide additional content information about the images, and therefore, will be rated higher on relevance when the image caption matches with the topic of a speech fragment. The study is described in the following section.

## IV. INFLUENCE OF IMAGE CAPTION STUDY

We asked the raters to provide a relevance score between an image and the speech fragment on the basis of: (i) image only and (ii) image with caption.

### A. Participants

We asked three native Dutch male participants, ranging in age from 27 to 42 years old (M = 34.4, SD = 7.5) to participate in the study. None of them were part of the image evaluation user study described in the Section III.

### B. Materials

For these 3 raters, we used the transcript of each of the 4 audio fragments from the first study and provided them with the same 40 images for an assessment task. We decided to use speech transcripts of the audio fragments so that the

 ISBN: 978-1-61208-452-7 

participants could clearly identify the spoken words, which might have been misheard in listening to the audio fragment.

### C. Tasks

Similar to the image evaluation user study, we presented all the images to all the 3 raters in the same order for each transcript of the audio fragment. However, here the task was performed in two steps; (i) all the raters were asked to read through a transcript of each of the audio fragments and then they were asked to assess the images without caption, (ii) after they had assessed the images, they were given the same images with caption, and were asked to assess them again. The instruction given was: *"From the following images, which ones do you think are suitable for the narrative"*. Furthermore, we instructed the raters to perform a binary assessment (choosing either yes or no) for the images.

### D. Results

The results of the study again show a low average relevance score for the images. Only 10.5% of the images without captions were considered relevant to the topic of the speech, when measured on a binary scale. For the image with captions, the relevance score was 17.5%. The difference between the results suggests the influence of captions in providing additional information about the images. However, because of the small sample size of the relevant images, the results remain inconclusive. Furthermore, to know the agreement among the participants, we computed the inter-rater reliability. We used Fleiss' kappa, which calculates the degree of agreement in classification over that which would be expected by chance [21]. We found the value of Fleiss' kappa ($\kappa$) = 0.64, which is considered as "substantial agreement" among the raters.

## V. Conclusions

Our main observations from the evaluation studies are that the overall relevance score for images is very low. We found that merely the presence of similar keyword(s) in the image caption and the speech content is not a good predictor of image relevance. We found that only the images whose caption and visual content appeared to be related to the topic of the speech fragment were rated higher on relevance. This suggests that taking into account the topic and other aspects of speech may improve image relevance score. Furthermore, the caption seems to help people see the relevance of an image. We consider this study as a first step towards understanding audio enrichment with images from a user viewpoint. The study has given some insights on the relevance of images to speech content. We noticed that the topic of speech plays an important role for improving image relevance score. In future work, we intend to further explore the user aspects on audio enrichment and compare user experience with different modalities.

## Acknowledgment

This research was supported by the Dutch national programme COMMIT (project P1-INFINITI).

## References

[1] L. Nixon, "Introducing linked television: a broadcast solution for integrating the web with your tv content," in ACM International Conference on Interactive Experiences for Television and Online Video, Brussels, Belgium, June 2015, pp. 1–2.

[2] S. Mukhopadhyay and B. Smith, "Passive capture and structuring of lectures," in Proceedings of the 7th International Conference on Multimedia. New York, USA: ACM, 1999, pp. 477–487.

[3] P. Chiu, J. Foote, A. Girgensohn, and J. Boreczky, "Automatically linking multimedia meeting documents by image matching," in Proceedings of the 11th Conference on Hypertext and Hypermedia. New York, USA: ACM, 2000, pp. 244–245.

[4] A. Popescu-Belis, E. Boertjes, J. Kilgour et al., "The amida automatic content linking device: Just-in-time document retrieval in meetings," in Machine Learning for Multimodal Interaction, ser. LNCS, A. Popescu-Belis and R. Stiefelhagen, Eds., no. 5237. Springer, 2008, pp. 272–283.

[5] E. Brown, S. Srinivasan, A. Coden et al., "Towards speech as a knowledge resource," in Proceedings of the 10th International Conference on Information and Knowledge Management, ser. CIKM '01. New York, USA: ACM, 2001, pp. 526–528.

[6] P. Ferragina and U. Scaiella, "Tagme: On-the-fly annotation of short text fragments (by wikipedia entities)," in Proceedings of the 19th International Conference on Information and Knowledge Management, ser. CIKM '10. New York, USA: ACM, 2010, pp. 1625–1628.

[7] R. Mihalcea and A. Csomai, "Wikify!: Linking documents to encyclopedic knowledge," in Proceedings of the 16th Conference on Conference on Information and Knowledge Management, ser. CIKM '07. New York, USA: ACM, 2007, pp. 233–242.

[8] D. Milne and I. H. Witten, "Learning to link with wikipedia," in Proceedings of the 17th Conference on Information and Knowledge Management, ser. CIKM '08, New York, USA, 2008, pp. 509–518.

[9] D. Odijk, E. Meij, and M. de Rijke, "Feeding the second screen: semantic linking based on subtitles," in OAIR, 2013, pp. 9–16.

[10] S. Chakrabarti, B. Dom, D. Gibson et al., "Automatic resource compilation by analyzing hyperlink structure and associated text," in Proceedings of the 7th WWW conference, 30 (1-7), 1998, pp. 65–74.

[11] C. Y. Wei, M. B. Evans, M. Eliot et al., "Influencing web-browsing behavior with intriguing and informative hyperlink wording." J. Information Science, vol. 31, no. 5, 2005, pp. 433–445.

[12] F. Abel, Q. Gao, G.-J. Houben, and K. Tao, "Semantic enrichment of twitter posts for user profile construction on the social web," in Proceedings of the 8th Extended Semantic Web Conference: Research and Applications, ser. ESWC'11. Springer-Verlag, 2011, pp. 375–389.

[13] W. Bosma, "Image retrieval supports multimedia authoring," in Linguistic Engineering meets Cognitive Engineering in Multimodal Systems, E. Zudilova-Seinstra and T. Adriaansen, Eds., 2005, pp. 89–94.

[14] R. Mihalcea and C. W. Leong, "Toward communicating simple sentences using pictorial representations." Machine Translation, vol. 22, no. 3, 2008, pp. 153–173.

[15] R. Aly, K. McGuinness, M. Kleppe et al., "Link anchors in images: Is there truth?" in Proceedings of the 12th Dutch Belgian Information Retrieval Workshop, Ghent, Belgium, 2012, pp. 1–4.

[16] D. Nadeem, R. Ordelman, R. Aly, and F. de Jong, "User perspectives on semantic linking in the audio domain," in 10th International Conference on Signal-Image Technology and Internet-Based Systems, SITIS 2014. IEEE Computer Society, November 2014, pp. 244–247.

[17] D. Stein, E. Apostolidis, V. Mezaris et al., "Enrichment of news show videos with multimodal semi-automatic analysis," in Networked and Electronic Media, Istanbul, Turkey, October 2012, pp. 1–6.

[18] J. Vanattenhoven, C. van Nimwegen, M. Strobbe, O. Van Laere, and B. Dhoedt, "Enriching audio-visual chat with conversation-based image retrieval and display," in Multimedia. New York, USA: ACM, 2010, pp. 1051–1054.

[19] D. Schuurman, L. D. Marez, and T. Evens, "Content for mobile television: Issues regarding a new mass medium within today's ict environment." in Mobile TV: Content and Experience, ser. HCI, A. Marcus, A. C. Roibás, and R. Sala, Eds. Springer, 2010, pp. 143–163.

[20] A. F. Hayes and K. Krippendorff, "Answering the call for a standard reliability measure for coding data," Communication Methods and Measures, vol. 1, no. 1, 2007, pp. 77–89.

[21] J. L. Fleiss, "Measuring nominal scale agreement among many rater," Psychological Bulletin, vol. 76, 1971, pp. 378–382.

 ISBN: 978-1-61208-452-7 