VIDEO TRANSLATION: WEAVING SYNTHETIC VOICES INTO THE MULTILINGUAL PRODUCTION WORKFLOW

S.A.K. Weber and X. Bai

BBC News Labs, NBH Great Portland Street, London, W1A 1AA, UK

ABSTRACT

The production of media content across several languages and platforms is both time consuming and complex. Microphones, sound booths and arrays of editing software are typically required to generate translated audio tracks. This paper presents a one-stop solution to simplifying this workflow. With a particular focus on the translation of audio tracks contained in video files, this paper describes an innovative workflow that leverages commercialised Text-To-Speech voice synthesis and a prototypical system running in production. This workflow bypasses the need for microphones, video or audio editing software and allows a single editor to generate multiple mixed-gender voice-overs. A lightweight markup language is presented which helps editors to fine-tune synthetic voices. The balance between automation and editorial and linguistic quality will be also examined. The majority positive feedback received from journalists and audiences indicates that the prototype and its underlying language technology have the potential to become part of the multilingual video production process.

INTRODUCTION

The plethora of digital platforms makes information available in a great number of languages, and the expectation of audiences to be able to consume media in their own languages is growing. International broadcasters and streaming services, in return, increasingly reach out to their global audiences in multiple languages. In global newsrooms, for example, multilingual journalists not only produce original news reports, but they also re-version existing video content into the language of their audiences. In order to meet the growing demands from our audiences, innovative production workflows and new tools must be developed to assist language editors in the translation of video content.

Current translation workflows are complex, reliant on a variety of resources and can be expensive and time consuming. This paper presents a simplified process that introduces Text-To-Speech (TTS) voice synthesis and computer-assisted translation into the re-versioning of video content. Most of us have come across either of these technologies: online machine translation, ‘digital personal assistants’ and translation apps for smartphones, in language learning tools and many others. The quality is now advanced enough to be trialled within the production process and gauged by producers and audiences. This paper begins with a brief description of the typical re-versioning workflow to identify the steps that can be rationalized. Then, with a focus on voice synthesis, this paper examines a prototypical system, developed for a pilot online service, which successfully integrated language technology. Particular attention is paid to certain linguistic aspects of voice synthesis and this paper examines how this prototype deals with differences in the quality of phonetic and prosodic voice performance. A lightweight

markup language is presented which has been designed and implemented for editors to fine-tune voices. The paper also describes how the prototype handles the generation and balancing of audio tracks in this workflow which no longer relies on any video editing software, studios or recording equipment.

The paper finally presents the user feedback received during the course of this pilot. The prototype was tested by two user groups in Japanese and Russian: language journalists (the test users) who used the prototypical voice-over tool and the audiences who watched the videos online. They provided feedback on the quality and intelligibility of the voices and on how they were perceived in voice-over tracks of news videos. The paper closes with concluding remarks.

VIDEO TRANSLATION

This paper exemplifies the new technology with the re-versioning of news video packages as were used during the pilot. These packages are illustrated news reports (usually between 2 to 3 minutes long), each of which is a composition of interviews, vox pops, edited footage and a journalist's introduction and links. The audio contained in news packages is usually a mix of the natural sound track (e.g. crowd noise, soundbites) and a pre-recorded voice-over track which narrates the story. Depending on the video content, the voice-over track can contain several different voices.

In the current workflow, shown below, the re-versioning process begins with the manual translation of a video-transcript into the target language. This is usually done with the help of a text editor (1). The translated script is needed for the subsequent voice-over recordings (2): for each voice contained in the video a voice in the target language is required, while matching the gender of the original voices. Native speakers must be available for the voice-over recording. The recording itself requires recording equipment and a sound booth or studio.

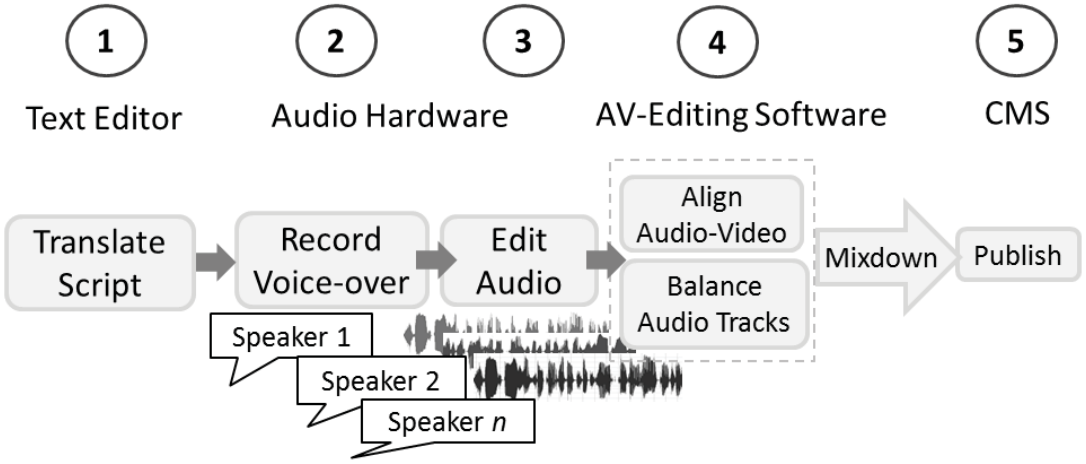

Figure 1: Typical workflow for the translation of news video packages

In step (3) the journalist edits the recorded audio clips to remove disfluencies (“er”, “um”, retakes) and adds the new audio files to the original video in a multi-track audio-video editing application. The editor then aligns the audio with the relevant video segments to ensure that the audio is in sync with the number of video frames (4). Often, the translated script results in a significantly longer text than the original English version. This requires the editor and/or native speaker to paraphrase the translated text and re-record a section

until the new audio track matches the duration of the original audio; therefore steps (1), (2) and (3) are often repeated several times. The editor then balances the audio levels of the various voices to match each other, and then mixes the new voice-tracks with the natural sound of the video. Finally (5), the video is mixed down and published using a Content Management System (CMS).

A wide range of skills, hardware and software applications are involved in the re-versioning process: editorial and linguistic skills for the text translation; dubbing skills to read and record the transcript; sound-engineering skills for audio recording, editing and balancing; text editing software for script translation; microphones, cables, sound booth; external and internal mixer for voice recordings; video editing software for audio editing and video-alignment; a CMS for video publication.

INTEGRATING LANGUAGE TECHNOLOGY

The above workflow can be simplified by reducing the number of hardware and software components that are typically involved, and by moving the whole sequence into a single application. Several steps in the traditional workflow can be merged by automating some of those tasks. To do so, during the development of the prototype presented in this paper, the focus was laid on streamlining the linguistic and audio components. Two types of language technology were identified to contribute to the simplification: Machine Translation (MT) and Text-To-Speech (TTS) voice synthesis. For the prototype, off-the-shelf products for MT and commercially available speech synthesis were used. Both these technologies were integrated into a single web application and deployed onto a cloud server. The language technology integration resulted in the amalgamation of steps (1) to (4): translation, voice-over recording, editing, video alignment and audio track mixing. In the new model, shown below, text translation, voice-over generation, video alignment and audio balancing are carried out semi-automatically and within the same window of the application. This allows each of these steps to be repeated easily, whenever and as often as needed. Video mixdown and publishing are also done within this application. Each step in the new workflow will be described in detail in the next section.

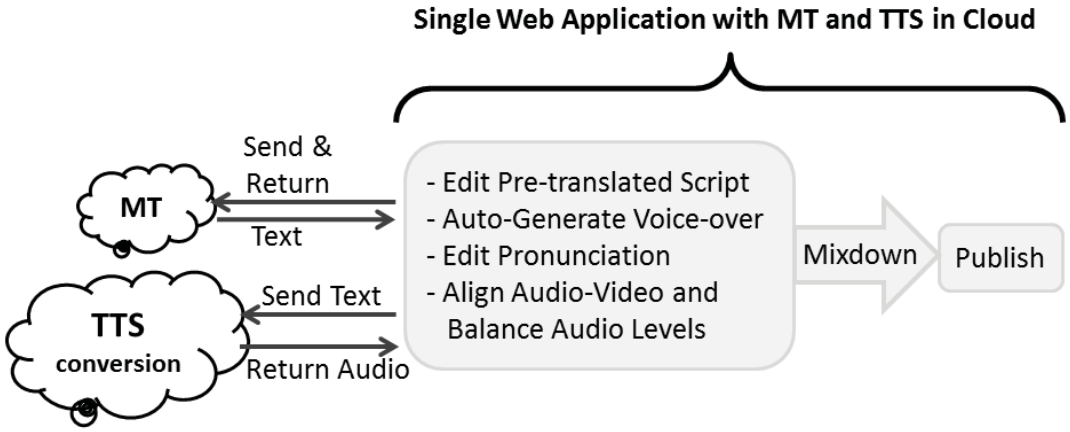

Figure 2: Integration of MT and TTS

COMPUTER-ASSISTED TRANSLATION

First, the web application sends the original English transcript to a machine translation server which immediately returns a translation. The user post-edits it to correct any

linguistic errors. MT is not used in full automation, but for computer-assisted translation, because of the high number of resulting linguistic errors. While short and factual sentences are translated fairly accurately from English into Russian, for instance, the MT engine struggles with idioms, word ambiguities and cultural context of the language. Inaccurate translation means inaccurate representation of the original text and falsifies its content. Post-editing is therefore essential in this process to preserve editorial correctness (Barrachina et al, 2008).

One by-product of using computer-assisted translation is the parallel text view which was integrated into the user interface (see Figure 3: Screenshot below). The original transcript is always visible on the screen for the user to refer back to during the whole re-versioning process. The users of the prototype found this feature to be very useful for making sure the translation doesn't deviate from the original script. [1]

TEXT-TO-SPEECH VOICE SYNTHESIS: GENERATING THE VOICE-OVER

Once the translated script has been post-edited, the user (still working in the same window) now focuses on turning sentences and paragraphs into audio with the help of TTS. The TTS synthesis technology used in this pilot is based on concatenative synthesis, also called unit selection. This technique "puts together pieces (acoustic units) of natural (…) speech to synthesize an arbitrary utterance." (Mitkov, 2009, pp.334-335). The synthesized voices that we hear are based on segmented recordings of real voices. These acoustic units (phones, diphones, half-syllables etc.) are concatenated when input-text is converted to waveforms. Because the acoustic units are based on pre-recorded human speech, unit selection speech synthesis tends to sound more 'natural' than synthesis based on other techniques, such as statistical parametric TTS. The latter is superior to unit selection in being more flexible and easier to build (Watts et al, 2013; Ekpenyong et al, 2013). However these TTS voices have a more robotic and mechanical sound quality and were therefore excluded from this pilot.

Figure 3: Screenshot: Translation, TTS, Video Preview

In the new virtual voice-over workflow, the user can choose from several mixed-gender voices in a drop-down list and assign a voice to a particular sentence or paragraph. If the video transcript requires several voices, the user can assign different voices to different

[1] For the benefit of presenting a detailed discussion of TTS technology in this paper, a comprehensive evaluation of MT will have to be the subject of a different paper.

sentences. All the user needs to do then is to press 'play' and the selected text is read aloud automatically. In the background, the text segment is sent to a cloud server where it is converted to an audio file and usually returned within up to a second. The return time depends on the length of the text. Empirically speaking, the longer a script, the longer it takes for TTS services to generate the corresponding synthesis voice. If the text consists of 5 or more sentences, the return time increases slightly. Usually, this does not exceed 2-3 seconds. The audio files are cached which allows instantaneous playback on the second play, after the whole text has been converted once. The audio file locations are cached via an in-memory hash map on the cloud server. Each TTS voice has a unique ID which can determine the voice provenance (a voice provider's API endpoint). The hash value is either a path or a URL pointing to the actual audio file returned by a TTS service.

Caching the returned audio files is essential for aligning the audio with the relevant video frames. Once they have been cached, the translated audio is played back in real time together with the video. The video play-head follows the position of the text highlighted by the user. This way the user can watch the video while listening to the synthesized audio in real time and re-edit or paraphrase the text if the audio duration doesn't match the video sequence. To align the audio with the video the user simply sets the start-time of a sentence or a paragraph. The audio editing process thus completely moves away from traditional splicing, moving, cutting and pasting of waveform blocks: audio editing is now done through script editing.

CONTROL OF PHONETIC PERFORMANCE

Often the phonetic performance of the synthetized voices requires manual correction. The TTS voices, for example, don't always stress the syllables in a word correctly which can at times be distracting or even unintelligible. The mispronunciation usually happens with proper nouns (names and places), but also applies to other word types. In unit selection voice synthesis the user has some options for controlling the TTS output, such as by re-spelling a word to generate the correct pronunciation.

Example 1) Russian: proper noun Йенс Столтенберг *(Jens Stoltenberg)*

The emphasis should be on first syllable *o* - Ст**О**лтенберг . The TTS voice, however, puts the emphasis on second syllable *e* - СтолтЕнберг. In order to trigger the correct pronunciation the user appends another *o* to the first syllable and splits the word: new spelling: Ст**ОО**лтен берг

Example 2) Russian: common noun Праймериз (*primaries)*

The emphasis should be on first syllable *a* – пр**А**ймериз . The voice wrongly emphasises second syllable *e* – праймЕриз. In order to make it say it more or less correctly the user appends a double *aa:* new spelling: пр**ААА**ймериз

Example 3) Japanese: number 8万人の移民 (*80,000 migrants)*

The correct reading is: *Hachi**mannin** no imin,* but the voice reads *Hachi **banjin** no imin.* To trigger the correct reading the user writes the same word, which is typically written in Chinese characters (kanji), in a different and syllabic script instead (kana):

new 'spelling': 8**まんにん**の移民

USING SSML PROMPTS TO CONTROL VOICE PERFORMANCE

There are other ways to control the flow of the speech in order to simulate the pacing of human speech. One of them is to use a markup language at the backend of the system. It allows users to manipulate, to a certain extent, prosody (intonation, stress and rhythm) and speed as well to insert pauses. The TTS providers employed in this prototype support SSML (Speech Synthesis Markup Language). This is a W3C standard for annotating aforementioned aspects before sending content to TTS services. However, SSML support is not standardised across different providers. Therefore, a Lightweight Expressive SSML (LESSML) was designed and implemented for our application to accommodate voice providers' different interpretations. The goal was to create a markup language with simple and unobtrusive syntax easily read and understood by editors. The basic LESSML grammar in Extended Backus-Naur Form (EBNF) is described below.

```
    Script ::= (Text Markup)* | Text
  Markup ::= Prosody | Emphasis | Silence | Substitute
  Prosody ::= "[[" Speed ( "," Volume)? "|" Text "]]" | "[[" Volume ( "," Speed)? "|" Text "]]"
Emphasis ::= "[[" "emphasis:" Number "|" Text "]]"
  Silence ::= "[[" "silence:" Number "|" Text "]]"
Substitute ::= "[[" "substitute:" Text "|" Text "]]"
    Speed ::= "speed:" Number
  Volume ::= "volume:" Number
  Number ::= [0-9]+ "." [0-9]* | "." [0-9]+
    Text ::= (!"]]".)*
```

Figure 4: LESSML grammar

Each annotation is sandwiched between a pair of double square brackets. All literal keywords (e.g., *silence*, *emphasis* and *substitute*, etc.) are case-insensitive and can be abbreviated to their first three letters, respectively (e.g., *sil*, *emp* and *sub*, etc.). Combinable keywords such as speed and volume can be applied to the same piece of text with a comma delimiter. Two examples of the use of LESSML are shown below.

スコットランドで5日に議会選挙が実施され、スコットランド国民党[[sil:0.2]]（SNP)が3期連続で勝利しました。

Example 4) silence (0.2 seconds) was inserted

Data from [[emp:strong|radar]] satellites will be used to map the ground

Example 5) strong emphasis on the word "radar" was inserted

The manual insertion of *silences* (pauses) between words is necessary to improve the overall flow and intelligibility of the TTS performance, because it allows the listener to 'digest' more easily what they're listening to. This became obvious through the user feedback: when voices just 'plough through' the whole text at a consistently slow pace

then they sound robotic and distract from the content. The manual insertion of *emphasis,* however, does not make a significant difference to the prosody of a word. It was observed that the TTS engine merely increases the audio level of the particular syllable that has been tagged with '*emphasis*'. The actual lexical stress remains unaffected.With the markup language tag '*substitute'* (see LESSML grammar above) an alternate text can be specified for the TTS voice to speak. This way, the user can save the re-spelling of recurring words for the purpose of phonetic correction and can thus construct a pronunciation dictionary. At the time of writing this paper, developing a pronunciation dictionary for this prototype is work in progress.

BALANCING AUDIO LEVELS AND MIXDOWN OF VIDEOS

In this workflow, the user does not perform any track mixing or track balancing in the traditional sense – not least because the entire editing process is now text-based. Most of the audio balancing is automated. The audio levels of all voices were calibrated and hard-coded during software development, so that the voices' audio levels match each other and don't create any digital clipping. Because the voices are computer generated their audio levels are consistent. The only inconsistent audio levels are in the original audio track. In the presented prototype the original audio is included in the final mix, with its levels lowered to serve as background track. This was in keeping with the broadcaster's audio style. Because the dynamic range tends to be different in each video, a fully automated, hard-coded level configuration for the original audio was not deemed very useful. Therefore, the user manually sets the audio levels of the underlying natural sound of the video to balance it against the TTS voices.

```
 INPUT: The original video, OV, a list of scripts objects, SO
        and a ducking configuration, DC.
OUTPUT: The mixdown video, MV.

begin
  stitched_a_track = Φ
    for each script in SO do
      if (script.silence ≠ true) then
        concat(stitched_a_track, script.a_file)
      else
        concat(stitched_a_track, script.sil_duration)
      end
    end
  ducked_a_track = duck_audio(get_a_track(OV), DC)
  original_v_track = get_v_track(OV)
  mixed_a_track = aa_mix(stitched_a_track, ducked_a_track)
  MV = av_mix(mixed_a_track, original_v_track)
end
```

Algorithm 1: Mixdown Algorithm

The video is mixed down automatically by the application. It first generates the translated audio track (see Algorithm 1 above). It concatenates all existing audio files in the correct order (a_file*)*, while including the duration of each silence in between (script.sil_duration). This chain of audio files plus silences is attached to the new audio track (stitched_a_track).

Now the original audio track is extracted and modified, with its audio levels lowered (ducked) according to the user settings. The ducked original track is then mixed with the translated audio track, with the TTS files stitched to it (aa_mix), and merged with the original video file (OV) to create a mixdown of the video (MV). The MV is now ready to move into the web-publishing pipeline.

USER FEEDBACK – LANGUAGE JOURNALISTS

The new workflow was seen by the test users as a vast improvement, because they can stay within the same web application for the entire re-versioning process from end to end. The absence of the traditional stack of editing software and hardware was considered very convenient. One of the other benefits identified by the editors was that they could create voice-over tracks with a variety of voices all by themselves through this application. If a video has only one person voicing it over, it can be confusing when, for instance, a male speaker is voiced over with a female voice, or when multiple speakers are conversing together but everyone ends up with the same voice; then it can be hard to work out which respondent is speaking at each time. Therefore, having several voices available while working alone improves the overall quality of the translated video.

The test users were impressed by the 'natural' quality of the TTS voices. Some of the voices were described as "pleasant" sounding, "soft" and "close to natural speech." In contrast, the lexical stress (word stress) was identified as the most prominent imperfection, and the manual fine-tuning of phonetic and prosodic voice performance by having to re-spell words was perceived as cumbersome. The voices were described by the users as emotionally neutral and with a lack of expression. The lack of emotional expression, however, was seen as both an advantage and disadvantage. The voices were considered inappropriate for videos with emotionally charged content, e.g. for the voicing-over of people in distress. At the same time, their neutral character was deemed very suitable for typical news segments and headlines. Nevertheless, having expressive voices available was regarded as something desirable.

USER FEEDBACK – AUDIENCE COMMENTS

The audience-facing webpages of this prototypical service displayed the videos in a carousel slider and a player featuring continuous video play (see URLs for "Today in Video", BBC Japanese and BBC Russian). They included a disclaimer to make the audience aware of the applied language technology. The audiences were also given the option to send in their comments. They were not asked any specific questions or given a questionnaire – viewers wrote their comments into a blank comment box in an online form.

Of all comments received from the viewers, 79.4 percent were positive and 20.8 percent negative about the new technology. The themes that emerged from these comments were: *TTS Quality, Access to Content versus Imperfection, and MT*. Other miscellaneous topics were not related to the new technology and are irrelevant for this review.

Breakdown of Comments Received

	Total	TTS Quality	Content v Imperfection	MT	Misc.
Positive comments	79.4%	25.4%	41.3%	4.8%	7.9%
Negative comments	20.8%	11.3%	1.6%	0.0%	7.9%

Table 1: Audience Comments

Among the positive comments, 25.4 percent explicitly addressed the TTS voice quality and described it as “really easy to understand”, “amazing”, “wonderful”. They were “surprised at the pronunciation and intonation quality”. Many of the comments said they were impressed by how natural the TTS sounded – some viewers remarked that they either forgot or “would never have guessed” that they were listening to TTS voices. 11.3 percent described the voices as unnatural sounding and criticised the wrong emphasis of syllables or pitch patterns. 42.9 percent of comments touched on the trade-off of imperfections in the TTS voices and access to news content. Among those, only 1.6 percent felt that the voices were unpleasant and difficult to listen to which stopped them from watching the content. 41.3 percent, however, felt that despite the lack of emotion in the voices and noticeable prosodic imperfections, they valued the increased offering of news content thanks to this prototype. The fact that “some words were a bit unclear” was not critical to them; although the voices sometimes “break up”, the content was “still good”. The majority supported the idea of this technology and appreciated being able to view more video content in their own language because of it and asked for this video service to continue.

CONCLUSIONS

This paper has shown how the translation of video content benefits from integrating language technology for the simplification of the translation workflow. By introducing MT and TTS into the traditionally complex sequence of video translation, several steps like text translation, voice-over generation, video alignment and audio balancing have been joined into a single web application. Furthermore, the use of TTS technology enables the semi-automatic generation of audio tracks without the reliance on studios and recording equipment. This opens up the possibility of providing language journalists with purely laptop-based video translation – creating voice-overs ‘on the go’.

This paper discussed the performance of unit selection TTS voices and how it can be controlled in the new workflow. Technically, a full automation is already possible by simply creating a text, selecting a voice and letting the system mix it down automatically. This however, renders the resulting audio potentially unintelligible, if words are wrongly pronounced and if the voice ‘ploughs through’ the text without any suitable pauses. These limitations and imperfections make the manual correction of phonetic and prosodic qualities necessary. This task could be improved to a certain extend by enhancing the user interface of the web application. The more fundamental improvement, however, is reliant on more research into TTS voice technology. Once the speech technology has improved sufficiently – perhaps through a hybrid between unit selection and statistical parametric TTS – then the task of having to manually correct the voice performance will no longer be necessary (Nuorivaara). More research in this direction is desirable.

The paper has also shown that a different set of skills is needed in this new translation workflow. The user is no longer required to 'read' waveform blocks, to cut and edit them. With the integration of TTS technology, the editing process is now driven by the script and no longer by abstract waveform blocks. Instead, the focus is shifting towards the editing of language. This becomes obvious during the correction of pronunciation errors: a thorough understanding of phonetics and orthography in the target language is essential to successfully control the TTS voice performance.

Finally, the user feedback examined in this paper has shown that even when 'unnaturalness' was observed by the audience, it didn't stop them from watching the videos. For the majority of viewers who replied, access to more news video content in their language seemed to outrank imperfection in the speech delivery as it currently occurs – as long as the TTS technology does not render the audio unintelligible. This underlines the inherent potential of this new re-versioning workflow and its underlying language technology.

REFERENCES

1. Barrachina, S., Bender, O., et al, 2008. Statistical Approaches to Computer-Assisted Translation. MIT Press Journals, Computational Linguistics. March 2009, Vol. 35, No. 1, pp. 3 to 28.
2. Mitkov, R., 2009. The Oxford Handbook of Computational Linguistics. OUP. ISBN 978-0-19-927634-9.
3. Watts, O., Stan, A., Clark, R., Mamiya, Y., Giurgiu, M., Yamagishi, J., and King, S, 2013. Unsupervised and lightly-supervised learning for rapid construction of TTS systems in multiple languages from 'found' data: evaluation and analysis. Proceedings of 8th ISCA Workshop on Speech Synthesis. August, 2013. pp. 121 to 126.
4. Ekpenyong, M., Urua, E., Watts, O., King, S. and Yamagishi, J., 2014. Statistical parametric speech synthesis for Ibibio. Speech Communication. January 2014. vol. 56, pp. 243 to 251.
5. BBC Japanese "Today in Video" (pilot webpage): http://www.bbc.com/japanese/video_and_audio/today_in_video
   BBC Russian "Today in Video" (pilot webpage): http://www.bbc.com/russian/video_and_audio/today_in_video
6. Nuorivaara, T. Breakthrough in Speech Synthesis. Aalto University School of Electrical Engineering, Department of Signal Processing and Acoustics.

ACKNOWLEDGEMENT

The authors would like to thank Toby Bladen, Andriy Kravets and Rob Squires (BBC Digital & Technology and BBC News Labs) for their work on this project, without which this paper would not have been possible.