Asynchronous Microphone Upload and Speech Recognition for Pronunciation Assessment — Supporting Language Instruction, High-Quality, Low-Bandwidth Voice, Transcription, Translation, and Speaker Verification

Conversational Applications Workshop Position

by James Salsman, jim@talknicer.com

Submitted to the World Wide Web Consortium (W3C) Ubiquitous Web Domain Workshop on Conversational Applications, 18-19 June 2010, Somerset, NJ, US.

Abstract

This position urges implementation of the HTML `input type=file accept=audio` form specification with Speex and/or Ogg Vorbis codecs, implementation and standardization of the more detailed feature extensions described at http://www.w3.org/TR/device-upload in HTML 5 updated with contemporary codecs, adding a phonetic alternative syntax for edit distance scoring to the W3C Pronunciation Lexicon Specification (PLS) Recommendation, and adding the recognized phoneme string, per-phoneme acoustic scores, and per-phoneme time endpoints as recognition results in the W3C Semantic Interpretation for Speech Recognition (SISR) Recommendation. Use cases and many of the more than 150 endorsements of the device upload proposal are described.

Introduction

For about a decade and a half, browsers have allowed the producer of an HTML form to request information, including files of data, from the person using the form. However, no HTML browser has yet implemented forms to the HTML 3.2, 4, and XHTML specifications, using `input type=file accept='audio/*'` to ask the operator to submit an audio file recorded from a microphone. If that specification were implemented, allowing the user to record, play back, and submit the recording for upload, then a wide variety of useful applications would be enabled, including pronunciation assessment for language instruction, high quality voice transcription, high-quality asynchronous speech transmission under low-bandwidth conditions, speech translation, and high-quality speaker identification and verification. Moreover, W3C speech recognition recommendations have been designed to support limited

vocabulary command-and-control applications, and not language learning, dictation, or open vocabulary applications. Dictation and open vocabularies are not addressed here.

Proposals

Along with implementing the existing HTML form specifications of `<input type=file accept='audio/*'>` for microphone audio upload using open codecs such as Speex and Ogg Vorbis, browser authors should implement and the W3C should include the 1999 device upload proposal [http://www.w3.org/TR/device-upload] in HTML 5 updated with Speex, Ogg Vorbis, Ogg Theora and/or Google VP8 codecs, to allow for a full range of secure, asynchronous, easy to use, device independent uploads from devices and/or helper applications.

To support pronunciation assessment for language instruction with speech recognition, an alternative phoneme syntax such as [phoneme1|…phonemeN] should be added to the W3C Pronunciation Lexicon Specification (PLS) Recommendation's section on the <phoneme> element [http://www.w3.org/TR/pronunciation-lexicon/#S4.6]. And finally, the phoneme string actually recognized and the acoustic score and beginning and end time points should be added, as an option, to the recognition results specified in the W3C Semantic Interpretation for Speech Recognition (SISR) Recommendation's section on “Accessing Variables Associated with a Grammar Rule or Referenced Grammar Rule” [http://www.w3.org/TR/semantic-interpretation/#SI3.3.3].

Alternatives

Related solutions currently available using Adobe Flash and early proposals for the HTML 5 device element [http://dev.w3.org/html5/html-device/] require the use of additional network bandwidth, out-of-band, proprietary, and synchronous protocols involving orders of magnitude more developer time than implementations possible if this proposal were adopted as a W3C standard and implemented in browsers. An open source vocoder (speex) was not available from Adobe Flash until late 2009 with the advent of Flash version 10 and Adobe Flex version 4. While Flash uses TCP port 1935 to send microphone information using the RTMP protocol, which has a published specification, that port is not always available through firewalls, and its port 80 HTTP-tunneled extension, RTMPT, and its secure extension RTMPS, are both still proprietary and only documented by the source code of emulations. Application developers should have the freedom from such proprietary solutions and be able to use ordinary `multipart/form-data` HTTP and HTTPS POST form submissions as an alternative to the currently available proprietary means of port 80 and secure microphone upload.

Use Case Scenarios

Please consider the following uses:

- A spoken language instruction service using pronunciation assessment, such as English Central [http://englishcentral.com], may want to obtain high-quality recorded speech from language learners and allow them to play it back for review in noisy environments to insure intelligibility and student satisfaction before submission for assessment. Doing so with existing Adobe Flash or the proposed HTML 5 `device` element would involve multiple times the amount of network bandwidth than would be necessary if this proposal were implemented.

- A voice transcription service similarly may want to obtain recorded speech from web users and allow them to play it back for review to insure accuracy and intelligibility in the presence of background noise, again without the extra bandwidth doing so with Adobe Flash or HTML 5's proposed `device` element would require.

- Any application requiring high-quality asynchronous voice transmission and operating under low-bandwidth or background noise conditions, such as voicemail or other kinds of voice messaging, would similarly benefit from asynchronous audio upload.

- Speech translation services could similarly benefit as described above, and from the ease of development this proposal allows compared to Adobe Flash or HTML 5's proposed `device` element.

- Applications requiring high-quality voice transmission for speaker identification and verification would benefit from asynchronous transmission, because recordings with higher audio quality than could be transmitted in real time over, for example, cellular telephone networks, could be enabled.

Endorsements

W3C Director Tim Berners-Lee wrote on 31 March 2000:

> This is a question of getting browser manufacturers to implement what is already in HTML.... HTML 4 does already include a way of requesting audio input. For instance, you can write:
>
> ```
> <INPUT name="audiofile1" type="file" accept="audio/*">
> ```
>
> and be prompted for various means of audio input (a recorder, a mixing desk, a file icon drag and drop receptor, etc). Here "file" does not mean

> "from a disk" but "large body of data with a MIME type."
>
> As someone who used the NeXT machine's "lip service" many years ago I see no reason why browsers should not implement both audio and video and still capture in this way. There are many occasions that voice input is valuable. We have speech recognition systems in the lab, for example, and of course this is very much needed.... So you don't need to convince me of the usefulness.
>
> However, browser writers have not implemented this!
>
> One needs to encourage this feature to be implemented, and implemented well.

In January, 2000, the device upload feature request had been endorsed by more than 150 people, including:

- Michael Swaine — Editor-in-Chief of *Dr. Dobb's Journal;*
- David Turner and Keith Ross of Institut Eurecom — in their paper, "Asynchronous Audio Conferencing on the Web";
- Integrating Speech Technology in Language Learning SIG (InSTIL) — and InSTIL's ICARE committee, both chaired by Lt. Col. Stephen LaRocca, a language instructor at the U.S. Military Academy;
- Dr. Goh Kawai — a researcher in the fields of computer aided language instruction and speech recognition, and InSTIL/ICARE founding member;
- Ruth Ross — IEEE Learning Technologies Standards Committee (LTSC);
- Phil Siviter — IEEE LTSC;
- Safia Barikzai — IEEE LTSC;
- Gene Haldeman — Computer Professionals for Social Responsibility (CPSR), Ethics Working Group Chair;
- Steve Teicher — University of Central Florida faculty; CPSR Education Working Group member;
- Dr. Melissa Holland — team leader for the U.S. Army Research Laboratory's Language Technology Group; and
- Tull Jenkins — U.S. Army Training Support Centers

Permission