Paper SAS4434-2020

Sound Insights: A Pipeline for Information Extraction from Audio Files

Dr. Biljana Belamarić Wilsey and Xiaozhuo Cheng, SAS Institute Inc.

ABSTRACT

Audio files, like other unstructured data, present special challenges for analytics but also an opportunity to discover valuable new insights. For example, technical support or call center recordings can be used for quickly prioritizing product or service improvements based on the voice of the customer. Similarly, audio portions of video recordings can be mined for common topics and widespread concerns. To uncover the value hidden in audio files, you can use a pipeline that starts with the speech-to-text capabilities of SAS® Visual Data Mining and Machine Learning and continues with analysis of unstructured text using SAS® Visual Text Analytics software. This pipeline can be illustrated with data from the Big Ideas talk series at SAS, which gives employees the opportunity to share their ideas in short, TED Talk–type presentations that are recorded on video. If you ever wondered what SAS employees are thinking about when they're not thinking of ways to make SAS products better, the answers lie in a pipeline for information extraction from audio files. You can use this versatile pipeline to discover sound insights from your own audio data.

INTRODUCTION

It is a commonly accepted idea that in the modern world of big data, the most frequent type of data is unstructured: rich media, including video and audio; free-form text, including medical notes and document collections, and so on. While some challenges, such as density and scale, are common to both structured and unstructured data, unstructured data presents additional unique challenges for analysis. One of these challenges arises because the unstructured data first needs to be converted to structured data (Bagga, 2013) and there are an almost infinite number of ways to do so, depending on the business question that is being asked of that data. But for those who are up to the task, unstructured data hides a wealth of insights.

Focusing specifically on audio data, the business value of insights from audio files is recognized by most industry analysts, who predict that the global speech and voice recognition market will reach $26B by 2025 (Marketwatch, 2019) and the speech-to-text application program interface market will reach $4.1B by 2025 (ReportLinker, 2019). This impact was foreshadowed by Donna Fluss of vendor-independent consulting firm DMG over a decade ago when she wrote: "When used properly and accompanied by best practices, speech analytics typically pays for itself in three to nine months" (Fluss, 2007). Using audio as a customer engagement channel and analytics tools to derive insights about the "voice of the customer," companies can, for example, evaluate the effectiveness of marketing campaigns, understand the experiences and sentiment of their customers, grow revenue (Salta, 2018), retain customers to increase market share, and increase efficiency by focusing on customer-reported differentiators (Kaplan, 2014). But you get no return-on-investment by just collecting data; the value comes from deriving insights and making them actionable items (Sage, 2013).

This paper describes how you can extract those insights from audio files, using five components of the speech-to-text capabilities of SAS Visual Data Mining and Machine Learning and four nodes of analysis of unstructured text using SAS Visual Text Analytics. We demonstrate using this pipeline with data from the Big Ideas talk series at SAS, which

gives employees the opportunity to share their ideas in short, TED Talk–type presentations that are recorded on video. However, the pipeline is equally applicable to contact and call center data, technical support and other conversational data, and even live streaming audio data (by connecting to SAS Event Stream Processing).

SCENARIO: BIG IDEAS

Since 2017, SAS has hosted a company-internal presentation series entitled Big Ideas. For each event, a dozen employees are selected as presenters from hundreds of applicants all over the globe. The audience consists of employees, but the event is also video-recorded, and those recordings are available on the company's internal site. The series supports the spirit of lifelong learning and provides a forum to share ideas, stories, passions, and to think about the potential to apply or achieve something spectacular. But because of the small cohort of presenters at each event, the selection process is very competitive.

Because the event is recorded on video but not transcribed, it presented an ideal use case for applying the speech-to-insights pipeline outlined previously. We focused on answering the question: "What are SAS employees passionate about outside of their daily jobs?" As extra motivation, we also wanted to use the findings for our own actionable insights: "What topics and terms should we include in order to write a winning application for the next Big Ideas event?"

TRANSFORMING AUDIO DATA INTO INSIGHTS WITH SAS

Because the source format was video, we used the open-source software FFmpeg to extract audio from the video files. The software converted the video to audio files in stereo WAV format (44.1 kHz, 16-bit stereo).

FROM AUDIO TO TEXT

The pipeline for transforming audio to text includes the following steps:

Figure 1. Speech-to-text Pipeline

For our project, the pre-processing step included these steps:

- converting the output from FFmpeg to 16 kHz 16-bit mono WAV, which is the standard supported input format for SAS speech-to-text pre-trained models
- segmentation of one long audio file into many short segments so that they could be processed in parallel and therefore faster

There are different methods for segmentation of long audio. For this project, we used a power-based algorithm to find pauses in the speech. Power is the absolute value of audio signals. Usually, an audio signal is higher in power when people speak than when there is silence. To segment long audio, we first chose a "low power threshold" that was based on the audio signal's power distribution. We considered values below this threshold to be silence. In addition, we specified a parameter called segment_len, which represented the maximum length in seconds that any audio segment could last.

The interaction between the low power threshold and segment_len for segmentation can be illustrated with the following example. We first specify segment_len of 30. In the first 30 seconds of the audio, we find the longest sub-sequence whose power values are all smaller than the low power threshold. We consider this sub-sequence as a pause in the speech, use

its center point as the breakpoint to split the audio, and, starting from this breakpoint, move to the next 30-second period. We keep segmenting this way until the end of the audio file.

These segments are used as input for the acoustic feature extraction step, in which we break the segments down into much smaller units of analysis, known as time frames. For the current project, the time frames were 25 milliseconds in length. For each extracted time frame, specific acoustic features are extracted as vectors. The acoustic model then uses these features to convert sound into characters. For the current project, we chose the 40-dimension set of mel-frequency cepstral coefficients (MFCCs) features, which is one of the standards in the field.

The third step in the pipeline is the acoustic model. In speech recognition, acoustic models map the relationship between acoustic features and linguistic units (labels) that comprise utterances. Usually, different human (or natural) languages have different label sets. You can train your own model or use a pre-trained model for this step. SAS provides pre-trained acoustic models, which are available for download from the SAS Visual Data Mining and Machine Learning webpage as two versions: one for use with central processing units (CPUs) and one for graphical processing units (GPUs). For the current project, we used a pre-trained acoustic model that had these features:

- used MFCC features for training
- was character-based and English-specific, which means that the label set consisted of the alphabet in upper case, the apostrophe character, and the space character
- relied on recurrent neural networks (RNN) with multiple Long Short-Term Memory (LSTM) layers as well as fully connected layers

Using the feature vectors extracted in the previous step as input, the acoustic model computed and produced as output the probability distribution over the labels mentioned previously for each time frame.

The fourth component of the pipeline is the decoder, which is also known as a language model. Like the acoustic model, the language model also assigns probabilities. However, where the acoustic model assigns probabilities for character labels, the language model assigns probabilities for tokens or words, based on their sequence. For this project, we used a unidirectional trigram model, which assumes that the probability of the occurrence of a word depends on the previous two words in that sequence. Like acoustic models, SAS also provides pre-trained language models for download from the SAS Visual Data Mining and Machine Learning webpage. In this project, we trained our own language model in order to augment the vocabulary for the topics that we expected to be covered. Besides the large corpus of news used to train the general model in English, our training corpus also included several SAS internal blogs, in order to introduce vocabulary related to SAS. For this project, the decoder used beam search, which is a heuristic search algorithm that expands all possible next steps and keeps a fixed number (beam) of active candidates at each time step in order to find the best guess for the word. The beam size is a parameter the users can specify themselves. Usually, increasing the beam size improves the accuracy but decreases the speed of the system. The output from the decoder is a sequence of characters comprising words that correspond to the short audio segment that was the input into the acoustic feature extraction.

Because the audio input is segmented in the first step, the output from the language model is often in units smaller than a sentence. In some use cases of text mining, this approach might work well. For this project, we were interested in the topics that speakers addressed.

In this case, we found that short snippets of transcripts overemphasized frequent words[1], which generated less-useful topics. Therefore, in the post-processing step of the pipeline, we concatenated these short sequences of words in the same order as the original audio input to create longer transcripts of text. Each observation represented one Big Ideas speech. We found that this approach gave us more intuitive groupings of topics in SAS Visual Text Analytics.

The speech-to-text pipeline described above is supported in three different SAS products: SAS® Cloud Analytic Services (CAS), SAS® Event Stream Processing (ESP), and deep learning Python (DLPy). However, there are differences between the capabilities and parameters available in these three products. Users who prefer actions can convert speech to text step-by-step using the Audio, Deep Learning, and Language Model action sets. CAS also enables users to train the acoustic and language models by themselves. For users who have streaming data events, SAS ESP enables conversion of speech to text in real time. Finally, users who prefer Python can use DLPy's end-to-end API that uses the file path of the WAV files as input and directly returns the transcript text.

FROM TEXT TO ACTIONABLE INSIGHTS

The speech transcripts, which were the output from the previous pipeline, are used as input into SAS Visual Text Analytics. To answer the research questions, we created a pipeline, which included the following nodes:

Figure 2. The Custom Pipeline in SAS Visual Text Analytics

We ran the entire pipeline and then started exploring the data by opening the Text Parsing node to find the most common terms and combinations of terms across all the speeches. The most common content terms included the terms "take," "year," "make," "work," and "people," each of which occurred in more than 92% of the speeches. We were also interested in the context in which these terms occurred. We were able to easily find, for example, that the kinds of things the speakers discussed "taking" included risks and chances, breaks and time off, lessons, and so on. Looking at the term "make," speakers talked about making choices and decisions, making things better, making a difference, making food such as casseroles and cookies, but also about making mistakes. As another example, the use of the term "year" was to provide the context of tenure at SAS, indicate age, and localize a point in time in the past when a particular event happened that the speakers shared. These contexts make sense when we think about the oratory strategy to tell personal stories to create a memorable emotional connection with the audience.

We also explored the data using the Topics node, which automatically detected common topics across the speeches and grouped them together. The pre-populated values (of 1) for the term and document density seemed to give the best results. As previously mentioned, we found that the Topics node autogenerated better topics with longer observations (such as one observation per speech) than with shorter snippets, which were broken up at pauses that speakers naturally made in intervals of 10 seconds or less.

[1] We estimated word frequency based on 11 words from the most frequent topics auto-generated by SAS Visual Text Analytics from short snippets and full speeches, using the word frequency from the 450-million-word corpus of Contemporary American English (on the website https://www.wordfrequency.info/free.asp?s=y). The frequency range of the words in the topics based on snippets was from 19 to 560 and the average was 199. The frequency range of the words in the topics based on the full speeches was from 157 to 3742 and the average was 941.

About two thirds of the speeches were grouped together automatically, out-of-the-box, into five topics, as illustrated in Figure 3. SAS Visual Text Analytics uses a term-document frequency matrix and reduces the dimensions with the singular value decomposition (SVD) method. Each topic is represented by the five most relevant terms. A plus sign in front of a term signifies that the software detected related forms, such as singular and plural of a noun or present and past tense of a verb, and automatically grouped them together. These related forms can be examined in the Text Parsing node.

Topic	Created by	Documents ↓
+word, husband, trip, +road, +die	System	8
+intelligence, +decision, learning, +machine, water	System	8
education, +student, data, +classroom, +school	System	8
innovation, trust, +culture, leadership, mindset	System	6
+dream, dream, +money, +change, +purpose	System	6

Figure 3. Topics in SAS® Visual Text Analytics

The speeches grouped in the first topic included descriptions of road trips with families and friends as well as recounting of family relationships. One of these speeches, for example, described a very stressful road trip and concluded that one person's perception of a stressful trip can be another family member's memory of a wonderful time. Another speaker recounted family trips as she talked about coping with the loss of a family member. Because our pipeline included the Sentiment node ahead of the Topics node, the results of the node included a visualization of sentiment per topic. The emotionally difficult events described in the speeches in this topic contributed to a predominately negative sentiment for the topic, as seen in the third column in Figure 4.

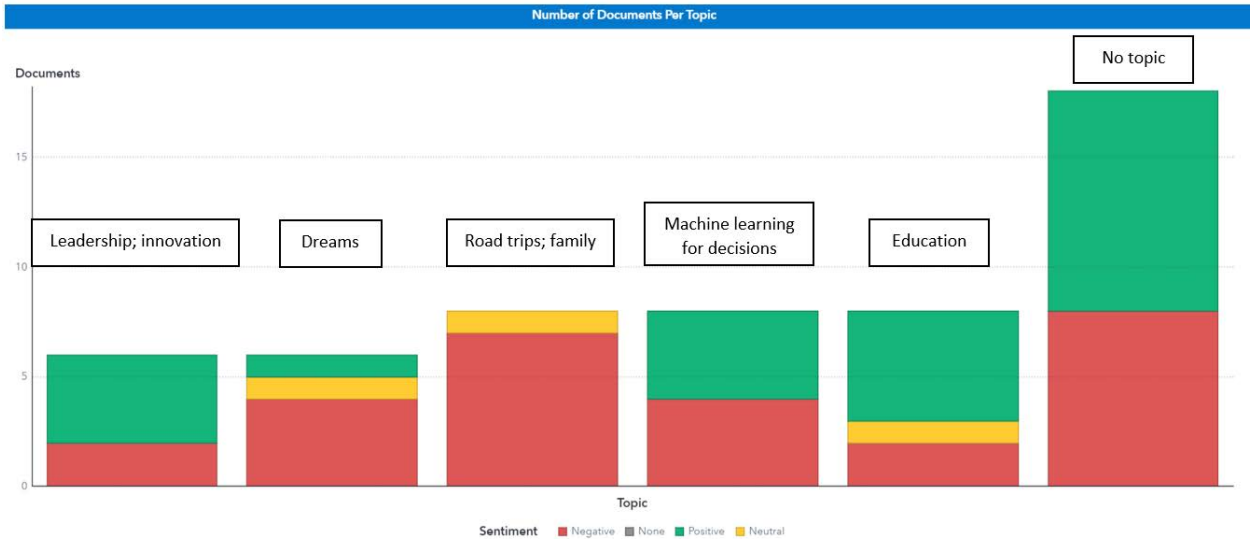

Figure 4. Sentiment of Topics in SAS Visual Text Analytics

The speeches grouped together in the second group centered on using machine learning (ML) and artificial intelligence (AI) for making better decisions. Two of them, for example, discussed how ML and AI can help solve the problems of human trafficking and shortages of clean drinking water. As seen in the fourth column of Figure 4, there were more speeches with positive sentiment on this topic than the previous topic. But there were also an equal number of speeches with negative sentiment.

The third topic centered on education, students, and data. For example, one speech discussed how we could potentially use data to better support students with disabilities and another talked about bringing data and analytics to students in classrooms. The sentiment of this topic was predominately positive, with one speech with neutral sentiment and two with negative sentiment.

The fourth topic combined ideas of leadership and innovation. Several speeches in this group focused on a culture of trust and showing vulnerability as part of successful leadership. Like the previous topic, the sentiment of most of the speeches in this topic was positive.

The fifth topic brought together speeches about dreams coming true, overcoming obstacles, persevering, and becoming successful. For example, one speech discussed the unique human abilities to dream and hope, whereas another provided tips for overcoming challenges and thriving through change. The sentiment of this group of speeches was predominately negative, probably because of the struggles inherent to achieving one's dreams.

The speeches that were not grouped in any of the topics above included unique speeches about, for example, "rocking a party" of AI geeks, creativity in work as in jazz, the power of numbers, the power of words, the value of taking chances, and others. There were more speeches with positive sentiment in this group than negative.

As we were browsing the speeches in the different topic areas, we realized that many talked about failures and struggles. Wanting to explore that area deeper, we added an additional node in the pipeline, the Concepts node (Figure 5).

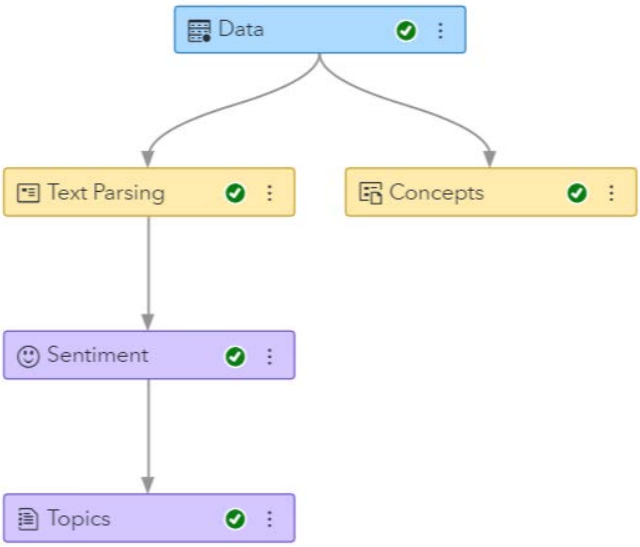

Figure 5. The final pipeline in SAS Visual Text Analytics

In the Concepts node, we created information extraction rules that captured terms, such as "problem," "difficult," "struggle," "disappoint", "hard", and "hardship." This part of the model building was a manual, iterative cycle of writing rules and examining output. We found that all but one of the speeches used these terms, with the exception being one speech about the courage to ask "what if" as the first step in encouraging innovation. We were also curious whether most of the talks also used positive terms such as "solution", "solve," "success" and "encourage." Therefore, we created a different concept with rules for extracting these terms. Less than two-thirds of the talks mentioned positive terms such as these. Therefore, the information extraction rules in the Concepts node showed that the speakers more commonly used terms describing struggling than succeeding.

CONCLUSION

In this paper, we explored how we can get insights from audio files of the Big Ideas talks at SAS. We used SAS Visual Data Mining and Machine Learning speech-to-text capabilities and SAS Visual Text Analytics to analyze unstructured data from speeches and derive insights for potentially applying to be selected for a future installment of the talk series.

Here are the data-driven insights we carry forward into writing our own abstracts for the Big Ideas series. Looking at frequently used content terms, we realized that the choice of words, including "take," "year," "make," "work," and "people," supported storytelling. Another insight the pipeline provided was that two-thirds of the talks were related to five main topic areas. Because these areas were popular in previous talks, chances are greater that proposing a future talk related to one of these topics might get accepted. In addition, nearly all talks specifically discussed a problem or a struggle, whether personal, societal, or global. Therefore, it seems very important that applicants for the talks mention a specific problem or hardship they are addressing, even more so than the solution.

This versatile speech-to-insights pipeline can be used to discover sound insights from your own audio data.

REFERENCES

Bagga, Simran. 2013. SAS Institute white paper. "Text Analytics: Unlocking the Value of Unstructured Data." https://www.sas.com/en/whitepapers/iia-text-analytics-unlocking-value-unstructured-data-108443.html.

Fluss, Donna. 2007. "Speech analytics converts call centers to profit centers." TechTarget. https://searchcustomerexperience.techtarget.com/news/1259565/Speech-analytics-converts-call-centers-to-profit-centers. Accessed on February 21, 2020.

Kaplan, Marcia. 2014. "Utilizing 'Voice of the Customer' for Competitive Advantage." Practical Ecommerce. https://www.practicalecommerce.com/Utilizing-Voice-of-the-Customer-for-Competitive-Advantage. Accessed February 21, 2020.

MarketWatch. 2019. "Speech and Voice Recognition Market Size, Growth, Opportunity, and Forecast to 2025." https://www.marketwatch.com/press-release/speech-and-voice-recognition-market-size-growth-opportunity-and-forecast-to-2025-2019-11-04. Accessed February 11, 2020.

Practical Cryptography. "Mel Frequency Cepstral Coefficient (MFCC) tutorial." http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/. Accessed February 18, 2020.

NetApp. "What is Unstructured Data?" https://www.netapp.com/us/info/what-is-unstructured-data.aspx. Accessed February 11, 2020.

ReportLinker. 2019. "Speech-to-text API Market by Component, Application, Deployment Mode, Organization Size, Industry Vertical and Region – Global Forecast to 2024." https://www.reportlinker.com/p05826804/Speech-to-text-API-Market-by-Component-Application-Deployment-Mode-Organization-Size-Industry-Vertical-And-Region-Global-Forecast-to.html. Accessed February 11, 2020.

Sage, Adele. 2013. "Avoid the "All Listen and No Action" VoC Program Trap." Forrester. https://go.forrester.com/blogs/13-04-12-avoid_the_all_listen_and_no_action_voc_program_trap/. Accessed February 21, 2020.

Salta, Marissa. 2018. "CallTrackingMetrics Named in Forrester's Overview of AI-Fueled Speech Analytics Solutions." CallTrackingMetrics.

https://www.calltrackingmetrics.com/blog/press-releases/calltrackingmetrics-ai-fueled-speech-analytics-solutions. Accessed February 21, 2020.

"SAS Visual Text Analytics." https://www.sas.com/en_us/software/visual-text-analytics.html. Accessed February 11, 2020.

"SAS Visual Data Mining and Machine Learning." https://support.sas.com/en/software/visual-data-mining-and-machine-learning-support.html. Accessed February 11, 2020.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Biljana Belamarić Wilsey
SAS Institute
biljana.belamaricwilsey@sas.com

Xiaozhuo Cheng
SAS Institute
xiaozhuo.cheng@sas.com