



AudioSoft

White Paper

Audio Mining – How to locate vital information automatically

A review of technologies that will allow you to optimise your analysis and transcription resource

Revision 1.2
September 2008

Authors:

Robert Wright (robert.wright@ultra-audiosoft.com), BSc (Hons)
Dr. Paul Rocca (paul.rocca@ultra-audiosoft.com), PhD, BSc (Hons)

Contents

1.0 Introduction	4
2.0 Voice Detection	5
3.0 Automated Speech Transcription	7
4.0 Key-word spotting	11
5.0 Textual searches of transcripts	12
6.0 Language Identification	13
7.0 Speaker Identification.....	14
8.0 Auto Gisting.....	16
9.0 Audio Analytics – Collection, Classification to Analysis	17
10.0 Recommended Solution:.....	19
11.0 Conclusions.....	17
12.0 References.....	18
13.0 Acronyms	19
14.0 Appendix	23

Executive Summary

With ever increasing communications and the decreasing costs of mass storage huge quantities of data from a variety of sources are being recorded. The result is that most organisations do not have sufficient manual resource to analyse the recorded data and may be missing vital information and intelligence. Best in class audio analytic tools allow rapid identification of important information that may be buried within thousands of hours of these recordings. Whether you are searching for a word, a phrase, a person or just trying to understand complex relationships audio analytics can greatly enhance your operational capability and save significant costs. However these technologies only produce effective results in certain scenarios.

This paper reviews when various data mining technologies such as key word spotting, text to speech, speaker identification, language identification, should be used to achieve the best results. The paper reviews the practical factors that affect the success of these analytic technologies and demonstrates that data collection, classification and analysis needs to be considered as an integrated end to end process.

The paper also discusses practical ways in which Supervisors and Managers can use this knowledge to maximise manual and automated resources.

Audience

This paper is relevant to Supervisors, Managers and Analysts who need to maximise manual and automated resources and are seeking to improve operational capabilities.

In particular it is relevant to those responsible for recording solutions and whose budgets are weighed down by the labour costs of either trawling through thousands of hours of recordings or transcribing thousands of hours of audio data.

1.0 Introduction

1.1 Understanding the recording environment

Situations such as surveillance and lawful intercept benefit significantly from modern digital audio recording. Recording solutions can record a variety of different media for tens of thousands of hours. Depending on the environment, the recordings may be continuous or may be triggered by certain conditions. For example, a surveillance recorder may record everything or may be triggered by the amplitude of the signal. In a telephony environment, call recording may be triggered by vox, the handset being uncradled (on hook / off hook detection), keys being dialled or, in VoIP, through call initiation packets. These situations are best described as uncontrolled environments; analysts have no control over the content of the recordings.

Typical questions that may be asked are:

- ▶ “Is there a security threat?”
- ▶ “What was said?”
- ▶ “Who is that?”

Existing commercial packages offer audio mining in controlled environments. For example, many cinemas now have automated telephone booking. This prompts users to speak from a very restricted vocabulary set and allows words to be re-spoken if they cannot be interpreted with confidence. Another form of commercial software that has received mainstream attention is that of the dictation package used as an alternative to typing documents on the home PC. In this case, significant amounts of training data are used to teach the software about the user’s voice; the training process needs to be repeated for each new user.

In uncontrolled environments, these methods for data mining are not effective. Sophisticated analytical tools that have been designed for uncontrolled environments are required and these need to be integrated with the recording process to be effective. Without proper integration, false negatives can lead to incorrect decisions and false positives can lead to wasted time and effort. However, with proper integration and the correct tools used in the appropriate situations, analytical tools can enhance operation capability whilst significantly reducing manpower in the following markets;

- ▶ Courtrooms
- ▶ Air traffic control communications
- ▶ Lawful Intercept by Law Enforcement Agencies
- ▶ Police and Immigration interviews
- ▶ Emergency Services command and control positions
- ▶ Surveillance by Law Enforcement Agencies

This paper is based upon AudioSoft’s experience of investigating all the different available audio analytics tools on the market and applying the best of breed tools to real data from customers in the above market places.

1.2 How we can analyse recordings to elicit information about conversations

As well as parameters of the recording (such as start time, end time, duration, channel, field information) any conversation has the following characteristics:

- ▶ Conversation parameters: Start time, end time, duration, speaker 1, speaker 2 (optional, and also possibly speakers 3, 4, etc)
- ▶ Diction: Words and sounds that are spoken by each speaker (though they may overlap), including proper nouns and dialect(s)
- ▶ Speech characteristics: Language(s), accent(s), emotions
- ▶ Background noise

Only the recording parameters and conversation parameters are assumed to be fixed and absolute. Words, accents, languages are all subject to change and interpretation (the speaker may not even know himself what words with what accent in what language has been spoken). Each speaker has their own characteristics, which may have a causal relationship with the speech characteristics. For example, the place of birth of the speaker and their education will have a significant causal relationship on the language, dialect, accent and vocabulary used.

In most situations an analyst, whether recording covertly or overtly, will know some but not all of the above information and will seek to enhance their knowledge of the above information in order to answer the questions that they face. With all but the recording parameters and conversation parameters, perfect knowledge is not possible so we seek to assign the best information we can to these characteristics.

By assigning information to the characteristics of conversations/recordings and using an effective search one can find relevant conversations, interpret the information presented and then make a decision. The next sections describe some of the tools that can be used to help with this.

2.0 Voice Detection

One of the core tasks in all speech processing is identifying the presence of speech or voice in a signal. In most cases it is important to identify the starting and ending points of voice to allow further analysis. The voice component of a signal is a slowly time varying waveform that can be considered stationary in a short period of time (usually 5 to 100 ms). When an audio signal is broken up into frames (typically 10ms), this slowly varying property of voice signals, enables the classification of frames as noise or voice based on whether they can be modelled as short-term stationary processes.

In environments where there is reasonably good audio quality with little background noise or crosstalk, voice detection is relatively straightforward using standard algorithms that have been widely published in the literature, e.g. [1], [2]. There is a huge number of academic papers published comparing different algorithms, the simpler ones take account of the zero-crossings and the amplitude of the temporal signal, to detect the lower frequencies and higher amplitudes of voice segments when compared to noisy frames.

In noisy environments where the Signal to Noise Ratio is lower, research is ongoing, e.g. [3]. More sophisticated algorithms are based on spectral characteristics such as the energy contained within certain frequency bands within a certain time frame.

These algorithms can perform better than standard algorithms in environments where audio quality is poor, levels are either too quiet or clipped and where there can be large levels of background noise.

Benefits & Features
<ul style="list-style-type: none">▶ Detect speech in recordings that are primarily noise; reduces manpower costs and avoids missing key sections of speech▶ Mature technology allows relatively simple implementation
When can this tool be used?
<ul style="list-style-type: none">▶ When speech is interspersed with large gaps, where users currently manually search for data i.e. surveillance and lawful intercept solutions▶ When searching for data in recordings from noisy environments▶ Used as a pre-processor for automated speech to text tools, thereby reducing the amount of data that needs to be processed.

3.0 Automated Speech Transcription

Audio and video recordings from situations such as courtrooms, interviews, lawful intercept and surveillance now often need to be recorded because of legislation and transcripts produced to give an accurate record of proceedings. Currently the recordings are manually transcribed by human operators listening to the audio and typing in the speech as text. The goal of automated speech transcription is to automatically turn an audio recording of speech into an accurate transcription of that speech to reduce the time taken to perform this transcription and the manpower costs involved. To be effective in the uncontrolled environments under review in this White Paper that the solution must be speaker independent, require no training and produce minimum transcription accuracies of 50%.

The three main types of automated transcription are:

3.1 Grammar constrained recognition

Grammar constrained recognition works by constraining the possible recognized phrases to a small or medium-sized formal grammar of possible responses, which is typically defined using a grammar specification language. This type of recognition works best when the speaker is providing short responses to specific questions, like yes-no questions; picking an option from a menu or selecting an item from a well-defined list. The grammar specifies the most likely words and phrases a person will say in response to a prompt and then maps those words and phrases to a token, or a semantic concept. For example, a yes-no grammar might map "yes", "yeah", "sure", and "okay" to the token "yes" and "no", "nope", and "nuh-uh" to the token "no". If the speaker says something that doesn't match an entry in the grammar, recognition will fail and typically in commercial applications, the respondent will need to repeat themselves.

3.2. Natural language recognition

Natural language recognition is another type of system found in commercial use and much research is being carried out in this field. Natural language recognition uses statistical models. The general procedure is to create as large a corpus as possible of typical responses, with each response matched up to a token or concept. In most approaches, a technique called Wizard of Oz is used. A person (the wizard) listens in real time or via recordings to a very large number of speakers responding naturally to a prompt. The wizard then selects the concept that represents what the user meant. A software program then analyzes the corpus of spoken utterances and their corresponding semantics and it creates a statistical model which can be used to map similar sentences to the appropriate concepts for future speakers. The obvious advantage for call centres of natural language recognition over the grammar constrained approach is that it is unnecessary to identify the exact words and phrases. A big disadvantage, though, is that for it to work well, the corpus must typically be very large. Creating a large corpus is time consuming and expensive.

3.3. Dictation

In the Dictation approach, software is used to transcribe someone's speech, word for word. Unlike grammar constrained and natural language recognition, dictation does not require semantic understanding. The goal isn't to understand what the speaker meant by their speech, just to identify the exact words. However, contextual understanding of what is being said can greatly improve the transcription. Many words, at least in English, sound alike. If the dictation system doesn't understand the context for one of these words, it will not be able to confidently identify the correct spelling and will thus either put in an incorrect word or an absence of a word. Market leading software can use grammatical rules to clarify, which of two phonetically identical words is meant.

The models we consider in this white paper are dictation based since this is the most appropriate method of the three for the applications being considered. The models assume no prior knowledge of speaker identification or any training data, i.e. they do not require a speaker to have already spoken any words to train the software. They do, however, require various conditions to be fulfilled:

- ▶ The speech needs to be located in the audio file. Speech recognition is described in Section 2.
- ▶ The language of the speech needs to be known or the language needs to be identified as per Section 4.
- ▶ The signal-to-noise ratio needs to be sufficiently high for a meaningful transcription to be achieved. This is now discussed further.

If a human operator struggles to make out the speech then an automated transcription algorithm will fare little better. The metric we shall use for transcription accuracy is the Crampin Index, a modified version of the NIST criteria, as described in Section 14.0 Appendix.

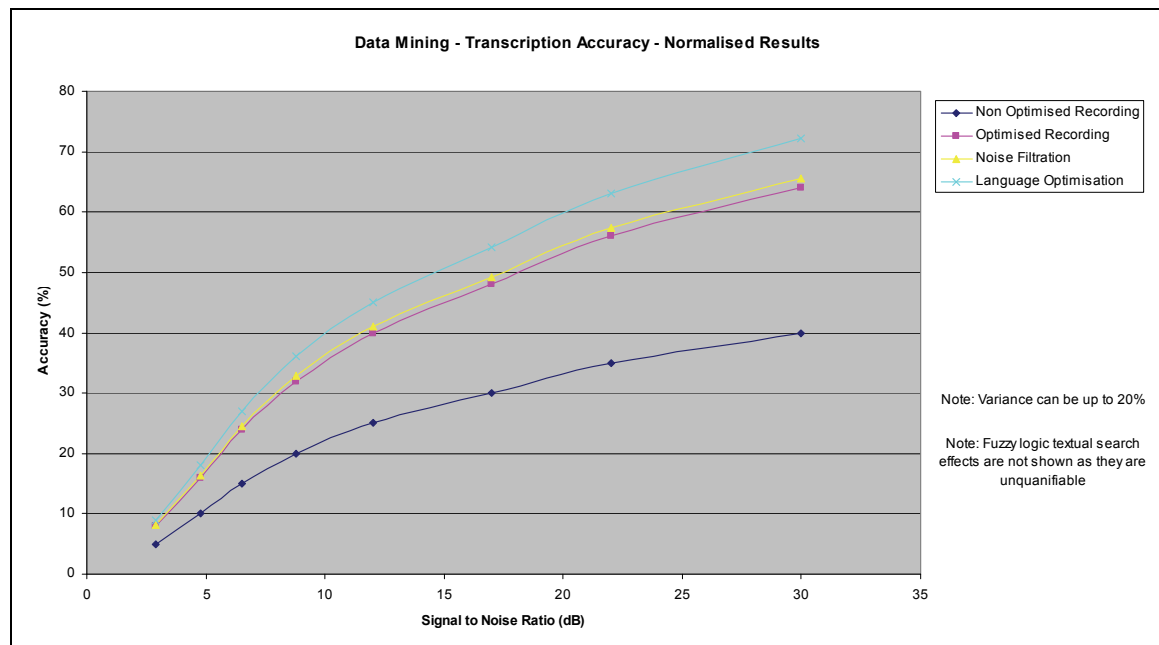


Figure 1: Transcription Accuracy as a function of Signal-to-Noise ratio

It can be seen from Figure 1 that, as a general rule, transcription accuracy increases linearly with signal-to-noise ratio. This is to a maximum of 80-90% accuracy depending on the models and vocabulary set used.

3.4 Issues that complicate transcription:

3.4.1 Poor audio quality

Even when recorded files are of broadcast quality, the actual audio content can be poor, due to signal noise, ambient noise interference, microphone location/quality, reverberation and echo.

3.4.2 Strong regional accents

Some pieces of audio are unintelligible to humans as well as computers.

3.4.3 Very fast speech, drunken/drugged speech

Most transcription systems have issues when speech is so rapid or slurred that boundaries between words become indistinct (segmentation problems)

3.4.4 Non standard patterns of speech

If the speaker uses words that are out of the transcription system's vocabulary it cannot recognize them. If the words are in the vocabulary but used in a non-standard context it may either fail to recognize them or it may simply slow the transcription process down depending on the situation.

3.4.5 "Cocktail party effect"

This is the name given to the effect that even humans have difficulty listening to recordings made using one microphone with more than one person speaking at once, a typical scenario of which is a social gathering such as a party, though it is also a common scenario in other settings such as recordings of meetings where people tend to interrupt and talk over each other, and where there may also be other typically quieter sub-conversations taking place at the same time.

Noise filtration can make small improvements to issues 3.4.1 and 3.4.5 as shown in Figure 1. However although the use of filters can improve the human intelligibility in some circumstances artificial artefacts can be introduced resulting in a reduction in transcription accuracy.

Language optimisation can make significant improvements to issues detailed in 3.4.2, 3.4.3 and 3.4.4 as shown in Figure 1. By varying the vocabulary set to suit the situation, one can improve transcription accuracy. For example, in air traffic control situations, the phonetic alphabet (alpha, bravo, charlie) is used commonly and so the expected probability of these words is increased. Appropriate language sets can be implemented depending on both language and dialect. Automated language detection is discussed more in Section 4.

However language optimisation is a labour intensive operation and typically requires 10 hours of labour per audio hour.

3.4.6 Speakers talking over each other

In order to accurately analyse a conversation it is important to be able to distinguish each speaker from the other(s). In the simplest case of a monologue no segmentation is required. In more complex cases, segmentation can be achieved by

taking into account voice characteristics (such as speed, pitch, accent, vocabulary, dialect); this can generally be achieved through automated techniques though accuracy is not necessarily perfect – two people talking at the same time can cause problems with automated (and manual) transcription techniques.

Benefits & Features

- ▶ In environments with reasonably clear speech, turns speech into text with an accuracy that facilitates easier transcription.
- ▶ Can significantly reduce the manpower taken to transcribe recordings.
- ▶ One can typically achieve approximately real time (1:1) transcription, depending on the model chosen and the quality of the audio file

When can this tool be used?

- ▶ When speech from a relatively noise-free environment needs to be transcribed to significantly reduce the costs of manual transcription, e.g. Lawful Intercept, Surveillance, Interview situations, courtrooms.
- ▶ When a summary of subject content within a recording (gisting) is needed.
- ▶ When a piece of intelligence may require thousands of hours of recorded data to be analysed.

4.0 Key-word spotting

Key-word spotting can be viewed as an addition or alternative to automated transcription described in Section 3. The aim is to prioritise a large number of recordings for review by a human operator or based on the occurrence of one or more key words or phrases, thus improving:

- ▶ The speed with which high-importance recordings are accessed
- ▶ The operational capability by increasing the amount of recordings that can be accessed
- ▶ Identifying which recordings need further analysis (either manual or automated).
- ▶ As an alternative to when transcription is either not necessary or not possible (at least automatically) due to the issues described in Section 3.

As an example, consider a surveillance or lawful intercept recording situation where the signal-to-noise ratio is relatively low, since the speakers will not be speaking directly into microphones and one needs to find the few important sentences of speech in thousands of hours of speech, noise and silence.

The two main approaches to Key Word Spotting are:

1. Large Vocabulary Continuous Speech Recognition. A large vocabulary Continuous Speech Recognition (LVCSR) engine is used to output text transcribed from speech. A textual search is then performed to search for the key words.

2. Phoneme Recognition based key word spotting. Speech is first broken down into phonemes, the basic elements of speech (such that several phonemes make a word). The software then matches sequence of phonemes that are phonetically identical to the key words. A similar approach can perform the key word spotting in one stage by searching through the transcripts for the key words or phonetic equivalents to them.

For the requirements under review in this paper, option 2 offers the best solution.

Benefits & Features
<ul style="list-style-type: none">▶ Highlights recordings that should be prioritised for further analysis▶ Effective in even when the signal to noise ratio is poor▶ Phonetically identical words can also be identified (e.g. shore/sure)▶ Can process up to 10x real time speed.
When can this tool be used?
<ul style="list-style-type: none">▶ As an aid to human operators by prioritising the order in which recordings are considered by the human operator▶ When searching for a key word or phrase only.▶ As a pre-processor before deciding whether recorded material warrants further analysis

5.0 Textual searches of transcripts

To make effective use of transcripts created by any process, whether transcribed by humans or by machine, there needs to be efficient ways of searching and categorising the textual information contained within a transcript.

However pure text string searching is unlikely to be of great use due to the flexibility of the way spoken language is used; instead a textual logic toolkit can:

- ▶ Find words with the same root as the search term e.g. searching on "drive" would find "drove", "driven" and "driving" too.
- ▶ Find synonyms of the search term e.g. "dog" would also find "hound", "mutt", "spaniel" and "lassie"
- ▶ Allow for searches which contain words (and stems and synonyms) in association with each other, e.g. a search for "drive NEAR airport" should find "drive to the airport" and "drove back from Heathrow" but not drive on one page and airport on the next.
- ▶ Undertake phonetic searching, allow for the spelling independent searching of transcripts for particular patterns of phonemes, e.g. searching for "recognise speech" should match "wreck a nice beach"

However textual search of transcripts does require manual effort in order to input all the data.

Benefits & Features
<ul style="list-style-type: none">▶ Database tools can be used to further enhance transcription or key word spotting capabilities.▶ A combination of language capabilities are available to find similar words easily
When can this tool be used?
<ul style="list-style-type: none">▶ To enhance automatic transcription and key word spotting, so that one can easily search for a subject of interest without having to specify a multitude of different variants of the same word

6.0 Language Identification

The goal of language identification is to automatically segregate audio data into identified languages. This would enable appropriate processing of audio data, whether this is to specify which model an Automatic Speech Recognition system should use to transcribe a file, or to flag the file for transcription by an appropriately skilled human transcriber, or to indicate that the file is of a profile that warrants further evaluation and should be prioritised above others for action.

A combination of language identification, vocabulary models, automated transcription and automated translation can be used in complex scenarios but careful attention should be paid to the loss of accuracy at each stage of the processing. For specific requirements consult your audio analytics provider.

Benefits & Features

- ▶ Language identification can quickly inform the analyst of the languages being spoken in different recordings (or in the same recording if the language changes)
- ▶ Language identification allows the appropriate language model to be used for transcription or key word spotting or for the appropriate analyst to review the document
- ▶ Can alert analysts when the language being spoken in a conversation changes

When can this tool be used?

- ▶ Language identification can be used in situations where the language is variable to infer information about the recording
- ▶ Language identification makes the transcription process easier in environments such as immigration or police interviews
- ▶ In a lawful intercept environment, language identification can provide valuable information about the content of recordings

7.0 Speaker Identification

The aim of Speaker Identification is to output the identity of the person most likely to have spoken that speech from a known population. Speaker identification is primarily used within audio analytics to differentiate multiple speakers when a conversation is taking place and to identify an individual's voice based upon previously supplied data regarding that individual's voice.

Speaker identification may be aided by prior knowledge (e.g. an analyst recognising a voice), through recorder information (e.g. CLI or DDI information being logged with a call) through someone identifying themselves by saying their name or through comparing the voice print against a known database of speakers (this is often also called voice biometrics). This last method is often automated, for which the voice print (a unique biometric of a person's mouth, nose and larynx) is enrolled through 30 seconds or more of audio data. These different methods may be used independently or combined depending on the application. In each case we are trying to ascertain information about the speaker, if not their identity then their gender and age. Speaker identification should be independent of medium (e.g. landline, microphone, mobile, VoIP), language and diction in the applications we consider.

The performance of Speaker Identification naturally decreases as the population size increases.

Test and reference patterns (i.e., acoustic features associated with a person's speech) are extracted from speech utterances statistically or dynamically. Various statistical acoustic features may be used in a matching stage. The comparison in the matching stage is performed using fairly complex statistical maths, typically probability density estimation or by distance (dissimilarity) measure. After comparison, the test pattern is labelled to a speaker model at the decision stage.

Two modes for Speaker Identification are the text-dependent and the text-independent. In the text-dependent mode, utterances of the same text are used for training and testing. Time alignment is needed for this mode. In the text-independent mode, training and testing involve utterances from different texts. Much more speech utterance is needed for this mode to increase accuracy. Statistical features are better for the text-independent case. Unless one can guarantee that recordings would always feature a particular piece of text (e.g. a radio call sign, commonly-used slang, a reference to a deity or similar frequently quoted piece of text), one is reliant on text independent mode though this is less accurate. For the purposes of the uncontrolled environments under review in this White Paper, text-independent speaker identification will be considered.

Benefits & Features

- ▶ Speaker identification distinguishes different speakers to allow partitioning of conversations, making interpretation of conversations easier
- ▶ Text-independent speaker identification can alert analysts when a person of interest (from an existing database) is speaking

When can this tool be used?

- ▶ Speaker identification can be used as an aid in any recordings where multiple speakers are being recorded on the same channel
- ▶ In lawful intercept/surveillance situations, a database of speakers can be used to match with the speakers from the recording, although the accuracy of such techniques depends on the size of the database and the amount of speech available

8.0 Auto Gisting

The aim of auto gisting is to output is to extract key sentences (highlights), show the key phrases in context and provide a summary of the document.

The text from a transcribed recording (See Section 3) is fed to the auto gisting software. A sequence of words with their attributes is extracted from the text according to linguistic rules and specified dictionaries. Semantic analysis is performed with the help of neural network technology. The software creates a knowledge base that contains all the information about the text structure, which serves as a basis for carrying out retrieval and analysis functions requested by the user.

Auto gisting products can deliver functionality such as:

- ▶ A list of the most commonly used words and phrases in their original context
- ▶ Word stemming and transforming to a generic form
- ▶ A frequency-based dictionary of concepts from the text (including words and word combinations)
- ▶ A filtration of words from a given stop list
- ▶ Lists of words semantically related to the selected words in the investigated text
- ▶ Numerical evaluation of the semantic significance of different concepts, and the weights of relations between these words, in the investigated text

In more complex situations the gist can be found through a careful combination of “techniques in speech recognition, speaker identification, natural language analysis and topic statistical classification” [5]. In these situations, complex language models, including phonetic and grammatical understanding, are required to facilitate high accuracy.

Benefits & Features
<ul style="list-style-type: none">▶ Auto gisting can summarise recordings by pulling out several key sentences and words, allowing quick comprehension of the content▶ The most commonly used words and phrases are highlighted and their context is given, thus allowing prioritisation of calls quickly and efficiently▶ Traffic light-based system can be implemented to prioritise recordings for further human analysis based on key criteria
When can this tool be used?
<ul style="list-style-type: none">▶ Auto-gisting can be used to prioritise recordings where the content is previously unknown based on the gist (or general meaning). For example, in telephone conversations.▶ Auto gisting is a maturing technology that is likely to provide future benefits in conjunction with other audio analytics.

9.0 Audio Analytics – Collection, Classification to Analysis

The effectiveness of the audio analytic technologies described in this document are hugely dependent upon the characteristics of the recorded data. These characteristics in turn affect which technologies will be applicable and the results they can produce.

Therefore it is vital that the characteristics of the recordings are classified, either in real time or part of a post processing operation.

In a real time scenario classification process can be used to improve the quality of the recordings by, for example, adjusting the signal gain to reduce clipping or increase the amplitude.

These characteristics are then saved as meta-data and can include:

- ▶ An estimate of a signal to noise ratio of the audio data
- ▶ An estimate of a length of speech content in the audio data
- ▶ Detection of periods of silence in the data

The meta-data can be displayed to a user or used to determine the next step of the analysis or used as part of an automated workflow process to route the data to;

- ▶ Automated transcription
- ▶ Key Word spotting
- ▶ Language identification
- ▶ Contextual search
- ▶ Human analysis

By adopting such an approach automated audio analytic technologies will only be applied to data when they will produce effective and meaningful results.

This approach therefore relies upon a flexible server farm resource. For instance this week's operation may lend itself to 5 servers automatically processing speech to text, with another 3 undertaking key word spotting and another 2 undertaking speaker identification. The following week may require a completely different approach if for instance the signal to noise ratio drops below a threshold resulting in 8 servers being used for key word spotting and the other 2 for speaker identification.

A flow-chart for the process from collection through classification to analysis is included in Figure 2.

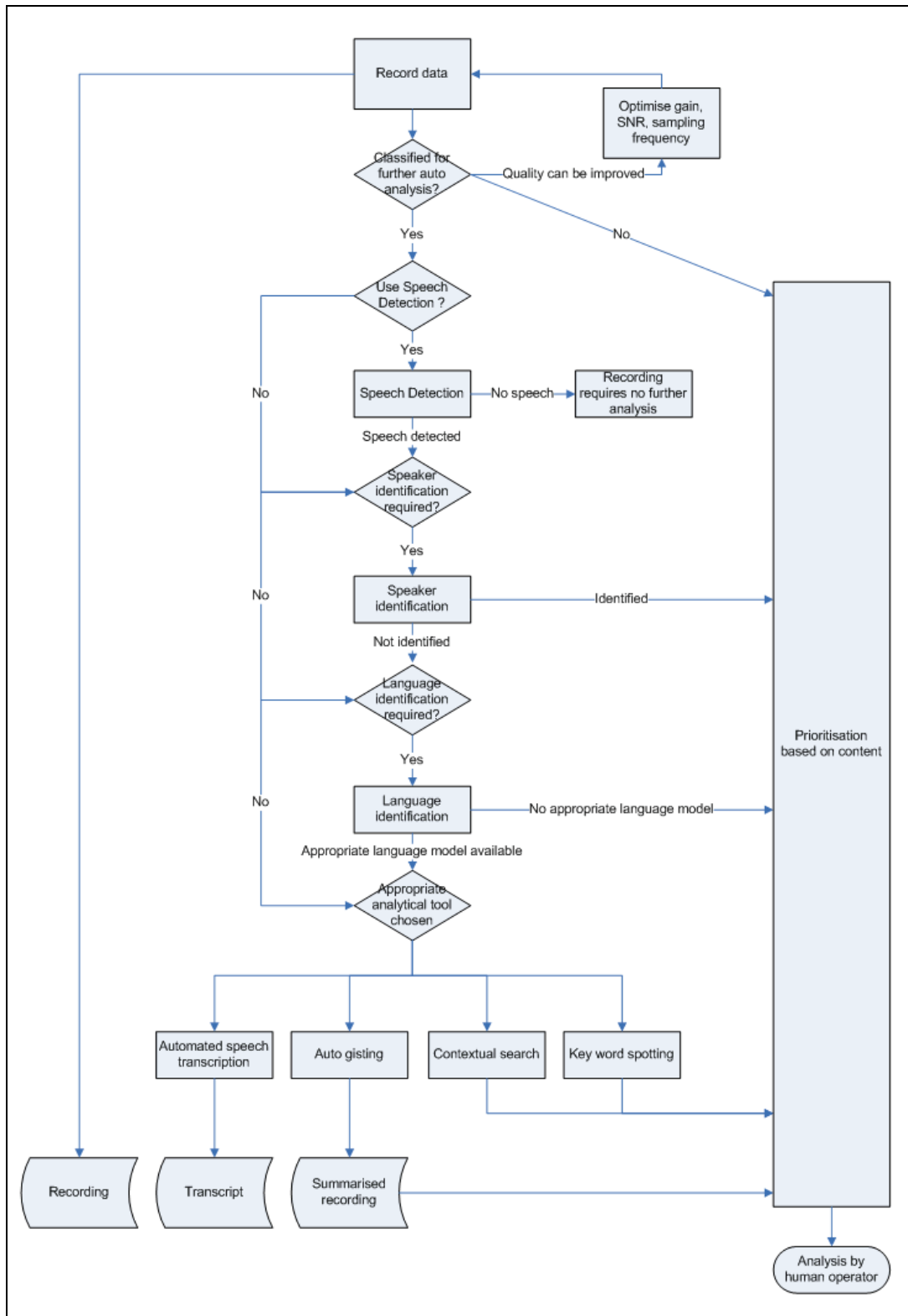


Figure 2: The process from collection through classification to analysis

10.0 Recommended Solution:

It is recommended that analytical tools form part of an end-to-end process and are considered with the recording solution. The table below describes what analytical tools are appropriate for different situations:

	Voice Detection	Automated speech Transcription	Key word spotting	Databasing and textual searches of transcripts	Language Identification	Speaker Identification	Auto gisting	Classification for further analysis
Courtrooms	✓✓	✓✓	✓	✓	✗	✗	✗	✓
Air traffic control communications	✓✓	✗	✓✓	✓	✗	✓✓	✗	✓✓
Lawful Intercept by Law Enforcement Agencies	✓✓	✓	✓✓	✓✓	✓✓	✓✓	✓	✓✓
Police / Immigration interviews	✓	✓✓	✓	✓	✓	✗	✓	✓✓
Emergency Services command and control positions	✓	✓✓	✓	✓	✓	✗	✓	✓✓
Surveillance by Law Enforcement Agencies	✓✓	✓	✓✓	✓✓	✓	✓	✓	✓✓

Table 1: Analysis of different options for analysing recordings

✗ = not appropriate, ✓ = appropriate, ✓✓ = essential

11.0 Conclusions

Best of breed audio analytics tools can now deliver real operational benefits and cost savings. However these benefits can only be realised when the characteristics of the recordings meet certain criteria and are matched with the most appropriate analytic tools. When correctly applied they can augment the current manual resource to greatly improve operational capabilities and identify vital intelligence.

To achieve the best results it is recommended that solutions should be implemented with an end to end systematic approach which includes data collection, recording, classification and analysis.

It is important to realise that any automatic solution should be seen as an enhancement, not a replacement for a human operator, as there will still be some manual input required.

In order to see how you could best reduce your costs and increase your operational capabilities:

- ▶ Speak to your audio analytics provider to discuss the details of your requirements.
- ▶ Supply a sample audio file to facilitate testing of algorithms and tuning to your environment.

12.0 References

1. El-Maleh, K, and Kabal, P, “Comparison of voice activity detection algorithms for wireless personal communications systems”, 1997.
2. Stahl, V, Stadermann, J, and Rose, G, “Voice activity detection in noisy environments”, September 2001.
3. Martin, A and Mauuary, L, “Robust speech/non-speech detection based on LDA-derived parameter and voicing parameter for speech recognition in noisy environments”, *Speech Communication* 48 (2006) 191–206.
4. Denenberg, L, Gish, H, Meteor, M, Miller, T, Rohlicek, J R, Sadkin, W, and Siu, M, “Gisting conversational speech in real time”, *Acoustics, Speech, and Signal Processing*, 30 Apr 1993, Volume: 2, pp 131-134 vol.2.
5. Rohlicek, J R, “Gisting Continuous Speech”

13.0 Acronyms

KWS: Key word Spotting

LVCSR: Large Vocabulary Continuous Speech Recognition

NIST: National Institute of Standards and Technology

Phonemes: The basic distinctive units of speech

SNR: Signal-to-Noise Ratio. A unit-less measure of the amplitude of the signal (in this case speech) compared to the amplitude of the background noise. Many algorithms will factor in the SNR and some will make assumptions about the statistics of the noise.

Training data: Data collected according to a scheme to give software a reference when performing automatic transcription (or using another tool).

14.0 Appendix

The Crampin Index

Correct Words – how many of the words detected by speech engine matched the human transcription.

Phonetically Correct Words – how many of the words detected by speech engine matched the human transcription phonetically, e.g. “wear” instead of “where”.

Detection accuracy – a percentage of the number of words (right or wrong) detected by speech engine versus number of distinct words transcribed by the human operator for the same recording.

Word Accuracy – The percentage of Correct words and Phonetically Correct words against Total number of words found by engine. e.g. engine detects 50 words in total (correct and incorrect) – of these, 8 are correct and 2 are phonetically correct and so scores 20%

The Crampin Index: $\text{Score (\%)} = \text{Detection Accuracy} * \text{Word Accuracy}$