Improving the R in ASR
Tuning Speech Recognition

Executive Summary

- Automatic Speech Recognition (ASR) systems are used to significantly reduce customer churn and interaction costs by enabling voice-of-the-customer analysis.
- Strong out-of-the-box recognition accuracy makes this possible, and accuracy can be further increased through tuning.
- Recognition of non-dictionary product and brand specific terms can be improved using any of various tuning methods positioned at different cost/benefit levels.
- Multiple tuning methods can be applied simultaneously to deliver maximum accuracy.

Introduction

Automatic Speech Recognition (ASR) transcription systems can be used to improve customer retention, reduce customer handling time, expedite customer request routing, and provide the foundation of customer interaction analysis. Significant return on investment can be achieved using out-of-the-box ASR capabilities, but certain optimizations are required to maximize ROI. For example, the quality of the source audio has a significant impact on transcription accuracy. Audio damaged by lossy compression and/or down-mixing to a single channel yields lower-accuracy transcripts than audio not subjected to such data-discarding transformations.

Another important optimization is tuning. ASR tuning is a process that incorporates knowledge about specific characteristics of the speech being transcribed into the ASR process itself. Effectively, the ASR system receives training that brings it into better alignment with speech specific to your business, thereby increasing transcription accuracy.

A variety of tuning processes are available at different cost/benefit points along a spectrum. These can be used individually or in concert to deliver the level of accuracy desired, up to the maximum possible accuracy as limited by source audio quality.

Customer Language is Business-Specific

All businesses use highly differentiated terminology as part of their branding strategy. These include unusual phrases such as slogans and trademarks. Other than branding and company names, business language varies from vertical to vertical, such as health insurance, banking, and consumer products. Business terminologies represent different domains that vary not only in vocabulary but also in how frequently certain words are used as well as the frequency of certain word juxtapositions.

> **The word combination “giant hail” is common within the property insurance domain but is exceptionally rare within the banking domain.**

Language usage variation extends further than just the enterprise level. Different lines of business and different physical locations can also have language usage characteristics that differ significantly from other parts of the business. A single ASR deployment can service all levels and business lines within the enterprise by providing fine-grained and targeted tuning capabilities that make it possible to take advantage of language differences to drive higher accuracy.

(412) 621-9310 | info@vocitec.com | www.vocitec.com | @Vocitec

Deploying a single ASR solution across the enterprise is cost-effective, scalable, and minimizes burden on IT and development groups involved in deployment, use, and support. Support teams can focus their efforts more effectively on a single system and will experience a reduced training burden relative to having to support multiple inflexible solutions.

Tune ASR Like Any Other Engine

Selecting a single, powerful, and flexible ASR engine and taking the time to tune it for different scenarios is an investment that reduces long-term costs while expanding horizons. Many customization/tuning mechanisms and techniques are available, ranging from highly targeted/low cost to highly general/moderate cost. Each method has a different scope of benefits, cost, and development timeline. The following explores some common tuning methods.

Substitutions

Substitution is used in the field to make quick, low-cost improvements to recognition of specific words and phrases. With substitutions, the erroneous text produced by the ASR system is identified and replaced with the corrected version of the text. This enables automatic correction of specific errors that occur frequently and consistently for the audio being processed. Error text can be specified as literal words, or as regular expression patterns for broader error coverage. Errors resulting from pronunciation variations, noise, or audio compression artifacts are typically best addressed with substitution.

> If “date of birth” is frequently transcribed as “data birth”, this can be corrected with the “before : after” substitution rule of “data birth : date of birth”.

ASR transcripts must exist before substitution rules can be developed. These transcripts must be “mined” for errors, and searches across the transcripts are necessary to determine error consistency and frequency. It is also important to listen to call audio to verify that the same word or phrase is being said most of the time for any given error. A tool providing fast search and fine-grained audio playback is invaluable in facilitating the error mining process.

The amount of effort required to create substitution rules is variable depending upon the amount of improvement desired. Doing transcript mining with the intent of creating hundreds of rules can require several days of effort. A benefit of this approach is that applying substitutions has negligible impact on ASR latency and throughput.

Language Modeling

The ASR system utilizes machine-learning components known as “models” to represent knowledge about speech. This knowledge is applied during transcription. Two types of models used are known as acoustic models and language models. The acoustic model converts audio into a stream of basic sound symbols specific to a language, such as English or Spanish. The language model is responsible for converting the stream of sound symbols into text.

General-purpose language models are typically trained to understand domains like banking, health insurance, telecommunications, and voicemail. They provide a strong baseline capability that works well out-of-the-box.

A custom language model is one that has been created for a specific customer, or a specific line of business for a specific customer, using audio provided by the customer. Such a model includes vocabulary specific to the target audio and captures how language is used within that audio, plus it provides a significant boost in overall accuracy when transcribing the target audio. These benefits are achieved because a custom language model is optimally aligned with the speech it is expected to process.

Different levels of custom language modeling can be provided depending upon the customer’s needs and constraints. These can be roughly categorized as “light” language modeling and “full” language modeling.

A portion of the provided audio is transcribed manually and combined with existing language resources to create a custom language model that is better attuned to a specific customer’s calls. The advantage of “light” modeling is that it takes approximately half the time to complete as full custom language modeling.

The effort and cost to create a custom language model is considerably greater than the effort required to create substitutions. However, this method significantly increases overall accuracy with no latency or throughput reduction. In many cases, due to improved alignment and focus, custom language models actually result in improvements in latency and throughput.

Full Custom Language and Acoustic Modeling

The best ASR results are produced by combining a custom acoustic model that is specifically tuned to the customer’s call recordings with a full custom language model. This results in higher immunity from noise and better recognition of accents that occur commonly on the specific customer’s calls. Typically the acoustic model can be developed in parallel with the language model, so while the effort is significantly greater than just creating a full custom language model, the impact on the overall development timeline is nominal.

The combined effort associated with audio collection and model creation requires a significant investment of time before both models can be created and deployed. However, the custom model pair provides the most accurate transcripts. To benefit from the development effort while it is still in progress, in some cases the custom language model can be developed and deployed first, while the acoustic model is still being developed. When completed, the acoustic model can be integrated and deployed, improving transcription quality by delivering the final enhancement to the ASR system.

	REQUIRED AUDIO	LEVEL OF EFFORT	RELATIVE IMPROVEMENT
Substitution	1x	3x	Improved
Language Modeling (Light)	1x	10x	Good
Language Modeling (Full)	2x	15x	Better
Full Custom Language and Acoustic Modeling	3x	30x	Best

Conclusion

Automated Speech Recognition (ASR) enables you to leverage your call recording assets to improve customer retention while reducing the cost of customer care. While out-of-the-box ASR transcription solutions can provide significant value immediately, tuning them can provide greater value and maximize return on your ASR investment.

When selecting an ASR solution, ensuring that each option and alternative is evaluated based on the needs of the organization is critical. Be sure to examine the extent to which an ASR solution can be customized and fine-tuned to suit the needs of both the enterprise and individual users.

Accuracy improvements through tuning can be accomplished using a variety of methods that enable a crawl-walk-run strategy. Improvements to high-value word/phrase recognition can be made immediately using low-effort/cost methods for immediate gratification, with more sophisticated methods that can be applied over time to improve accuracy more generally and as aggressively as desired.

Better accuracy means better analytics insights, and insights drive ROI in the form of retained customers, reduced costs, and the improved strategic market positioning made possibly by quickly understanding what customers like and don't like about your products and those of your competition.

Your call center is a treasure chest of customer opinions, issues, perceptions, and competitive intel. ASR is the key to quickly getting this treasure into the hands of enterprise teams who can use it to take your business to the next level.

About Voci Technologies

Voci Technologies combines artificial intelligence (AI) and deep learning algorithms to deliver the best-in-class enterprise speech analytics platform. Voci's innovative technology and strategic partnerships enable contact centers of all sizes to extract actionable intelligence from voice data to improve customer experience, operational efficiency and compliance requirements.

voci

(412) 621-9310 | info@vocitec.com | www.vocitec.com | @Vocitec