LumenVox©

Streamline Speech Transcription and Reduce Errors

Introducing Next Generation Automatic Speech Recognition (ASR)

Introduction

Businesses using speech-enabled customer experiences via legacy automatic speech recognition (ASR) systems can struggle with expensive errors from poor accuracy, time-consuming processes to train the system to recognize specific words and phrases, and hard-to-distinguish accents and local dialects. With a next-generation ASR engine, companies can use state-of-the-art speech recognition processing technology to serve a more diverse base of users.

This white paper intends to educate companies on use cases for next-gen ASR and how you can incorporate it within your business and deliver more value to your end-customers by:

- Learning how our ASR solution achieves **above-average transcription accuracy at 98.7%**
- Discovering how to **reduce word error rates by up to 52.3%**
- Seeing how to **reduce English dialect errors by over 2x**

At a glance, our new engine ASR software provides your end users with:

- Faster cycle for supporting new dialects and accents, without needing to add new acoustic models
- Faster cycle for new language support
- Ability to handle lengthy transcription
- Increased transcription speed
- Reduced footprint in storage and memory without degradation in speed and accuracy
- Continued support for customer investment in grammars

Highlights of the New Engine Capabilities

End-to-end acoustical modeling with Deep Convolutional Neural Networks (Deep CNN). This extends the benefits of neural networks from the sound-recognition stage all the way to the production of text, removing the need for separate acoustic models to accommodate dialects and accents.

Transfer-learning techniques from existing models to new models, greatly improving learning efficiency.

Use of quantized **Statistical Language Models (SLM).** Efficiently compressed in storage, quantized SLMs enable higher performance in less memory.

Performance and accuracy aided by traditional **SRGS grammars.**

A streaming model that operates online, thereby boosting responsiveness and performance.

End-to-End Acoustical Modeling with Convolutional Neural Networks

Our Deep Neural Network (DNN) engine uses a Convolutional Neural Networks (CNNs) implementation, **a more advanced algorithm that uses patterns to learn individual words.**

In the visual-pattern recognition world, CNN builds on dots, lines, and small areas to identify shapes. In the speech recognition world, CNN "hears" and recognizes small bits of sound; then it builds on sounds to recognize phonemes. Finally, it builds on phonemes to identify words. In contrast, the legacy model does the following:

1. ASR produces a series of phonemes (specific pronunciations).
2. The technology builds an acoustic model, which is then used to make sense of these phonemes and build them into words.

The legacy speech recognition approach may work until the engine encounters a new dialect or a new accent, or sometimes idiosyncrasies in the way a person talks. At that point, it can stop customer service in its tracks because it doesn't know how to map the new phoneme to a word, requiring a new acoustic model.

The new speech engine's end-to-end neural net solves the problem of handling different variations in speech by using **training data to learn how to map different pronunciations to the same word automatically.** The more examples, the more accurate the recognition. Plus, the engine can handle an unlimited amount of dialects and accents — all in one extensive model. Ensuring words are put correctly into phrases and sentences is a separate effort, performed at the end of the acoustic task.

Significant accuracy improvement

Faster cycle for supporting new dialects and accents, without needing to add new acoustic models

Ability to handle lengthy transcription

Transfer Learning Techniques

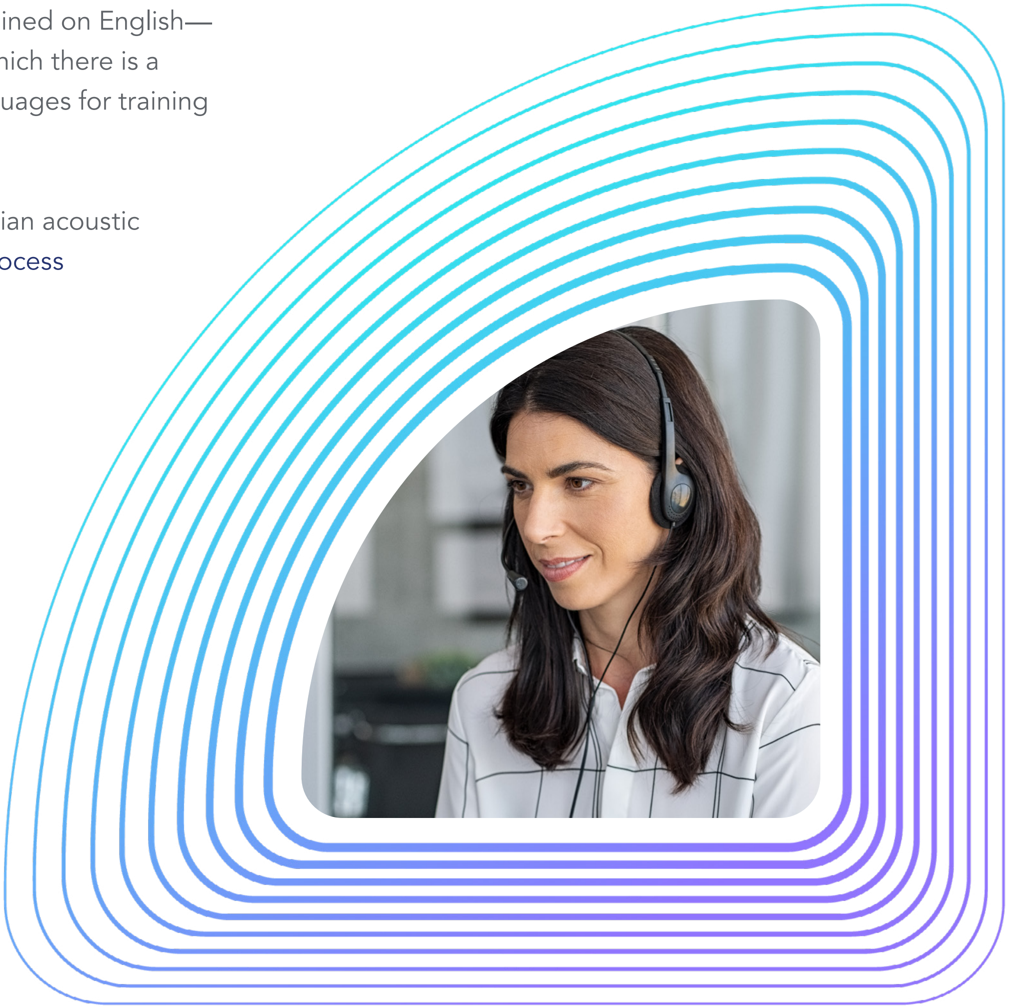
In our case this means, for example, that we can reuse the acoustic model trained on English—for which there are thousands of hours of training available—for Italian, for which there is a reduced amount of training. There is enough similarity between the two languages for training to overlap.

This results in better accuracy and performance compared to training the Italian acoustic model from scratch. Furthermore, transfer learning speeds up the training process considerably (days versus months), reducing the cost to learn languages.

Significant accuracy improvement

Lowered training model costs for new languages

Faster cycle for new language support



Support for the Use of Traditional SRGS Grammars

SRGS (Speech Recognition Grammar Specification, a W3C standard) grammars, written in either BNF or XML style, specify patterns of words that are expected to appear together and the order in which they can appear. They are used as part of the speech recognition process to help predict what combination of words are likely, in the case of ambiguous or unclear pronunciations. Such grammars are typically used by interactive voice responses (IVRs) to also interpret combinations of words to recognize a limited set of possible meanings. For example, "yes," "yes, please," and "okay" are all interpreted as "yes."

The LumenVox next-gen system has the ability to use such grammars in lieu of using a complete language model, for customers and applications that only expect limited words and phrases.

Speech recognition of a limited set of potential words is always more accurate than having the entire language in the pool of candidate words. This works very well in applications where users don't expect the freedom to say whatever they want, but rather they know they have to stick to limited options. Using the new learning capability enables significantly more accurate processing of sounds into text, when using grammars.

Bring word error rate to below 1%

Increased transcription speed

Quantization of Sorted Speech Language Models

We previously discussed the linguistic task of verifying how words are put together, that follows the acoustic task of recognizing the words. The linguistic task is governed by grammatical rules. These grammatical rules are specified in the language model (or in a custom grammar if used). The more complex the language, the more rules are contained in the language model. Storing the more complex language models, if inefficient, affects performance and speed of service when it takes up a great deal of memory.

With the new engine's technology, language models are stored using binary trees, which are optimized for large numbers. As more detailed models are added, the system compresses them, meaning they are faster and more efficient — doing more with less memory.

- Ability to handle lengthy transcription
- Increased transcription speed

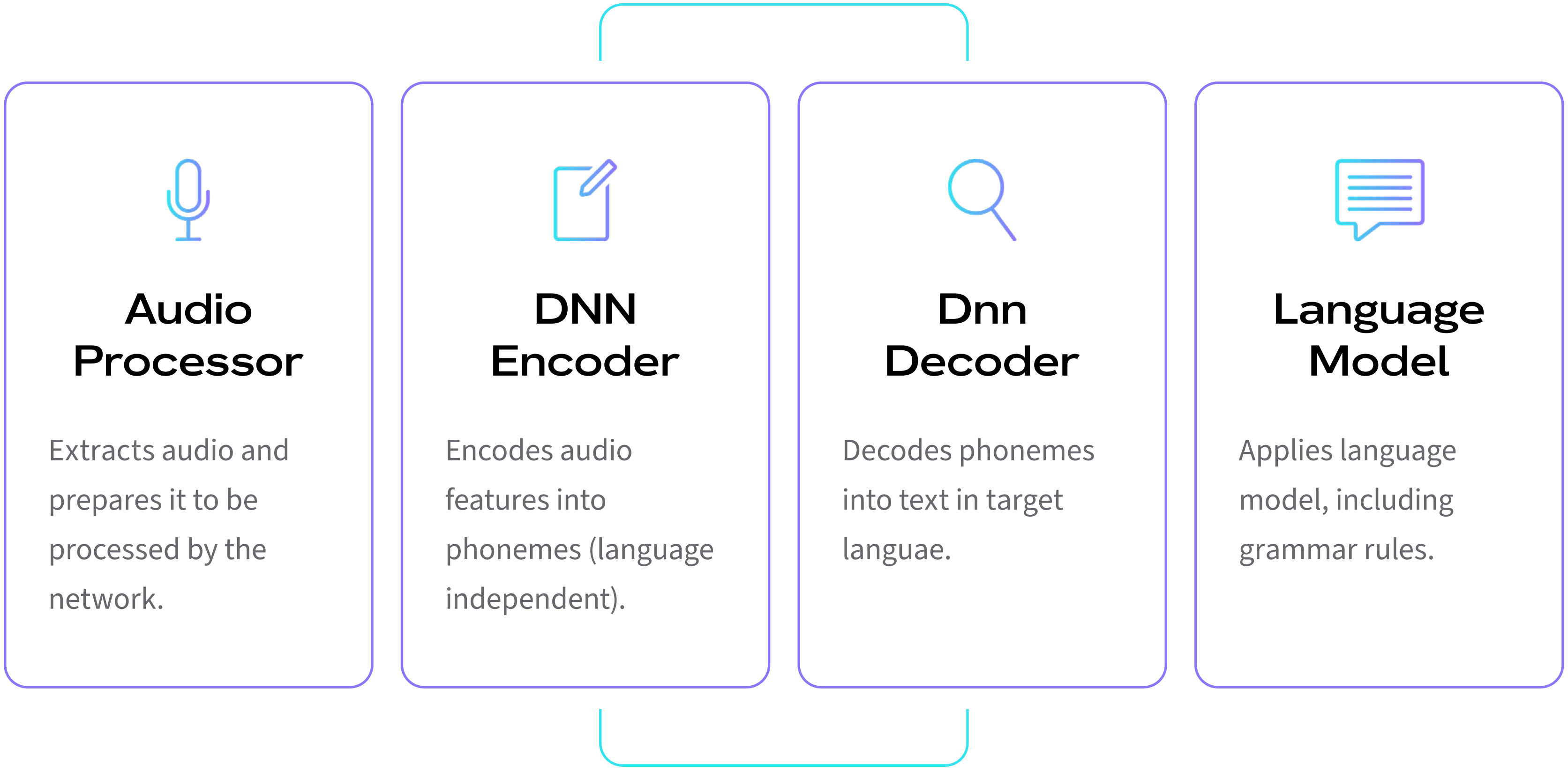
A Streaming Model

The new engine makes direct use of speech input in real-time. Incoming sounds are processed as received, and knowledge picked up during one pass is utilized right away. As a result, there are no multiple passes or extra-cycles of processing, and there is no latency or perceptible delay between the caller speaking and the system responding. The result is a system that is faster and more responsive.

- Increased transcription speed
- Reduced footprint in storage and memory without degradation in speed and accuracy

ASR Engine Architecture

The following block diagram shows the architecture of our engine. It consists of three main blocks, where blocks 1 and 2 perform the acoustic task, and block 3 performs the linguistic task:



Audio Processor

Extracts audio and prepares it to be processed by the network.

DNN Encoder

Encodes audio features into phonemes (language independent).

Dnn Decoder

Decodes phonemes into text in target languae.

Language Model

Applies language model, including grammar rules.

ASR Engine Architecture

Where Legacy HMM Falls Short

Traditionally, speech recognition engines rely on Hidden Markov Models (HMM). HMM algorithms implement statistical analysis using the behavior of an external factor to predict the behavior of an internal, hidden factor. Speech recognition engines look at the external data—the sounds produced—to predict the words intended by the speaker.

There is a disconnect between the process of creating phonemes from the audio, and the process of mapping the phonemes into words. This introduces limits to lexicon size, and also to the size of training data that was used to create them. As a result they fall short in modern times in performance, both in speed and accuracy.

Modern DNN: Training and Benefits of an End-to-End Approach

The next generation of technology is "Deep Learning," implemented through Deep Neural Networks (DNN). At LumenVox, our science researchers have created a modern machine learning system that incorporates all of the required activity that builds an extensible acoustic model that doesn't impose limits on the number of dialects and accents that can exist for each word and does not require separate lexicons and other resources to produce an accurate output.

The neural net is fed a large sample of raw data and corresponding answers and evaluates probabilities to compute an output that is then compared with the answer. This process may be repeated thousands of times. Once the model passes the accuracy checks, it's ready to be fed customer data and output the expected transcription.

This architecture is able to avoid the use of a separate acoustic lexicon because there is no need for a 'source of truth'. The truth is learned from the training data, and many truths are allowed to co-exist by allowing many inputs (audio bits) to map to an output (text) without intervening or limiting the mapping to and from internal phonemes, if the data says they should map. In essence we are replacing the use of linguistic resources with a data-driven learning method. Our state-of-the-art deep-learning system achieves with this end-to-end design significantly better results, that get even better with additional training, both in terms of accuracy and speed.

How Deep Learning Drives Accuracy: Breaking Down CNN

Convolutional Neural Networks (CNN) provide a still-deeper level of analysis, building up levels of comprehension and recognition. CNN recognizes and builds upon smaller patterns and phonemes to produce words.

Traditional HMM systems benefit from training but have less potential for accuracy than deep-learning CNN systems have. Our state-of-the-art deep learning systems achieve **better results and benefit even more from training.**

Language Modeling Starts with Understanding

Unlike acoustic models, our building of a language model has nothing to do with audio. It is all about understanding a language from its building blocks, and that can be done completely by using just the text. What we found is that **deep learning systems benefit from more programmed expertise** up-front. We reap major benefits by using our neural net technology to handle more of the overall process including the language modeling. It is thus truly end-to-end.

Our engine provides out-of-the-box statistical language models trained using millions of text lines from many sources, accounting for several use cases: generic transcription, spellings, digit string recognition, etc.

Acoustic Modeling Captures Linguistic Variations

An end-to-end speech recognizer links audio utterances and resulting text directly. With our new ASR engine, we do not need to use a phonetic dictionary (or lexicon) to transcribe the training text material. The engine itself handles interpretations of phonemes internally.

For each language for which we train the network, **we leverage hundreds or thousands of hours of audio,** covering dialectal variations and environmental conditions. This allows us to train a single model for one language instead of having to train individual ones (e.g., one English model would cover American, British, Australian, and other dialects). Given enough examples, the system will learn to recognize the word "tomato" for both American and British pronunciations.

Customizable Language Model Fits Unique Needs

We offer several options to customize the language modeling part of the recognition:

- Adding lists of words or phrases, which can be done on a per-request basis or caching such requests; this feature boosts the recognition of the specified words or phrases. A typical use case would be adding a list of employees that need to be recognized or specific product names (e.g., pharmaceutical drugs or specific restaurant dishes).

- Adapting the standard language model with a small amount of domain-specific text to boost recognition of in-domain audio. This is for those situations where there is not enough text data to fully train a new language model, but enough so that recognition of domain-specific text is enhanced.

Proof Through Performance: Putting ASR to the Test

Transcription Performance Testing Shows ASR Advantages

Up to **98.7%** transcription accuracy with LumenVox ASR

English Language Accuracy Rate Comparison with Cloud Providers

The performance advantage for the new LumenVox ASR engine is substantial. The transcription engine performed on par or better than the rest of the transcription services evaluated. When a grammar was used instead of free-form transcription for data sets where a closed grammar can be reasonably defined for the vocabulary, the subsequent accuracy was phenomenal.

Engine/ Dataset	Amazon Transcribe	Amazon Lex	Google Speech-to-API	Microsoft Cognitive Services	Deepgram	LumenVox ASR Transcript	LumenVox ASR Grammar
CSLU Digits	94.3%	91.5%	89.3%	98.4%	94.7%	98.5%	98.7%
CSLU Words	37.1%	31.1%	41.5%	62.1%	60%	58.8%	98.7%
CSLU Phrases	71%	66.5%	73.9%	86.3%	76.8%	84.6%	98.6%
Macrophone	Not possible to normalize results	93.4%	Not possible to normalize resutls	Not possible to normalize results	87.8%	95.4%	Not applicable to grammars

English Language Word Error Rate Comparison with Kaldi-Based Models

In performance tests comparing the LumenVox engine and LumenVox English acoustic model against multiple acoustic models that were created using the Kaldi platform and then run with the Kaldi-based engine, our new engine performs better than the Kaldi based engine and models.

Test/Engine	LumenVox ASR Transcript	Kaldi: UK Eng Model	Kaldi: US Eng Model	Kaldi: US Eng + Alpha Numerals
CSLU Multilangual	17.4%	32.2%	17.6%	21.2%
ICSI	43.3%	55.2%	43.7%	41.1%

As much as a **45.9%** reduction in English word error rates

Spanish Language Word Error Rate Comparison with Kaldi-Based Models

Our new engine performs well on all datasets, whereas the Kaldi-based performance drops especially for the open-source datasets. According to the internal README provided with Kaldi models, their engine is trained mostly in telephony speech and may be limited in terms of generalization to different channels and environments.

Test Name	LumenVox ASR Transcritp + SLM	Kaldi Engine + Model
Mexican Sala II	16.2%	24.8%
Voice Across Hispanic America (VAHA)	9.8%	19.3%
Multilangual Librispeech Spanish	15.7%	32.2%
Multilangual TEDX Spanish	21.3%	32.8%
Mozilla Common Voice Spanish	17%	35.7%
Voxforge (random sub-set of whole corpus, may overlap with training data, thus the unrealistically good results here)	4.6%	0.9%

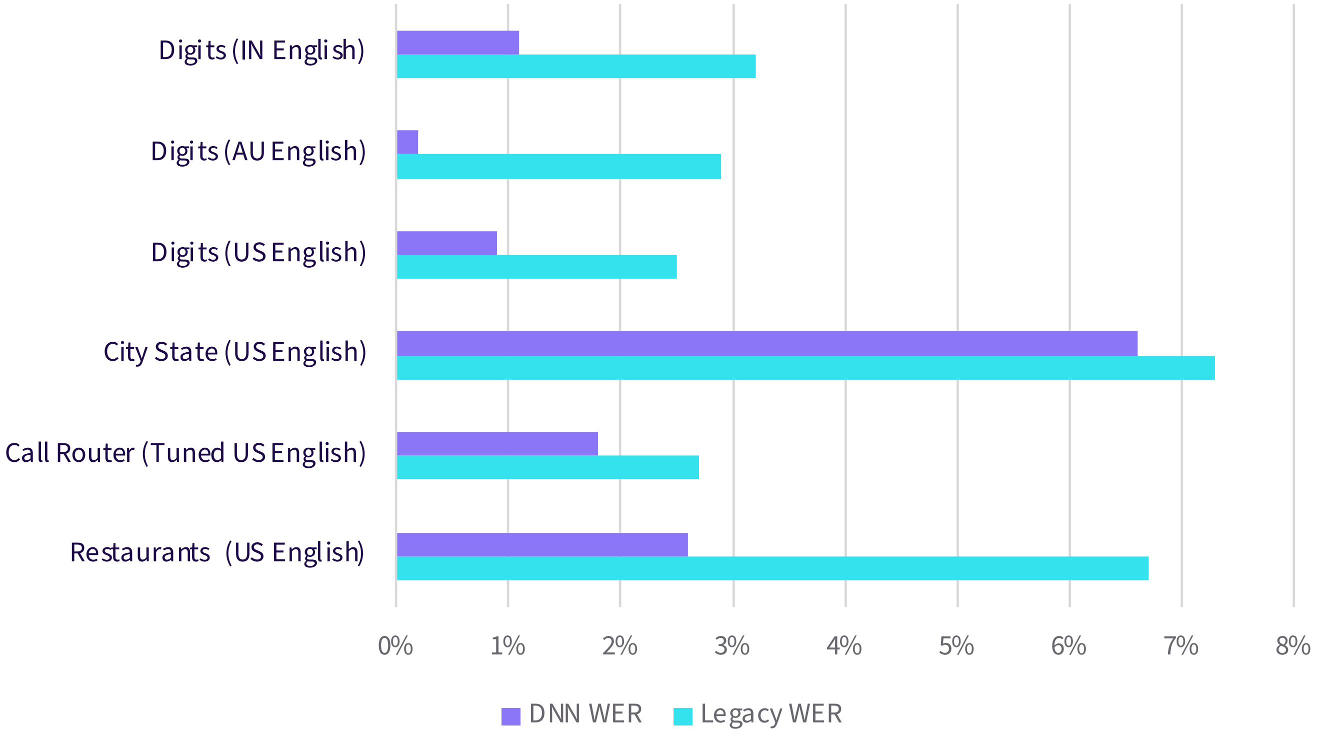
Up to a **52.3%** decrease in Spanish word error rates

Grammar-Based Testing Parses Dialects

Error Rate Comparison with our Legacy Engine

The new engine outperformed the legacy ASR on all the tests, achieving very low error rates. Furthermore, for the digits use case we show how well our model generalizes to the different dialectal variations (American, Australian, and Indian English dialects).

2x or greater
drops in many English variant error rates



The Difference ASR Can Make

- Technologically, there is no turning back. The new ASR delivers a level of accuracy in word recognition that the legacy technology cannot match.
- Legacy systems may perform well enough in the short run, but deep learning with CNN has the advantage of greater knowledge placed into the network up front to the benefits of customers and users. The result is a more expert, and more accurate, speech engine.
- The performance of legacy HMM models tend to be capped at a certain level; beyond that, additional training produces diminishing returns.
- Our state-of-the-art ASR engine also benefits from other technological improvements, such as compact storage of search trees and one-pass technology. Such improvements create a better performing, more responsive system without noticeable latency.
- End-to-end deep learning eliminates reliance on post-processing of phonemes to do the final translation into text. Therefore, a single DNN-based ASR can handle a large domain of input. There is no need for a separate model for dialects and accents. This is a significant benefit for partners, customers, and end-users alike.

Looking to Build Next-Generation Voice Experiences?

Request a demo today to try our ASR with transcription solution with your existing speech recognition software to see how it can save your customers time and money.

Request Demo →

LumenVox®

About LumenVox

LumenVox transforms customer communication. Our flexible and cost-effective technology enables you to create effortless, secure self-service and customer-agent interactions. We provide a complete suite of speech and authentication technology to make customer relations faster, stronger and safer than ever before. Our expertise is extensive— we support a multitude of applications for voice biometrics, inclusive of passive and active authentication for fraud detection. And we do it all by putting you and your customers first.

Interested in finding out more about this product?

Learn More →

www.LumenVox.com

Contact

LVsales@lumenvox.com

+1 (858) 707-7700