

Model Card: Einstein Call Coaching Language Model for Sales Cloud

Basic Information

A Language Model (LM) calculates the likelihood of a sequence of words. The model estimates the probability of a word given an observed word sequence and determines the next word probability. The LM calculates higher probabilities to “real” and “frequently observed” sentences than the ones that are wrong accordingly to natural language grammar or those that are rarely observed, which is how the final output is determined.

Since the LM captures the possibility of any word sequence, it also helps to distinguish between words with similar sounds. For example, it can predict the probability of the occurrence of light, night, fight, or right. Similarly, it can assemble phonemes to predict that “nowhere is” is, in fact, “no worries.”

A phoneme is one of the set of speech sounds in any given language that serve to distinguish one word from another. A phoneme may consist of several phonetically distinct articulations, which are regarded as identical by native speakers, since one articulation may be substituted for another without any change of meaning. Thus /p/ and /b/ are separate phonemes in English because they distinguish such words as pet and bet, whereas the light and dark /l/ sounds in little are not separate phonemes since they may be transposed without changing meaning.

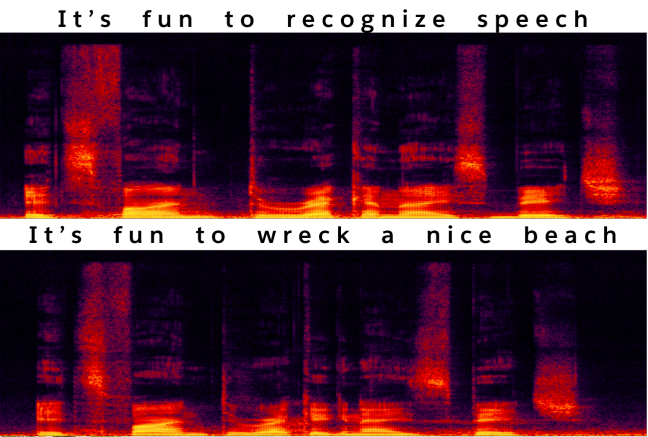
Basic Transcription Workflow

Speech recognition systems use both an acoustic model and a language model to generate a transcription from an audio input. These two models combined are responsible for analyzing audio signals, and outputting the speaker’s most likely utterance.

The **Acoustic Model** models the relationship between the audio signal and the phonetic units in the language. It determines the phoneme probability given an audio frame:

- **Phoneme** - a single unit of speech which is perceptible to the listener (a -> ae, ie)
- **Audio Frame** - a short segment of audio signal assumed to include at most one phoneme

The **Language Model** is responsible for assembling phonemes into meaningful words and sentences in the target language. For example, a certain set of phonemes may sound like “It’s fun to recognize speech” and “it’s fine to wreck a nice beach.” The Language Model is responsible for determining the most likely sentence.



Copyright: [Salesforce.com](https://www.salesforce.com)

Model Details

Person or organization developing model	Salesforce, Inc.
Model date	July 1, 2020
Input	Phoneme sequence
Training algorithms, parameters, fairness constraints or other applied approaches, and features	Kaldi speech recognition toolkit Fisher English Training Speech Part 1 Transcripts G2P (Grapheme-to-Phoneme)

Output	In language modeling for speech recognition the goal is to constrain the search of the speech recognizer by providing a model which can, given a context, indicate what the next most likely word will be including the start and end time of the spoken word in the audio file.
Licenses for training data	Kaldi Apache 2.0 Fisher English Training Part 2 LDC For Profit Membership Agreement g2p-en 2.1.0 - Apache Software License g2p-seq2seq - Apache Software License Phonetisaurus - BSD-3-Clause License
Send questions or comments about the model to:	ecc-ethics@salesforce.com

Intended Use

Primary intended use

The primary intended users of this model are Sales Managers. The use case is to create a transcription (internal) for the entire conversation and highlight important moments in sales call (external). Managers can zoom in to the right part of the conversation quickly in a call-recording player and listen to the important moments in a discussion. Then, managers can tailor and personalize coaching, by focusing on reps' moments of success and the areas for improvement.

Out-of-scope use cases

The software should not be used to track or profit from:

- financial information (such as credit or debit card numbers, any related security codes or passwords, and bank account numbers)
- specific people's names
- hate speech, harassment, and violence
- feeling, emotion and sentiment
- sensitive data, such as government issued identification numbers, racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, a information concerning sex life, and health information.

Training Data

The model is trained on:

- Fisher English Training Speech Part 1 Transcripts
- Words added by the Admin in the Call Coaching Setup page
- Sales conversations

Fisher English Training Speech Part 1 Transcripts represents the first half of a collection of conversational telephone speech (CTS) that was created at LDC in 2003. It contains time-aligned transcript data for 5,850 complete conversations, each lasting up to 10 minutes.

Out-of-Vocabulary words contains a list of product names, company names and custom trackers (any word or utterance) defined in the admin setup of Einstein Call Coaching by Salesforce customers actively using the product.

Sales conversations contain randomly selected, manually transcribed conversations from various sectors, scrubbed to remove personal data using commercially reasonable methods, and tagged with the relevant OOV keywords. These scrubbed conversations are treated as Customer Data and deleted when a Customer terminates their subscription.

Metrics (model performance measures)

Word Error Rate (WER)

Our primary measures for the model's performance are Precision and Recall. A major factor that contributes to the model's accuracy is Word Error Rate (WER), which is concerned with capturing the similarity between what was spoken and what is transcribed.

WER (Word Error Rate) is the number of errors divided by the total words.

$$\text{Word Error Rate} = \frac{\text{Insertions} + \text{Deletions} + \text{Substitutions}}{\text{Number of Words in Reference Transcript}}$$

*A **substitution** occurs when a word gets replaced (for example, "voice" is transcribed as "noise")

*An **insertion** is when a word is added that wasn't said (for example, "before" becomes "be four")

*A **deletion** happens when a word is left out of the transcript completely (for example, "take it out" becomes "take out")

Precision & Recall

The Call Coaching algorithm is looking in the transcription for keywords defined by customers and for out-of-the-box insights (such as pricing discussions, next steps etc.). The accuracy of the search results are evaluated using Recall and Precision measures. The results of the search may be impacted by various parameters: Word Error Rate (WER) - the quality of the transcription, as described above - is a major contributor, but also included are elements such as sound quality (e.g. background noises, distance from microphone), number of speakers, the methods we use to detect keywords and insights and more. Ultimately our Precision and Recall metrics optimize for the following question: Did we successfully find the keywords and insights customers were looking for?

The minimum committed threshold for moment detection accuracy (e.g. detection of specific keywords setup by the customer) is 80% precision and 40% recall. Our pilot results were **89% precision and 83% recall**.

PRECISION RESULTS: 89%

$$\text{Precision} = \frac{\text{Keywords Captured Correctly}}{\text{Correct Detection} + \text{Incorrect Detection}}$$

RECALL RESULTS: 83%

$$\text{Recall} = \frac{\text{Keywords Captured Correctly}}{\text{Total Relevant Keywords}}$$

Ethical Considerations

We recognize the potential for inaccurate transcription of voice data. Regional and social dialects differ in syntax (sentence structure), phonology (sound structure), and the inventory of words and phrases (lexicon). Background noise (e.g. driving etc.) may also cause inaccurate transcription. At this stage, the transcription is not visible to customers without filing a support ticket (this may change in future releases).


To mitigate potential misuse to the extent possible, we provide a built-in feedback mechanism in the Einstein Call Coaching application. Each voice call page has the option to report whether any mention highlighted in the call was captured correctly or not. This feedback is then sent to our AI solution team for further analysis, allowing us to normalize potential biases against end customers and reps.

We respect the privacy of employees, customers and partners personal data and information. During the setup process, admins are provided with guidance and resources to set up the voice product in a way that promotes our values of Trust, Customer Success, Innovation and Equality. For example, admins are advised to be conscious of the keywords and phrases they select to avoid flagging sensitive information and to limit keywords that might inadvertently

or inappropriately impact persons differently based on gender, religion, race, sexual orientation, income level, or any other sensitive category.

Similarly, admins are advised to limit keywords that could be construed as surveilling or monitoring employees; the feature should not be used to assess employee satisfaction, gauge performance, or define normative behaviors.

Competitor Mentioned



▼ Competitor Names

Add the names of up to 25 competitors.

Keyword 1

ACEM Technology

Keyword 2

Globex

Keyword 3

Soylent

Keyword 4

Initech

Keyword 5

Tools LLC

Keyword 6

Presidio Tech

Keyword 7

Technology Inc

Keyword 8

Indeed

Keyword 9

No Doubt

Keyword 10

NoDoubt

Show All

Tips for Success

Use specific keywords and phrases

Competitor Mentioned insights are the competitor names you want to track in voice conversations.

- Use exact company names, including numbers.
- Don't add punctuation, with the exception of periods and apostrophes.
- For example: Bob's Tires, Service 365, 8H

Limit bias

Be conscious of the keywords and phrases you select to avoid flagging sensitive information and to ensure the trust, safety, and privacy of your employees and customers. Limit words that might inadvertently or inappropriately impact persons differently based on gender, religion, race, sexual orientation, income level, or any other sensitive category.

Similarly, limit keywords that could be construed as surveilling or monitoring your employees; this feature should not be used to assess employee satisfaction, gauge performance, or define normative behaviors.

The details of each call and its recording file are stored in the Voice Call record, which is a standard object in Salesforce. An admin can manually delete a Voice Call record (via the Delete button) from the Voice Call record details page. This will delete all relevant data associated with the call across our technology stack. Admins can access Voice Call records via the Voice Call list view or a custom Voice Call report.

The language model is constantly evaluated internally by our AI solutions team. The team manually transcribes a random sample of calls on a weekly basis from a variety of customers. The sample is representative, balanced and contains sufficient mentions of each entity type. During the data annotation process, special entities such as competitor names, product names and Personally Identifiable Information (PII) are labeled. This labeled data is then scrubbed from the transcripts prior to further analysis or model training to protect our Customers confidential information. This ensures accuracy, and creates a gold standard to test our system's performance.

Caveats and Recommendations

One of the main challenges is the understanding and modeling of elements within a variable context. In a natural language, words are unique but can have different meanings depending on the context resulting in ambiguity on the lexical, syntactic, and semantic levels.