Generative AI for Dialog Systems
Text-to-Speech and Large Language Models

Omilia
Conversational Intelligence

Themos Stafylakis
Head of Machine Learning and Voice Biometrics

Outline

- Omilia Conversational AI
  - Omilia Cloud Platform (OCP)
  - Task-oriented Dialog Systems
    - ASR, NLU, Voice Biometrics, Dialog Manager, NLG & TTS
- Generative AI for Dialog Systems
  - Text-to-speech
    - From WaveNet to VALL-E
  - Large Language Models
    - Use-cases for task-oriented dialog systems

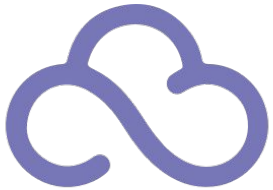
© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

About Omilia

- Founded in 2002 by Dimitris Vassos (CEO) and Pelias Ioannidis (CFO).
- It creates automated dialog systems for customer care via voice, text, social media, a.o. communication channels.
- About 340 people mainly in Greece (Athens and Chania), Ukraine (Kiev), Cyprus (Limassol), Azerbaijan, Czech Republic, and USA.
- Omilia's clients are some of the largest financial institutions, telecoms, healthcare and insurance organizations, retail, e-commerce, restaurants, a.o. in USA, Canada, EU, Ukraine, a.o.
- During the last few years, Omilia is focusing on its cloud solution, Omilia Cloud Platform (OCP).

Omilia
Conversational Intelligence

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd



Omilia Cloud Platform

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

What is OCP®?

OCP® is a Conversational AI platform for customer care that

- requires no-code,
- offers enterprise-scale customer service automation,
- works across voice and text channels,
- offers hyper-tuned out-of-the-box industry models,
- and ready-to-go integration with most leading CCaaS platforms.

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Architecture

Speech to Text

Translating the spoken language to accurate text

Natural Language Understanding

Understanding the semantics of the captured text (intent, entities, dialog acts, relations)

Voice Biometrics

Validating a user's identity through AI, using their voice characteristics.

Dynamic Dialog Management

Associating an intent with the desired outcome, planning dialog and tracking context

Pre-built Skills

Billions of real interactions, distilled into hyper-tuned ready to use models.

Text to Speech

Generating dynamic speech that sounds human-like

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Pre-built Tasks, Domains, and Industry Modules

Omilia's pre-packaged solutions come with a fully preloaded intelligent Virtual Assistant providing out-of-the-box recognition and understanding of all key concepts for a specific domain and language.
These pre-build packs come with Concept Annotation Dictionaries, Rules, and Intents

Industry Skills

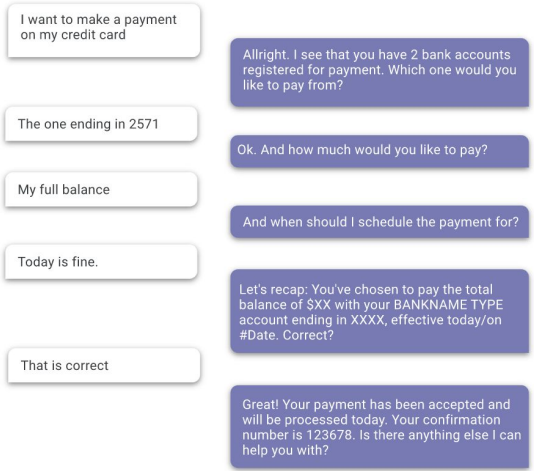
- Banking
- Healthcare
- Telecoms
- Retail
- Travel
- Insurance
- Utilities
- E-commerce

Banking Skills | Healthcare Skills | Telecom Skills | Ret...

+400 Intents Ready-to-Go

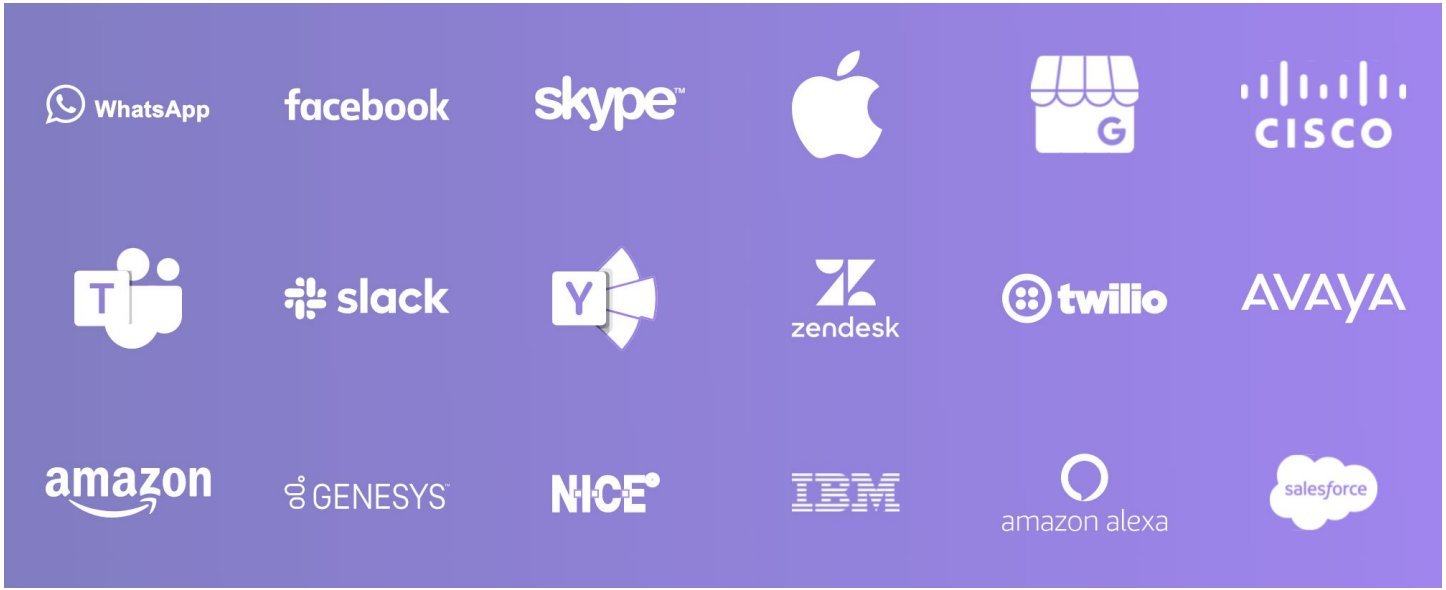
- Account_Balance
- List_Transaction
- Bill_Payments
- Activate_Card
- Loyalty_Rewards
- Transfer_Funds
- Branch_Locator

Virtual Agent Flows Ready-to-Go



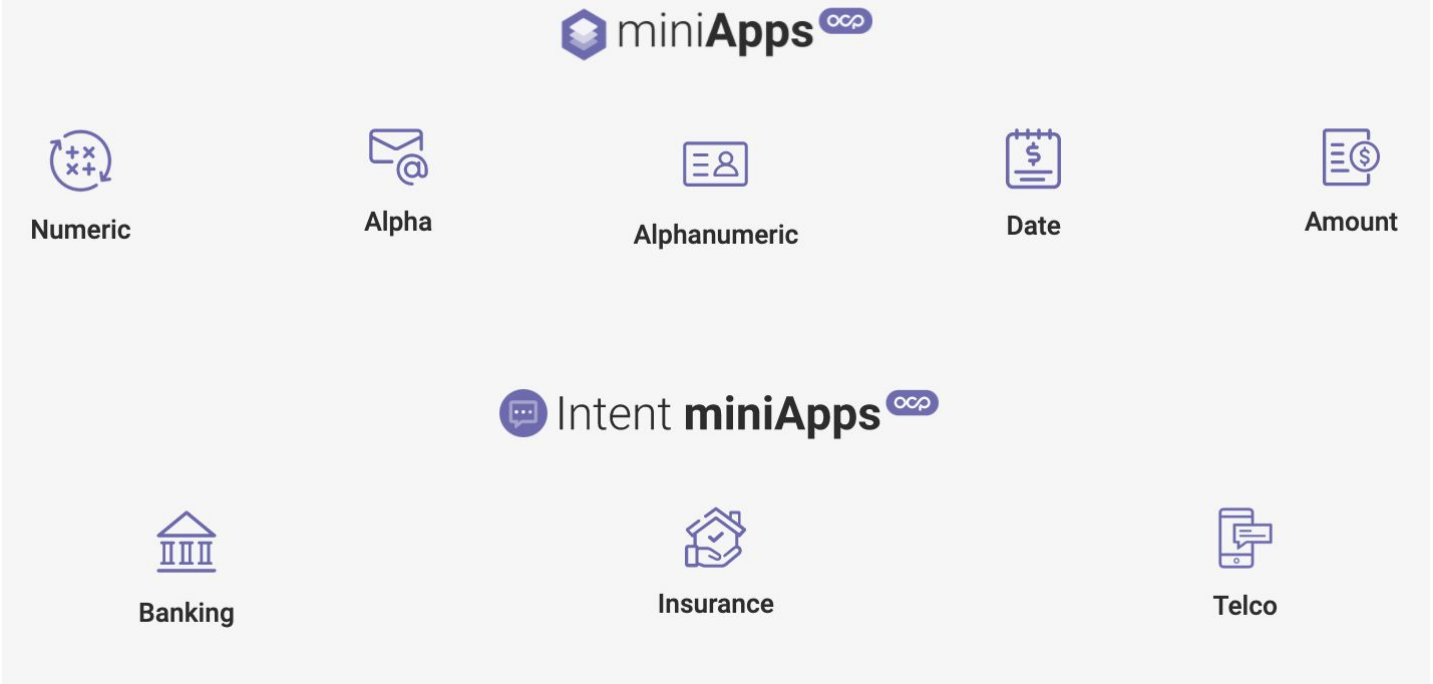
© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Unlimited channel integration



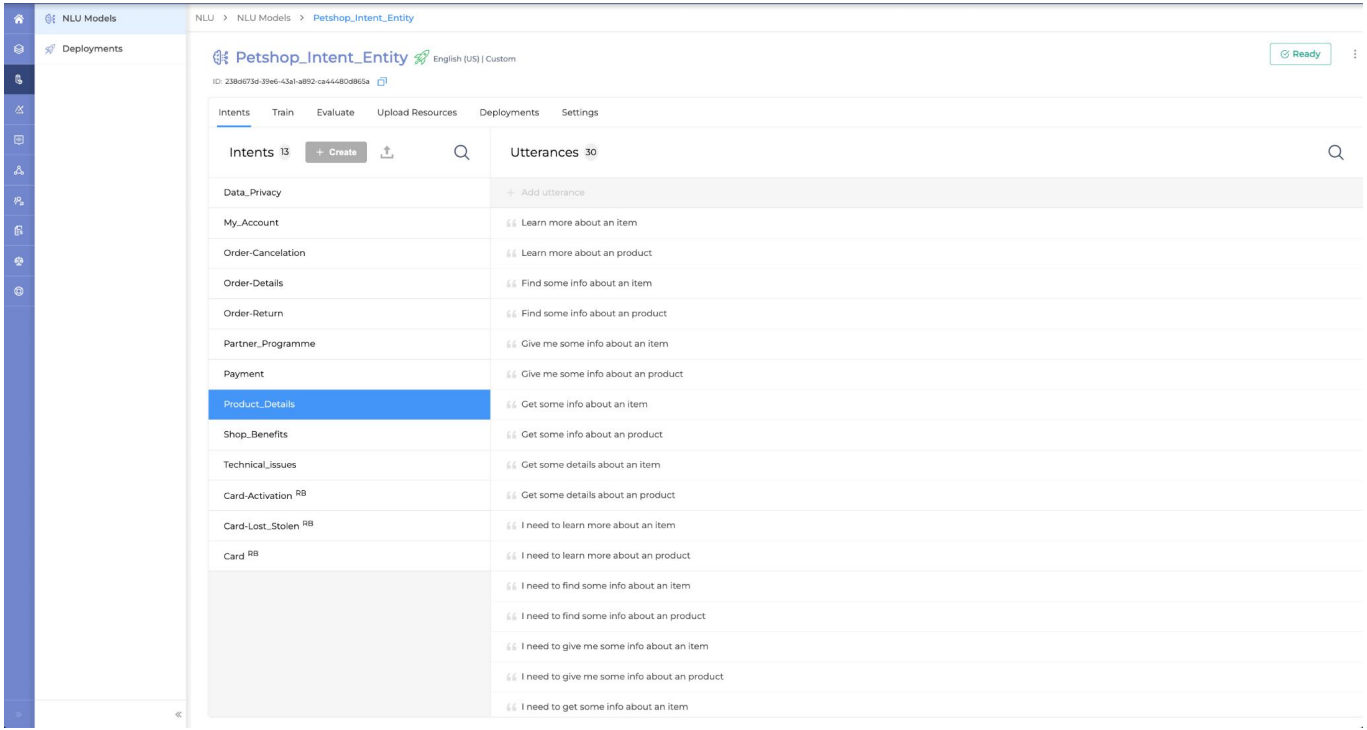
© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

miniApps®



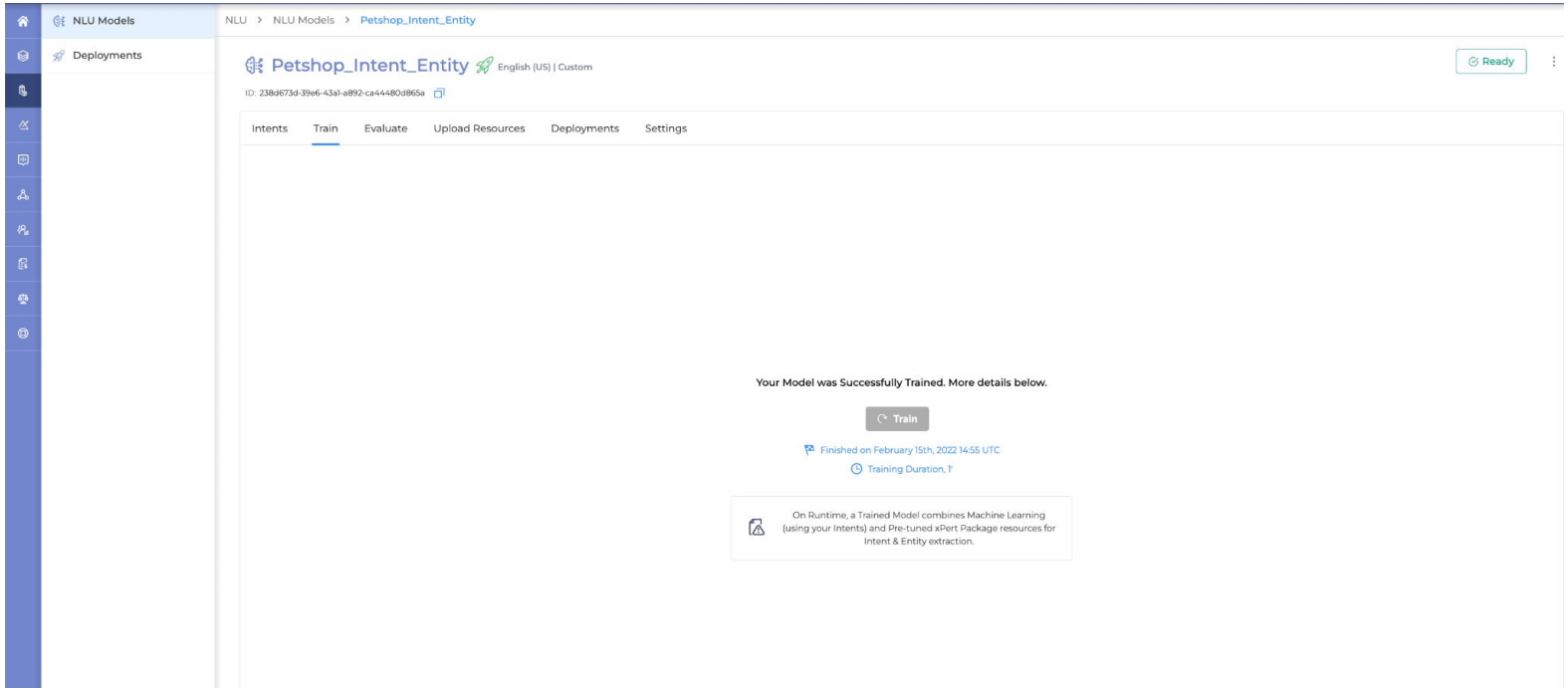
© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Intent MiniApps®



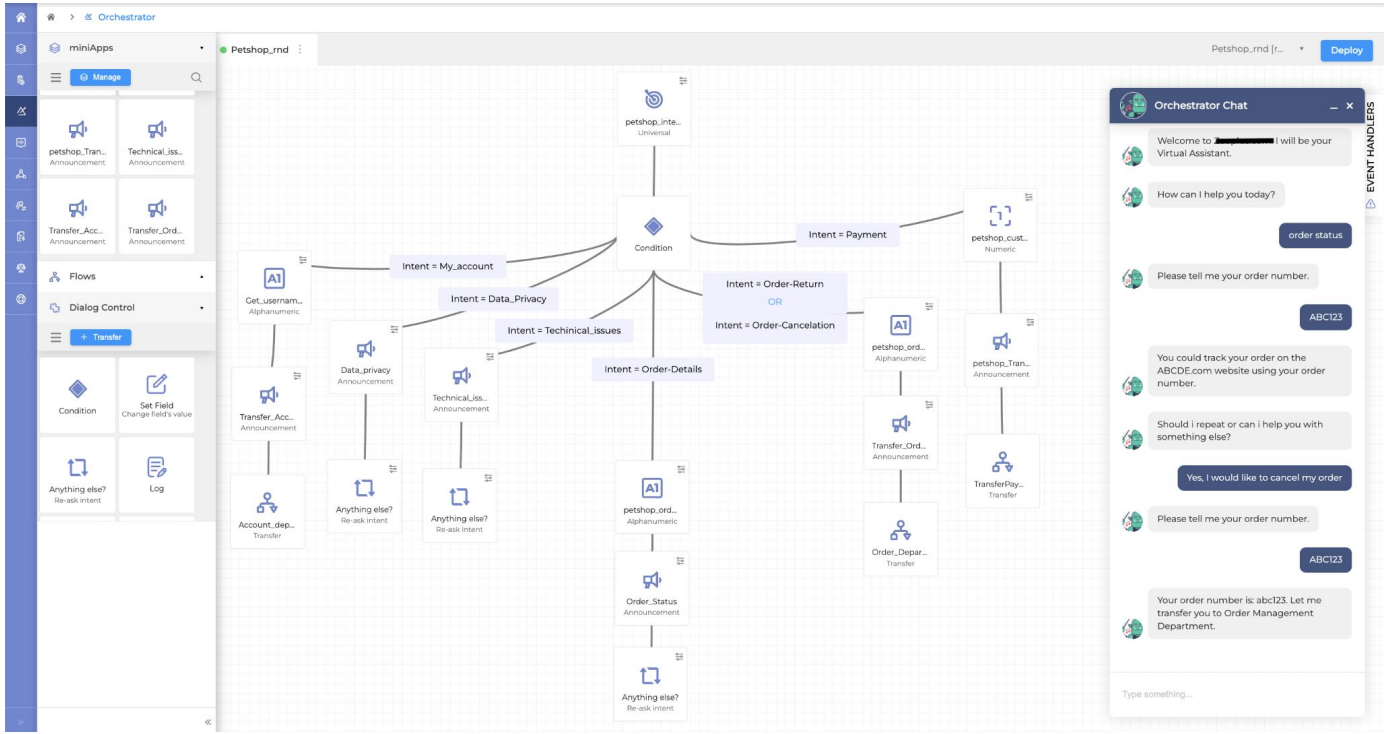
© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Intent MiniApps®



© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

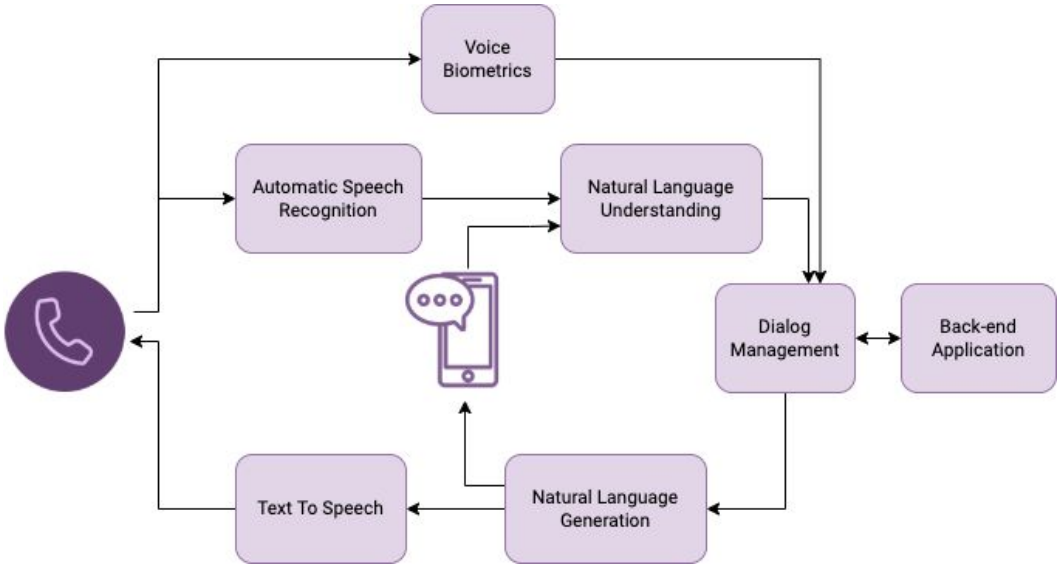
Orchestrator®



© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Conversational AI

Voice-enabled task-oriented dialog systems



© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Automatic Speech Recognition

Omilia ASR Services

deepASR®

- Conversational Speech-to-Text (real-time)
- Transcription Service (off-line)

Deployment: Cloud, On-Premises

Domains: Financial Services (Banking, Insurance, Investment), Telecoms, Healthcare, Government, Retail/Wholesale, Manufacturing, Transportation (Travel & Hospitality), Utilities, E-Commerce

Features: Context-sensitive & dynamic recognition grammars, Rich transcriptions (Time stamps, Confidence Scores)

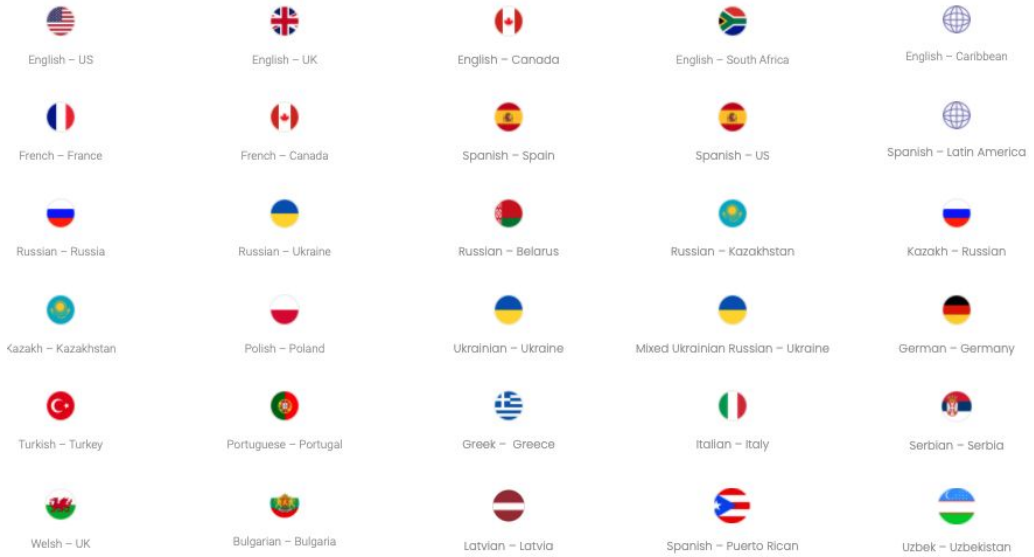
Customization: Acoustic Model, Language Model (word, sentence boosting), Lexicon (add new words)

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Automatic Speech Recognition

ASR models in 30 languages and dialects

deepASR®



© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Voice Biometrics

Authenticate customers in the background as they speak

Enrollment phase

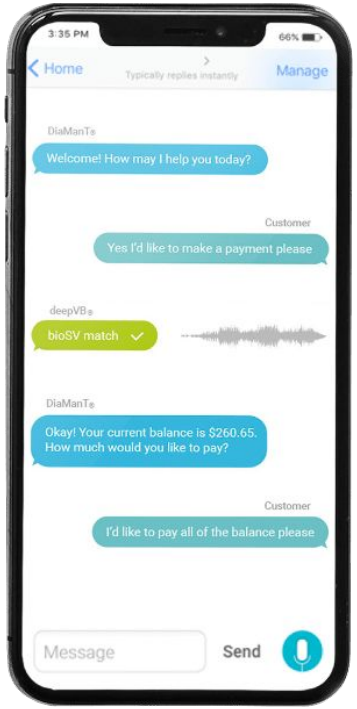
deepVB®

1. The callers are asked to enroll in the VB system.
2. If they opt-in, their voice sample is stored in the VB server and a **voiceprint** is created and stored in the database.

Note: The user may ask to unsubscribe from the VB system anytime.

Verification phase

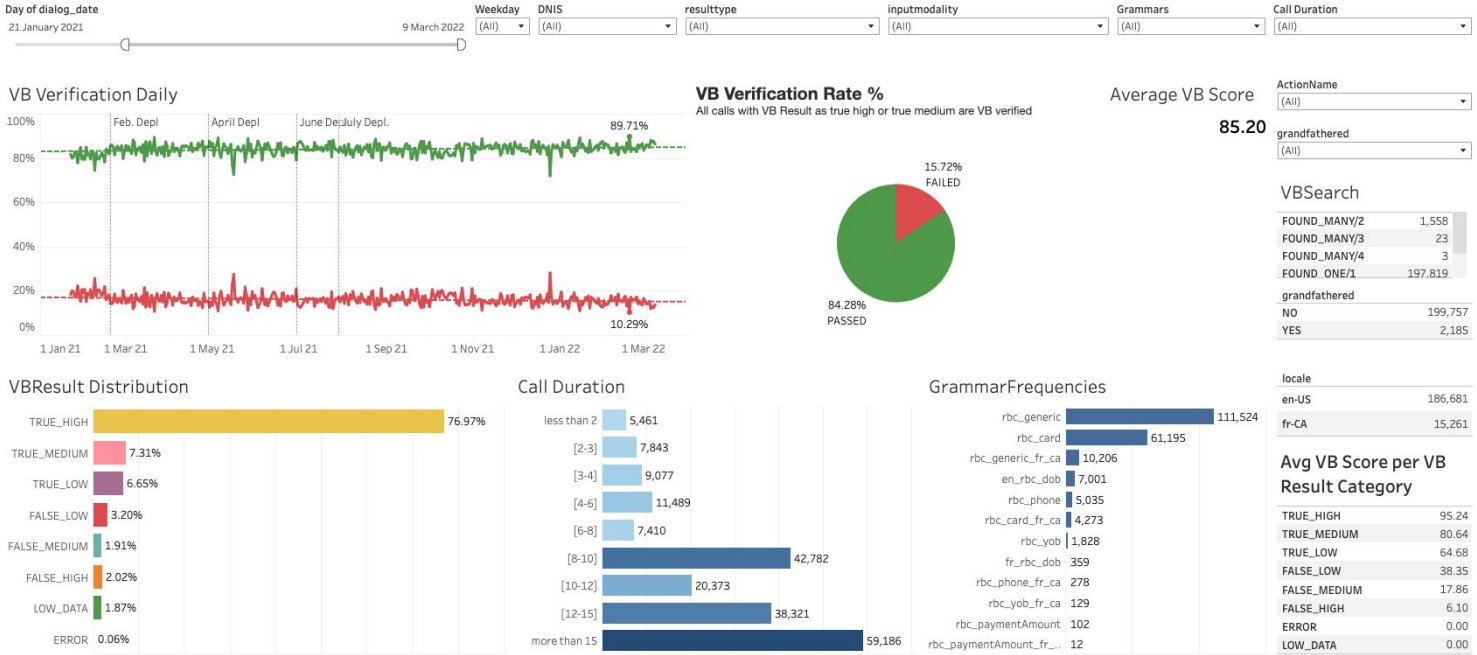
1. The VB system samples the speaker's voice and computes a voiceprint in real time.
2. It then compares the speaker's current voiceprint with the stored voiceprint and generates a "bioSV Confidence Score".
3. bioSV Confidence Score is used by the dialog manager to drive the conversation with the customer according to the business logic; deliver self-service, transfer to live agent, or ask additional traditional security questions.



© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Voice Biometrics

Tools for analysing VB results



© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Natural Language Understanding & Dialog Manager

DiaManT®

deepNLU®

xPert Packs®

Extracting meaning from utterances

- **Intents**
  - I want to access my bank account ⇨ Access_AccountType
  - I need online access ⇨ Access_TelcoChannel
  - I want to reactivate my accounts ⇨ Activate_Account
  - I call to cancel a warranty ⇨ Cancel_Warranty
- **Dialog Acts**
  - I need to speak to a customer service about logging into my account ⇨ AgentRequest
  - Balance again please ⇨ RepeatRequest
  - No, what I asked ⇨ Rejection
- **Entities**
  - Examples: AccountType, Airline, amountToRedeem, ComChannel, DayOfMonth, InsuranceType, OfferPeriod, SentimentDescriptor, TransactionPeriod, USZipCode
- **Relations**
  - Examples in the domain of food ordering will follow

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

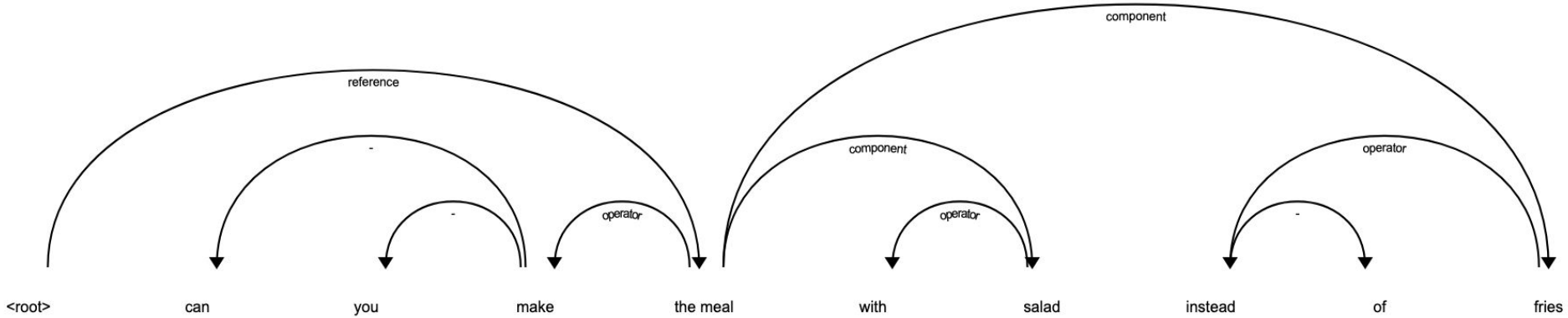
Natural Language Understanding & Dialog Manager

DiaManT®

deepNLU®

Extracting meaning from utterances

- **Relation extraction**
  - Examples in **food ordering:**
    - Can you make the meal with salad instead of fries?
- **Ontology**
  - Generic and/or restaurant-specific (e.g. derived from the menu)



© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Text-to-Speech

Transform text to speech

- **English**

- **Ukrainian**

- **Russian**

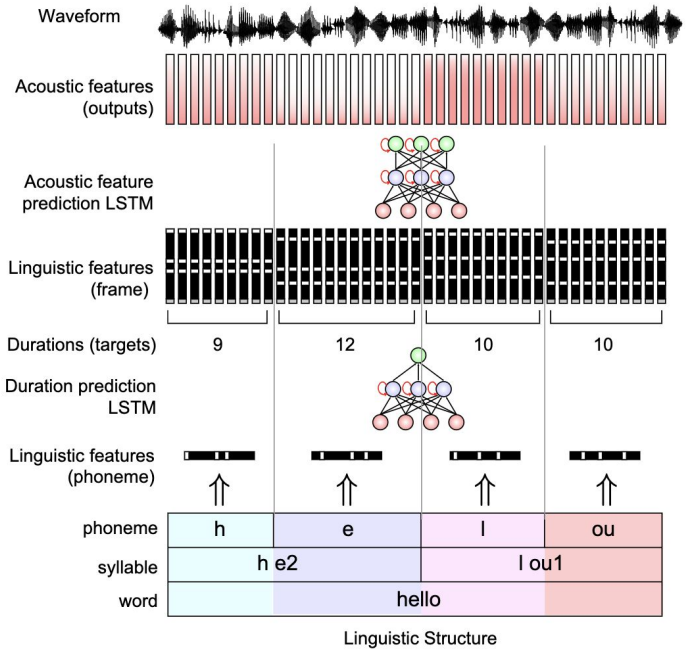
- **Neural TTS synthesis**
- **Conversational TTS**
  - The voices are geared to be used in conversation in real-world applications
- **Ability to control prosody based on punctuation**
  - Do you want to cancel your card because it is lost or stolen?
  - Do you want to cancel your card because it is lost, or stolen?
- **Ability to voice important information (dates, digits, addresses, etc.)**
- **Ability to control emphasis in order to highlight important information in a conversation**
  - E.g. The four digit number you need is **5432**.
- **French** and **Spanish** are also supported.

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Generative AI for Dialog Systems
Text-to-Speech

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Text-to-Speech - Statistical Parametric Speech Synthesis



H. Zen et al (Google AI). Fast, Compact, and High Quality LSTM-RNN-Based Statistical Parametric Speech Synthesizers for Mobile Devices, 2016

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Text-to-Speech - WaveNet (2016)

WAVENET: A GENERATIVE MODEL FOR RAW AUDIO

Aäron van den Oord **Sander Dieleman** **Heiga Zen**[†]

Karen Simonyan **Oriol Vinyals** **Alex Graves**

Nal Kalchbrenner **Andrew Senior** **Koray Kavukcuoglu**

{avdnoord, sedielem, heigazen, simonyan, vinyals, gravesa, nalk, andrewsenior, korayk}@google.com
Google DeepMind, London, UK
[†] Google, London, UK

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Text-to-Speech - WaveNet (2016)

Abstract

- A **deep neural network** for generating raw **audio waveforms**.
- It is fully probabilistic and **autoregressive**:
  - It models the predictive distribution for each audio sample conditioned on all previous ones:

$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1})$$

- It can be **efficiently trained** on data with tens of thousands of samples per second of audio.
- Significantly more natural sounding than the best parametric and concatenative systems.
- A single WaveNet can capture the characteristics of **many different speakers**:
  - It can switch between them by conditioning on the **speaker identity**.
- It can even be trained to **model music** or to do **phoneme recognition**.

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Text-to-Speech - WaveNet (2016)

Dilated Convolutions

- Increase the receptive field while keeping the number of parameters fixed.
- Also used in ASR, e.g. time-delayed NNet (TDNN in Kaldi).
- Convolutions are much faster in training than recurrent models (e.g. LSTMs or GRUs).
- Why? Teacher forcing.
- However generation is still very slow, due to its autoregressive architecture (no gain over RNNs).

$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1})$$

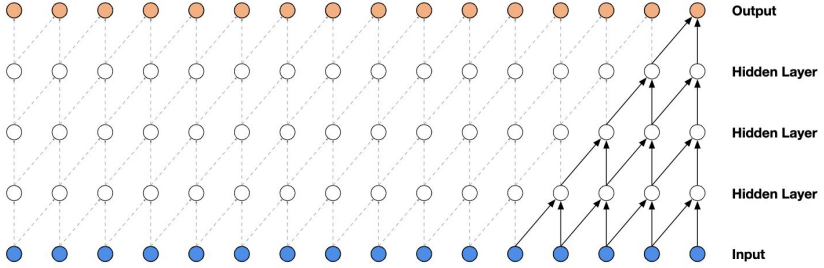


Figure 2: Visualization of a stack of causal convolutional layers.

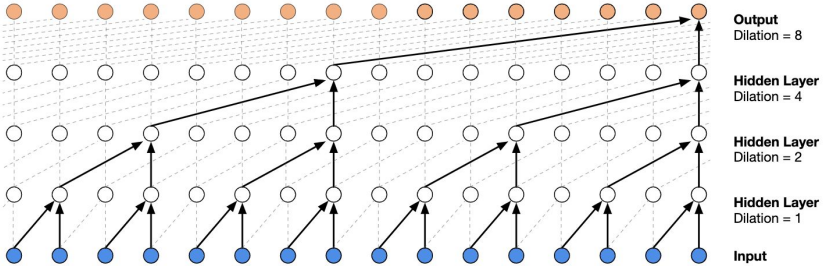
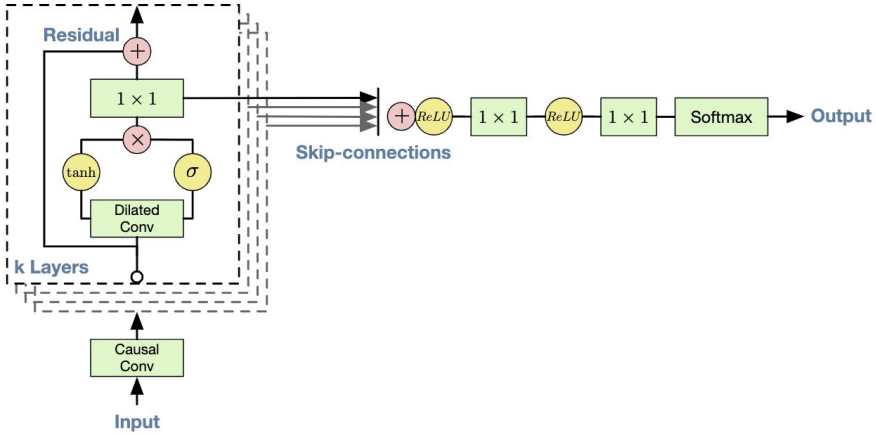


Figure 3: Visualization of a stack of *dilated* causal convolutional layers.

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Text-to-Speech - WaveNet (2016)

Architecture



- Residual Connections: Each layer adds details to the reconstructed signal
- Gated Activation Units: $\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x}\right) \odot \sigma\left(W_{g,k} * \mathbf{x}\right)$
- Softmax: Input and outputs are quantized to 256 levels/classes

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Text-to-Speech - WaveNet (2016)

Conditional WaveNet

$$p(\mathbf{x} \mid \mathbf{h}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}, \mathbf{h})$$

Condition Wavenet to mimic the voice of a speaker ($\mathbf{h}$ is a speaker embeddings):

$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x} + V_{f,k}^T \mathbf{h}\right) \odot \sigma\left(W_{g,k} * \mathbf{x} + V_{g,k}^T \mathbf{h}\right)$$

or to follow a given text:

$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x} + V_{f,k} * \mathbf{y}\right) \odot \sigma\left(W_{g,k} * \mathbf{x} + V_{g,k} * \mathbf{y}\right)$$

where $\mathbf{y}$ are the (time-synchronous) linguistic features of the desired text (a sequence indicating the phonetic class of each output sample, the prosody, a.o.).

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Text-to-Speech - WaveNet (2016)

Speech samples	Subjective 5-scale MOS in naturalness	
	North American English	Mandarin Chinese
LSTM-RNN parametric	3.67 ± 0.098	3.79 ± 0.084
HMM-driven concatenative	3.86 ± 0.137	3.47 ± 0.108
WaveNet (L+F)	$\mathbf{4.21} \pm 0.081$	$\mathbf{4.08} \pm 0.085$
Natural (8-bit μ-law)	4.46 ± 0.067	4.25 ± 0.082
Natural (16-bit linear PCM)	4.55 ± 0.075	4.21 ± 0.071

Table 1: Subjective 5-scale mean opinion scores of speech samples from LSTM-RNN-based statistical parametric, HMM-driven unit selection concatenative, and proposed WaveNet-based speech synthesizers, 8-bit μ-law encoded natural speech, and 16-bit linear pulse-code modulation (PCM) natural speech. WaveNet improved the previous state of the art significantly, reducing the gap between natural speech and best previous model by more than 50%.

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Text-to-Speech - WaveNet (2016)

WaveNet is very slow to be used in production, so why is it important?

- It was the first neural net to synthesize waveform (**features-to-waveform**).
  - Most TTS systems have a decoder (text-to-spectrogram) followed by a vocoder (spectrogram-to-waveform).
  - Prior to WaveNet the vocoder was typically **Griffin-Lim**.
- New **non-autoregressive models arrived**, overcoming the issue of slow inference:
  - **Parallel WaveNet** and **ClariNet** use a **pretrained WaveNet** for knowledge **distillation**.
  - 20 times faster than real time.
- **Offline generation** is still **useful**:
  - WaveNet was used to generate **Google Assistant** voices for US English and Japanese across all Google platforms.
- WaveNet can also be **conditioned on spectrogram**, i.e. it become a **vocoder**.
  - Many state-of-the-art TTS systems use a WaveNet-like architecture as a vocoder.
- It is **multispeaker**:
  - You may train a single WaveNet (with utterances of many speakers) and generate speech from any of them.

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Text-to-Speech - Tacotron 2 (2017)

NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS

Jonathan Shen[1], *Ruoming Pang*[1], *Ron J. Weiss*[1], *Mike Schuster*[1], *Navdeep Jaitly*[1], *Zongheng Yang*[*2], *Zhifeng Chen*[1], *Yu Zhang*[1], *Yuxuan Wang*[1], *RJ Skerry-Ryan*[1], *Rif A. Saurous*[1], *Yannis Agiomyrgiannakis*[1], *and Yonghui Wu*[1]

[1]Google, Inc., [2]University of California, Berkeley,
{jonathanasdf,rpang,yonghui}@google.com

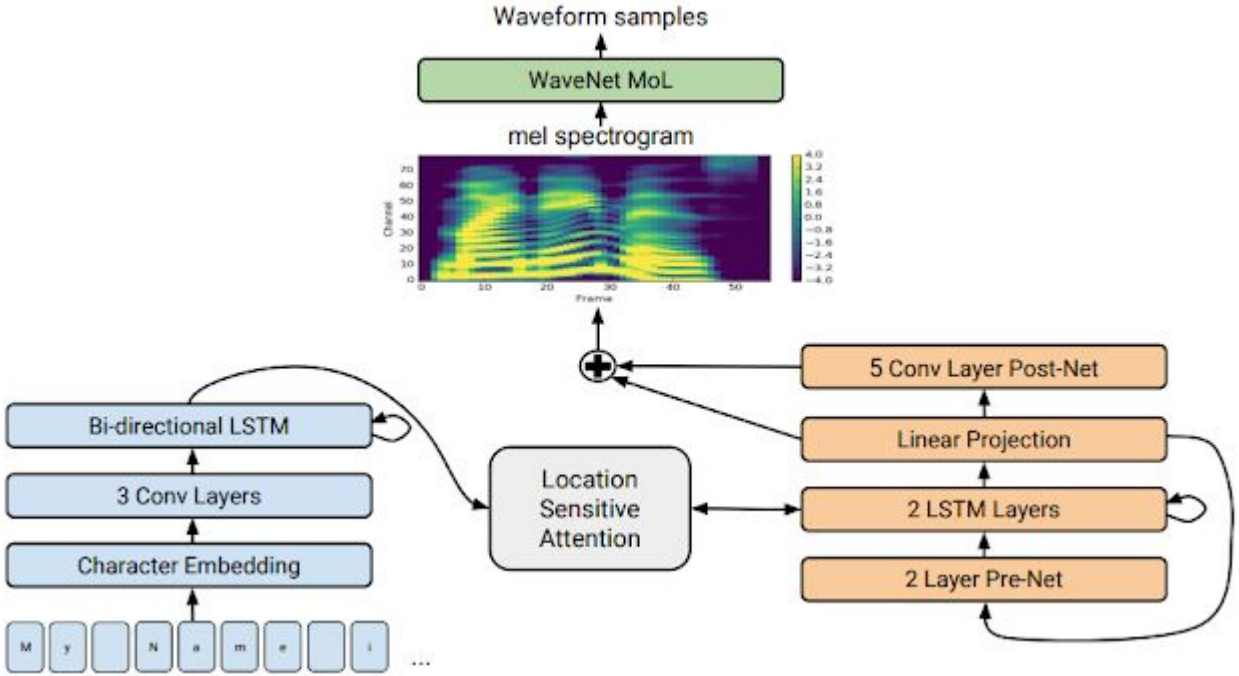
© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Text-to-Speech - Tacotron 2 (2017)

- It does not use complex linguistic and acoustic features as input.
- It comprises:
  - a recurrent **sequence-to-sequence** feature prediction network with attention which predicts a sequence of mel spectrogram frames from an input character sequence, and
  - a modified version of **WaveNet** which generates time-domain waveform samples conditioned on the predicted mel spectrogram frames.
- The alignment between the input character sequence and acoustic features is estimated by the **attention** mechanism.

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

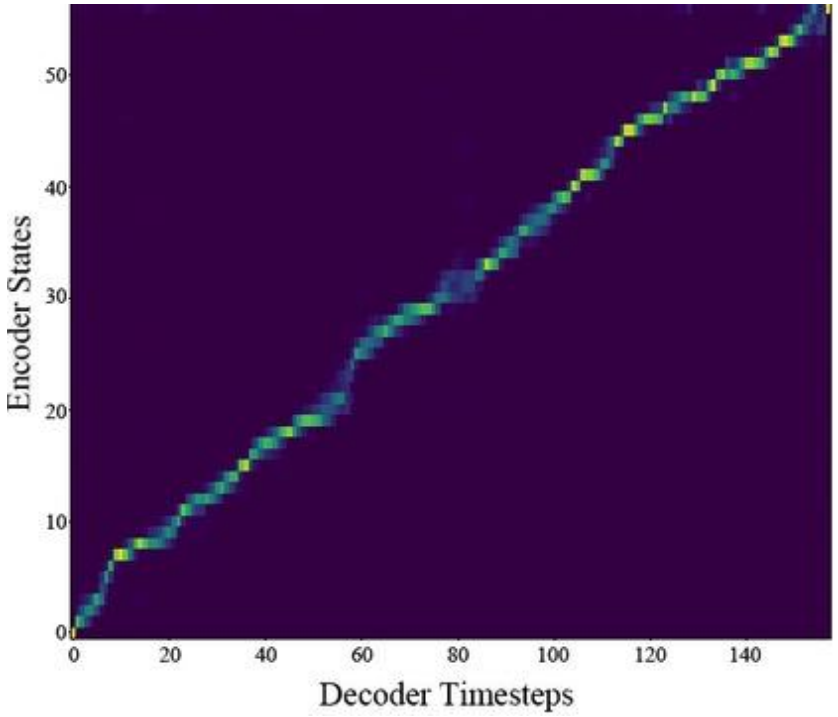
Text-to-Speech - Tacotron 2 (2017)



© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Text-to-Speech - Tacotron 2 (2017)

Alignment via attention



© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Text-to-Speech - Tacotron 2 (2017)

- It works well on out-of-domain and complex words.
- It learns pronunciations based on phrase semantics.
- It is somewhat robust to spelling errors.
- It is sensitive to punctuation.
- It learns stress and intonation.
- Its prosody changes when turning a statement into a question.
- It is good at tongue twisters.

Samples:
https://google.github.io/tacotron/publications/tacotron2/index.html

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Text-to-Speech - Zero-shot Multispeaker Tacotron 2 (2018)

Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis

Ye Jia* **Yu Zhang*** **Ron J. Weiss*** **Quan Wang** **Jonathan Shen** **Fei Ren**
Zhifeng Chen **Patrick Nguyen** **Ruoming Pang** **Ignacio Lopez Moreno** **Yonghui Wu**
Google Inc.
{jiaye,ngyuzh,ronw}@google.com

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Text-to-Speech - Zero-shot Multispeaker Tacotron 2 (2018)

It can mimic the voice of speakers unseen during training

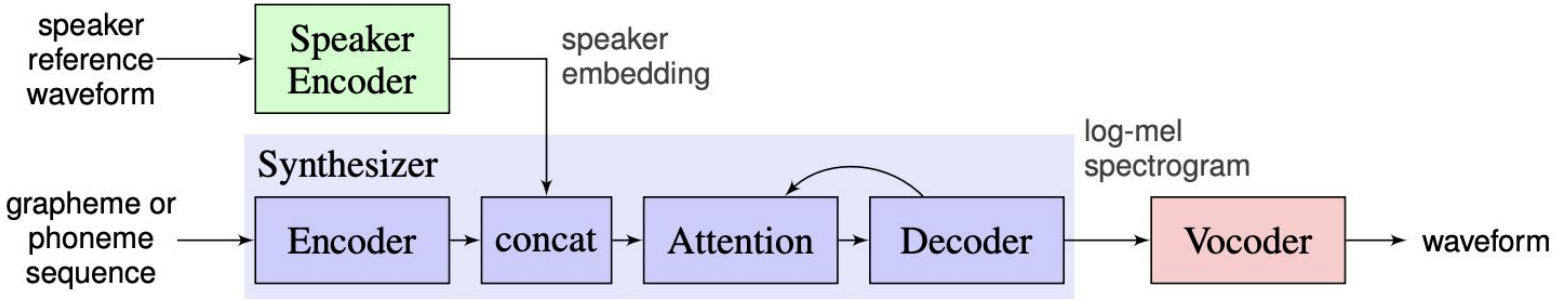


Figure 1: Model overview. Each of the three components are trained independently.

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Text-to-Speech - Zero-shot Multispeaker Tacotron 2 (2018)

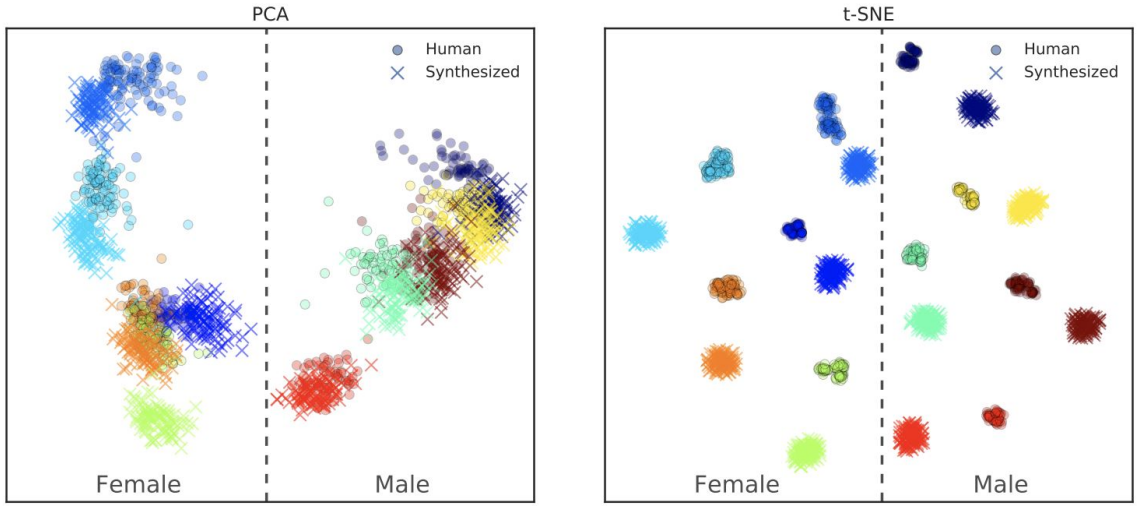


Figure 3: Visualization of speaker embeddings extracted from LibriSpeech utterances. Each color corresponds to a different speaker. Real and synthetic utterances appear nearby when they are from the same speaker, however real and synthetic utterances consistently form distinct clusters.

Samples: https://google.github.io/tacotron/publications/speaker_adaptation/

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Text-to-Speech - FastSpeech 1&2 (2018)

FastSpeech 2: Fast and High-Quality End-to-End Text to Speech

Yi Ren[1]*, Chenxu Hu[1]*, Xu Tan[2], Tao Qin[2], Sheng Zhao[3], Zhou Zhao[1]†, Tie-Yan Liu[2]

[1]Zhejiang University
{rayeren,chenxuhu,zhaozhou}@zju.edu.cn

[2]Microsoft Research Asia
{xuta,taoqin,tyliu}@microsoft.com

[3]Microsoft Azure Speech
Sheng.Zhao@microsoft.com

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Text-to-Speech - FastSpeech 1&2 (2018)

- **FastSpeech 1**
  - A **non-autoregressive** model trained using knowledge distillation from Tacotron 2.
  - As a non-autoregressive model it is very fast in inference.
  - However, **knowledge distillation** is slow in training (e.g. requires training a Tacotron 2).
  - The duration extracted from the teacher model is not accurate enough.
  - The target mel-spectrograms distilled from teacher model suffer from information loss due to data simplification.

- **FastSpeech 2**
  - **Trained** directly with **ground-truth** targets.
  - Introduces more variation information of speech (e.g., pitch, energy and more accurate duration) as conditional inputs.
  - A very popular network for real-time production systems.

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Text-to-Speech - FastSpeech 1&2 (2018)

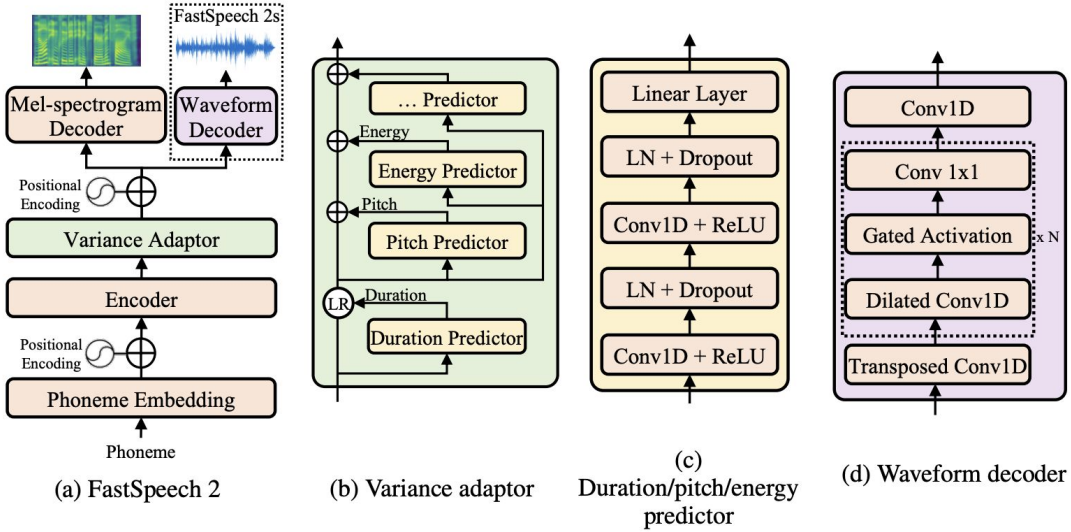


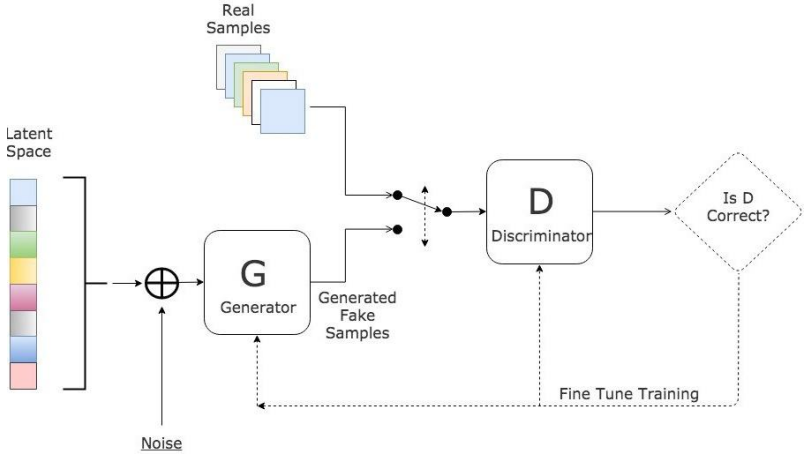
Figure 1: The overall architecture for FastSpeech 2 and 2s. LR in subfigure (b) denotes the length regulator proposed in FastSpeech. LN in subfigure (c) denotes layer normalization.

Samples: https://speechresearch.github.io/fastspeech2/

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Vocoders using Generative Adversarial Nets (GANs)

- K. Kumar, et al. “MelGAN: Generative adversarial networks for conditional waveform synthesis,” in NeurIPS, 2019.
- M. Binkowski, et al. “High fidelity speech synthesis with adversarial networks,” in ICLR, 2020
- J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” in NeurIPS, 2020.



Generative Adversarial Nets

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Text-to-Speech - VALL-E (2022)

Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers

Chengyi Wang* **Sanyuan Chen*** **Yu Wu*** **Ziqiang Zhang** **Long Zhou** **Shujie Liu**
Zhuo Chen **Yanqing Liu** **Huaming Wang** **Jinyu Li** **Lei He** **Sheng Zhao** **Furu Wei**
Microsoft
https://github.com/microsoft/unilm

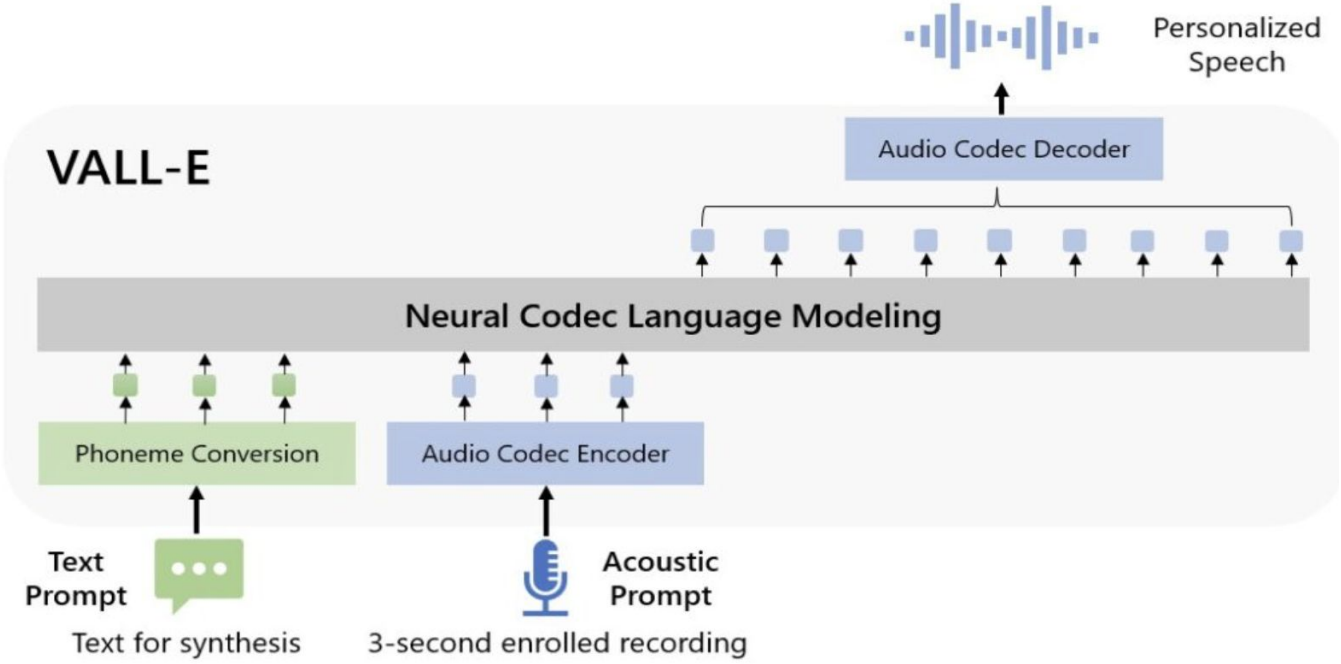
© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Text-to-Speech - VALL-E (2022)

- VALL-E is a TTS framework by **Microsoft**.
- It treats TTS as a language model task and uses **audio codec** codes as an intermediate representation instead of mel spectrograms.
- The framework leverages a large amount of **semi-supervised data** to build a generalized TTS system in the speaker dimension, highlighting the importance of scaling up semi-supervised data for TTS.
- It uses a **transformer decoder** (instead of recurrent of convolutional).
- VALL-E can produce diverse outputs for the same input text while maintaining the **acoustic environment** and the **speaker's emotion** from the acoustic prompt.
- In a zero-shot scenario, VALL-E demonstrates natural speech synthesis with **high speaker similarity** compared to other systems.

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Text-to-Speech - VALL-E (2022)

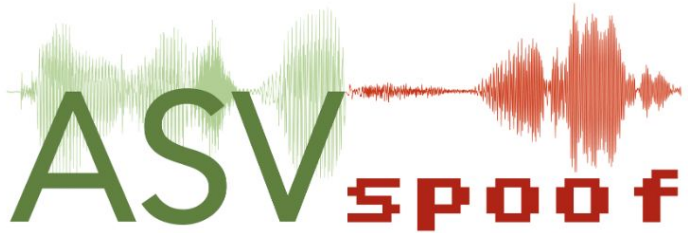


Samples: https://www.microsoft.com/en-us/research/project/vall-e/

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Can we detect synthetic speech?

- Doable with Deep Neural Networks trained to classify real vs synthetic speech
- Fairly easily for known and old TTS systems
- Harder for new systems which are unseen during training



asvspoof.org

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Generative AI for Dialog Systems
Large Language Models

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Large Language Models

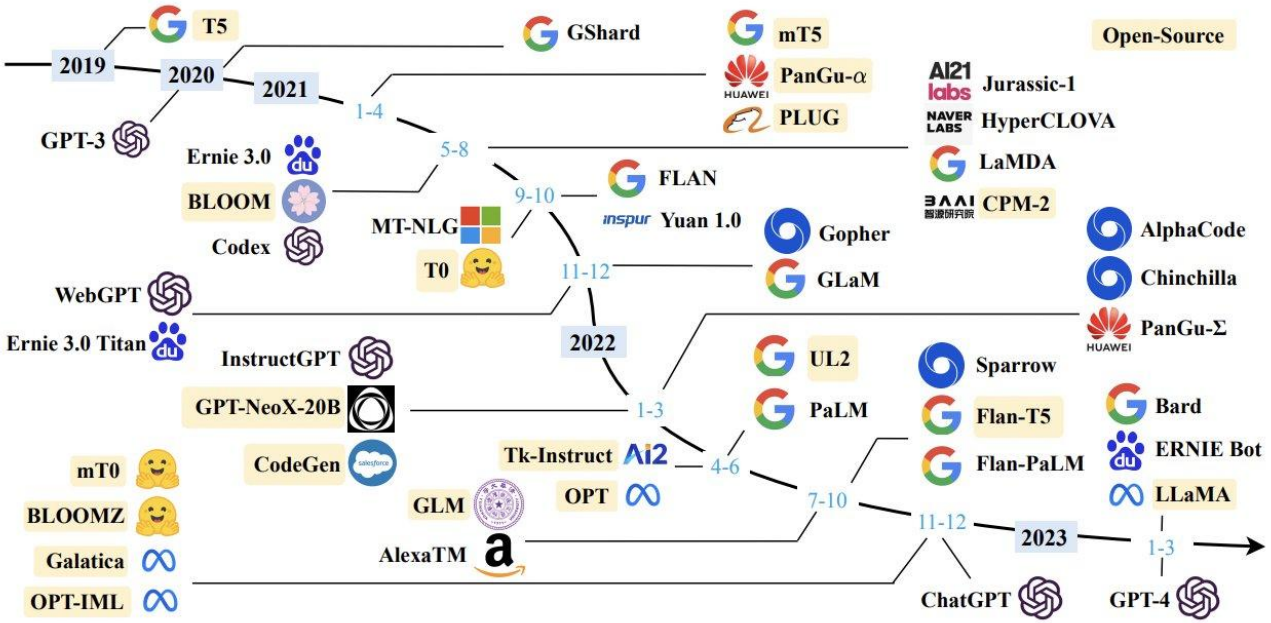


Fig. 1. A timeline of existing large language models (having a size larger than 10B) in recent years. We mark the open-source LLMs in yellow color.

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

LLMs for Task-Oriented Dialog Systems

Offline tasks

- **Assist dataset creation**
  - Suggest new sentences for intents or entities.
  - Suggest intent labels.

- **Assistance in dialogue testing**
  - Dialogue generation/completion.
  - Create “unhappy paths” of dialogues.

- **User simulator for automated testing**
  - Provide a pre-defined NLU Ontology (intents, entities, dialogue acts) and use LLM to interact with the system by mimicking a user.

- **Summarization**
  - Summarize a dialog (for archival purposes or to inform an agent)

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Suggest New Sentences

N

Give me more phrases like that: "I'm having problems in my basement i'd like to speak to a representative", "no lights electric's out", "i have a power failure"

- "I need help with a flooding issue in my basement"
- "I can't get any power in my house"
- "I'm experiencing an electrical outage"
- "I need to talk to someone about a problem with my electricity"
- "I'm having issues with my electrical system"
- "I'm having trouble with my electrical wiring"
- "I need assistance with a power outage"
- "I'm having issues with my electrical outlets"
- "I need help with a power failure in my home"

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Dialogue Generation

N

Agent: Thanks for calling ABC bank. How may I help you today?
Caller: pay bill
Agent: Would you like to pay today or schedule a future payment?
Caller: Today
Agent: Please tell me the amount you would like to pay.
Caller: one thousand four hundred eighty two dollars and thirty two cents
Agent: Got it, thanks. Please enter or say the payee's short number you'd like to pay. If you don't remember, I can list the registered payees for you.
Caller: one
Agent: To confirm: You have chosen to pay one thousand four hundred eighty-two dollars and thirty-two cents from your checking account, short number one effective today. Shall I proceed with the payment?"
Caller: yes
Agent: Your payment of one thousand four hundred eighty-two dollars and thirty-two cents has been successfully completed. Your confirmation number is zero nine eight eight.

Agent: You are a GPT-Agent, a state of art LLM. You are polite and kind. Generate a new dialog between a caller and an agent.

Agent: Thank you for calling ABC bank. How may I assist you today?

Caller: Hi, I'm calling to check on the status of a recent payment I made.

Agent: Sure, I'd be happy to help with that. Can you please provide the date and amount of the payment you made?

Caller: The payment was made yesterday for the amount of $500.

Agent: Let me look that up for you. Can you please confirm the last four digits of the account the payment was made from?

Caller: The last four digits are 1234.

Agent: Thank you. I've located the payment in our records and it looks like the payment was processed successfully. Is there anything else I can assist you with today?

Caller: No, that's all. Thank you for your help.

Agent: You're welcome. Thank you for choosing ABC bank for your banking needs. Have a great day!

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Dialogue Completion (1/2)

N

Agent: Thanks for calling the bank. How may I help you today?
Caller: pay bill
Agent: Would you like to pay today or schedule a future payment?
Caller: today
Agent: Please tell me the amount you would like to pay.
Caller: one thousand four hundred eighty two dollars and thirty two cents

Complete the above dialog

Agent: Great, I can process that payment for you today. Can you please confirm the account number or the biller name to which you would like to make this payment?
Caller: The account number is xxxx-xxx-xxx and the biller name is XYZ Company.
Agent: Thank you for the information. I have processed the payment of $1,482.32 to XYZ Company on account number xxxx-xxx-xxx. Is there anything else I can assist you with today?
Caller: No, that's all. Thank you.
Agent: You're welcome. Have a great day. Thank you for choosing our bank for your payment needs.

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Dialogue Completion (2/2) - “Unhappy path”

N

Agent: Thanks for calling the bank. How may I help you today?
Caller: pay bill
Agent: Would you like to pay today or schedule a future payment?
Caller: today
Agent: Please tell me the amount you would like to pay.
Caller: one thousand four hundred eighty two dollars and thirty two cents

Continue the above dialogue with the Caller disorientating the regular dialogue flow.

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Dialogue Completion (2/2) - “Unhappy path”

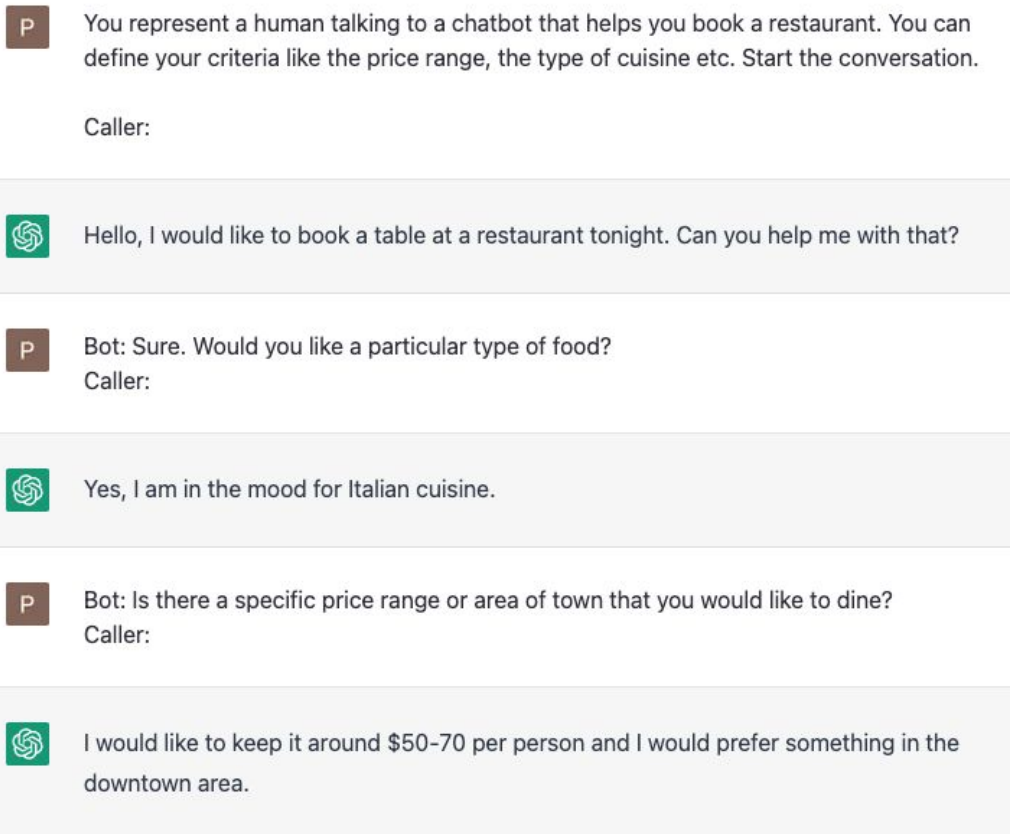
Agent: I understand, so you would like to make a payment of $1,482.32 today. Is that correct?
Caller: No, I meant I want to transfer the money to another account
Agent: I apologize for the confusion. To transfer money to another account, you will need to provide me with the account number and routing number of the receiving account. Are you able to provide that information?
Caller: Actually, scratch that. I just want to check my account balance
Agent: I understand, I can help you with that. Can you please provide me with your account number or verify your identity so that I can access your account information for you?
Caller: Never mind, I remember now that I have to pay the bill tomorrow. Can you just remind me tomorrow?
Agent: I apologize for any confusion. I can set up a reminder for you to pay your bill tomorrow. Would you like me to do that for you?
Caller: No, it's okay, I will just remember myself. Thank you
Agent: I understand, Is there anything else that I can help you with today?

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

User Simulator

Automated Testing

- Use LLMs to mimic a caller and converse with our system.
- User utterances are generated by LLMs using the dialogue history.
- We can make LLM aware of the domain (the supported NLU Ontology)
- We can ask the LLM to follow unhappy paths (e.g., unsupported intents, new intent before completion of the current and others).



© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

LLMs & Task-Oriented Dialog Systems

LLMs at runtime

- **Frequently-Asked Questions**
  - Retrieve parts relevant to the user query and copy them to the LLM context, together with the user's query.
  - The LLM can select the most relevant parts and rephrase the accordingly.

- **NLG**
  - Use an LLM to adapt the intent-specific system prompt to the dialog history:
    - "My father passed away yesterday and I need to close his account".
    - "I got married and I want to close my account to open a new one with my wife."
  - In both cases, the intent is "Close-Account", but the response should be different.
  -
- **NLU: LLM as intent classifiers**
  - Performs better than BERT especially in few-shot settings.
  - Narrow-down the intents by passing the user utterance to a weak-classifier (K-NN).
    - Keeping only the intents of the K neighbors and pass these candidates to the LLM.

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

LangChain

LangChain

- **Models:** Supported model types and integrations.
- **Prompts**: Prompt management, optimization, and serialization.
- **Memory:** Memory refers to state that is persisted between calls of a chain/agent.
- **Indexes:** Language models become much more powerful when combined with application-specific data - this module contains interfaces and integrations for loading, querying and updating external data.
- **Chains:** Chains are structured sequences of calls (to an LLM or to a different utility).
- **Agents:** An agent is a Chain in which an LLM, given a high-level directive and a set of tools, repeatedly decides an action, executes the action and observes the outcome until the high-level directive is complete.
- **Callbacks:** Callbacks let you log and stream the intermediate steps of any chain, making it easy to observe, debug, and evaluate the internals of an application.
- https://python.langchain.com/

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd

Thank you!
Happy to address your questions

tstafylakis@omilia.com

© 2020 Omilia Natural Language Solutions Ltd. This document is proprietary and CONFIDENTIAL. No part of this document may be disclosed in any manner to a third party without the prior written consent of Omilia Natural Language Solutions Ltd