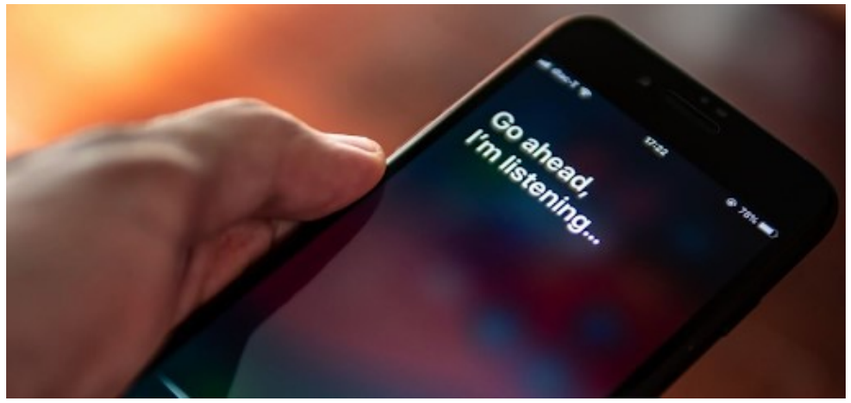
«Ok, Google: facciamo due chiacchiere?»



Cigna Gaia

Di Mauro Davide

Falcone Chiara

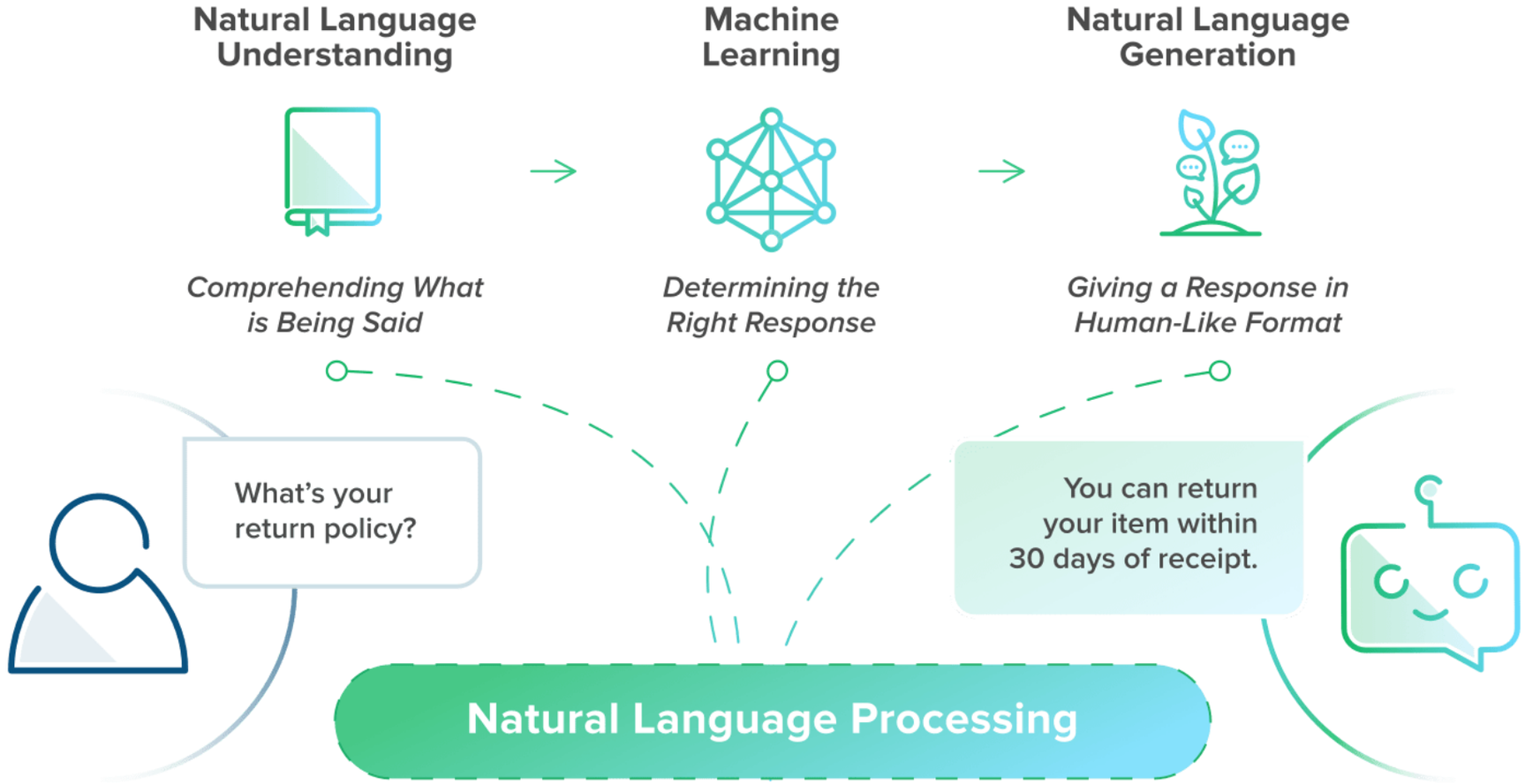
Progetti
Informatica Musicale 2020/21

Indice

- Elaborazione del linguaggio naturale e NLP
- Riconoscimento vocale
- Sintesi vocale
- Analisi del caso WaveNet
- Applicazioni

Elaborazione del linguaggio naturale e NLP

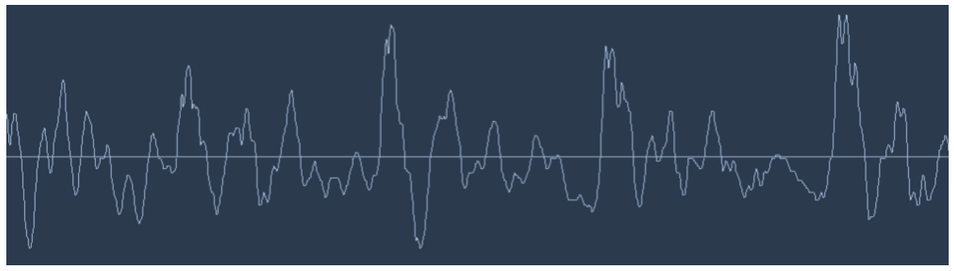
- Natural Language Processing: comprensione e generazione
- Comprensione: part-of-speech (POS)
- Generazione: la traduzione del linguaggio artificiale di un computer in testo e/o voce.



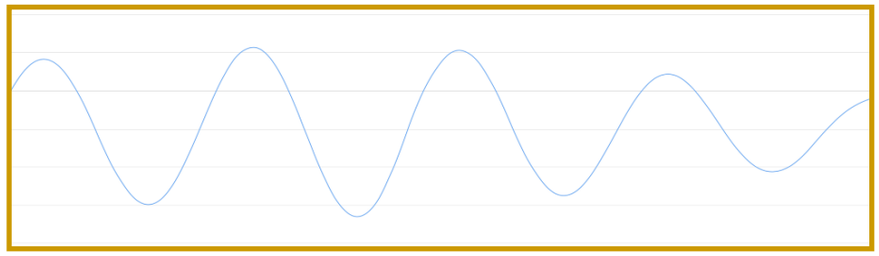
Riconoscimento vocale

- Deep Learning: word embeddings e reti convolutive ricorrenti
- La nostra voce come onda sonora e il campionamento

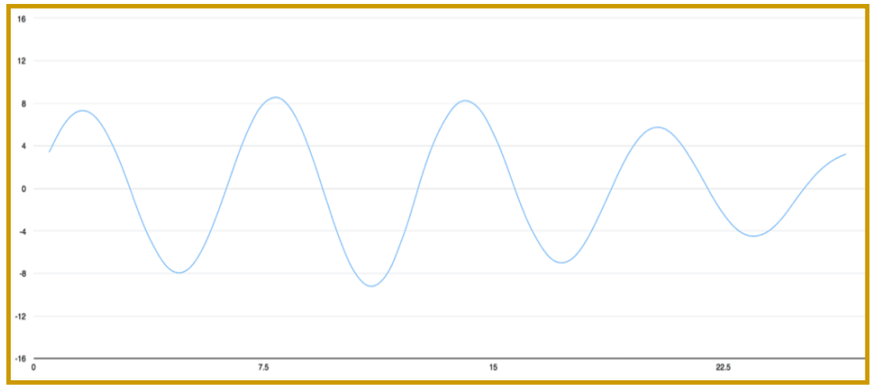
1. L'onda sonora generata dalla pronuncia della parola "ciao":



2. Lo zoom di una piccola parte dell'onda sonora:



3. Per trasformare questa onda sonora in numeri è sufficiente registrare l'altezza dell'onda per punti equidistanti, come vediamo di fianco:



4. Ecco i primi 100 numeri del campionamento dell'onda sonora della parola ciao pronunciata 16.000 volte al secondo:

```
[-1274, -1252, -1160, -986, -792, -692, -614, -429, -286, -134, -57, -41, -169, -456, -450, -541, -761, -1067, -1231, -1047, -952, -645, -489, -448
, -397, -212, 193, 114, -17, -110, 128, 261, 198, 390, 461, 772, 948, 1451, 1974, 2624, 3793, 4968, 5939, 6057, 6581, 7302, 7640, 7223, 6119, 5461,
 4820, 4353, 3611, 2740, 2004, 1349, 1178, 1085, 901, 301, -262, -499, -488, -707, -1406, -1997, -2377, -2494, -2605, -2675, -2627, -2500, -2148, -
1648, -970, -364, 13, 260, 494, 788, 1011, 938, 717, 507, 323, 324, 325, 350, 103, -113, 64, 176, 93, -249, -461, -606, -909, -1159, -1307, -1544]
```

5. Pre-processiamo questi numeri, li raggruppiamo in gruppi che rappresentino 20 millisecondi di audio ciascuno. Ed ecco i nostri primi 320 campioni:

```
[-1274, -1252, -1160, -986, -792, -692, -614, -429, -286, -134, -57, -41, -169, -456, -450, -541, -761, -1067, -1231, -1047, -952, -645, -489, -448
, -397, -212, 193, 114, -17, -110, 128, 261, 198, 390, 461, 772, 948, 1451, 1974, 2624, 3793, 4968, 5939, 6057, 6581, 7302, 7640, 7223, 6119, 5461,
 4820, 4353, 3611, 2740, 2004, 1349, 1178, 1085, 901, 301, -262, -499, -488, -707, -1406, -1997, -2377, -2494, -2605, -2675, -2627, -2500, -2148, -
1648, -970, -364, 13, 260, 494, 788, 1011, 938, 717, 507, 323, 324, 325, 350, 103, -113, 64, 176, 93, -249, -461, -606, -909, -1159, -1307, -1544,
-1815, -1725, -1341, -971, -959, -723, -261, 51, 210, 142, 152, -92, -345, -439, -529, -710, -907, -887, -693, -403, -180, -14, -12, 29, 89, -47, -
398, -896, -1262, -1610, -1862, -2021, -2077, -2105, -2023, -1697, -1360, -1150, -1148, -1091, -1013, -1018, -1126, -1255, -1270, -1266, -1174, -10
03, -707, -468, -300, -116, 92, 224, 72, -150, -336, -541, -820, -1178, -1289, -1345, -1385, -1365, -1223, -1004, -839, -734, -481, -396, -580, -52
7, -531, -376, -458, -581, -254, -277, 50, 331, 531, 641, 416, 697, 810, 812, 759, 739, 888, 1008, 1977, 3145, 4219, 4454, 4521, 5691, 6563, 6909,
6117, 5244, 4951, 4462, 4124, 3435, 2671, 1847, 1370, 1591, 1900, 1586, 713, 341, 462, 673, 60, -938, -1664, -2185, -2527, -2967, -3253, -3636, -38
59, -3723, -3134, -2380, -2032, -1831, -1457, -804, -241, -51, -113, -136, -122, -158, -147, -114, -181, -338, -266, 131, 418, 471, 651, 994, 1295,
 1267, 1197, 1291, 1110, 793, 514, 370, 174, -90, -139, 104, 334, 407, 524, 771, 1106, 1087, 878, 703, 591, 471, 91, -199, -357, -454, -561, -605,
-552, -512, -575, -669, -672, -763, -1022, -1435, -1791, -1999, -2242, -2563, -2853, -2893, -2740, -2625, -2556, -2385, -2138, -1936, -1803, -1649,
 -1495, -1460, -1446, -1345, -1177, -1088, -1072, -1003, -856, -719, -621, -585, -613, -634, -638, -636, -683, -819, -946, -1012, -964, -836, -762,
 -788]
```

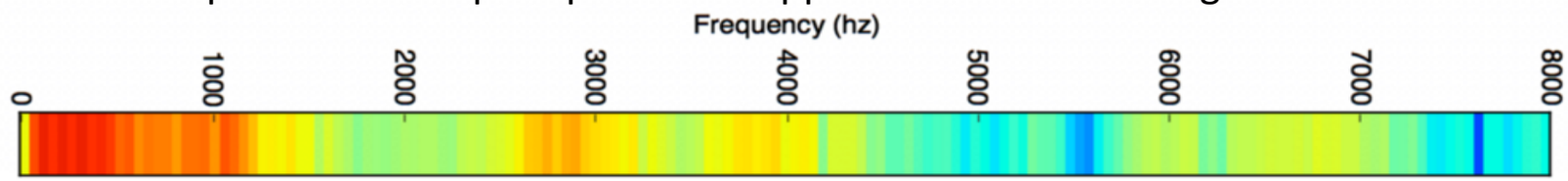
Riconoscimento vocale

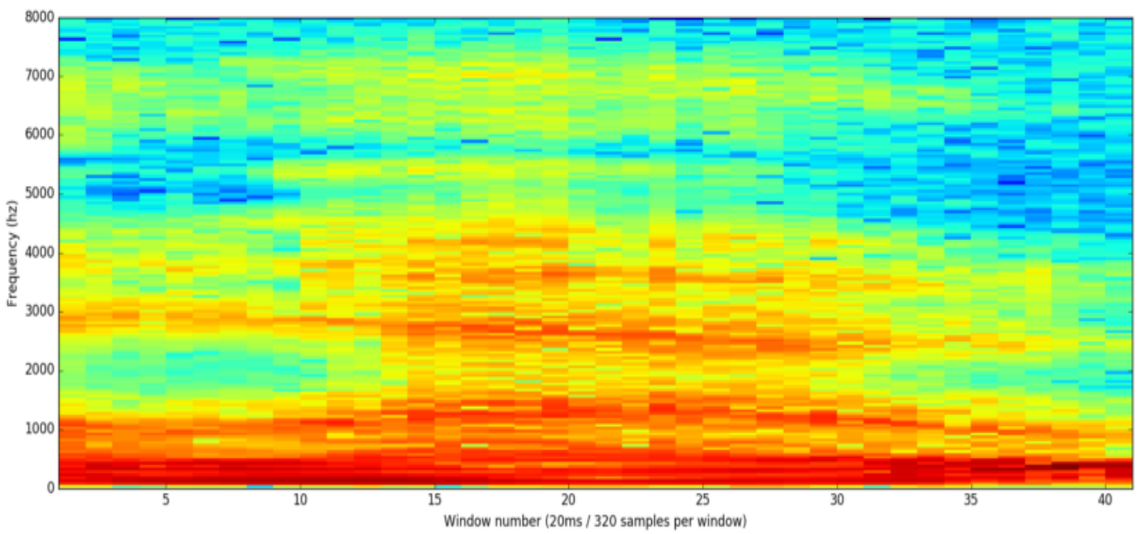
- L’impronta univoca della sequenza sonora con la trasformata di Fourier

Il risultato finale è un punteggio di quanto sia importante ogni gamma di frequenza, dai toni bassi (cioè note basse) a quelli più alti. Ogni numero che segue rappresenta la quantità di energia che è contenuta in ogni banda 50Hz della nostra clip audio di 20 millisecondi:


```
[110.97481594791122, 166.61537247955155, 180.43561044211469, 175.09309469913353, 180.0168691095916, 176.00619977472167, 179.79737781786582, 173.53025213548219, 176.87177119846058, 170.42684732853121
, 159.26023828556598, 163.24469810981628, 149.15527353931867, 154.34196586290136, 151.46179061113972, 152.99674239973979, 143.98878156117371, 156.6033737693738, 155.78237530428544, 157.17930941017838, 146.28632297509679, 164.37233032929228, 158.1282656446088, 147.23266451005145, 133.26597973863801, 116.51701000288331, 116.85501120577126, 115.40519005123537, 120.85619013711488, 112.48406123161091, 111.80244759457571, 92.590676871856431, 105.75863927434719, 95.673146446282971, 90.391748128064208, 79.355818055314899, 86.080143147713926, 84.748200268709567, 83.050569583779065, 86.207180262242758, 90.252031938154076, 89.361567351948437, 90.917307309643206, 90.746777849123049, 86.726552726337033, 85.709412745066928, 95.938840816664865, 99.092545759170 69, 96.632437741434885, 103.23961231666669, 105.80328302591124, 109.53029281234707, 116.46408227060996, 129.20890691592615, 130.43460361780441, 138.15581799444712, 128.25056761852832, 138.14492240466387, 140.0352714810314, 128.15138139429752, 123.93018478493934, 121.19289035588113, 119.03159255422509, 114.23027889344033, 119.17173421549 97, 101.02560719093093, 110.91192243698025, 106.04872005953503, 100.86977927980999, 92.123301579000341, 94.376766266598295, 97.850709698634489, 113.37126364077845, 110.24526597732718, 113.72249347908021, 120.63960942628063, 122.06482553759932, 117.96716716036715, 120.87682744817975, 125.06097381947157, 111.57319012901624, 115.54483708595507, 116.99850750130265, 114.40659619324526, 79.869543980883975, 104.83111191845597, 104.66218602004588, 104.91691734582642, 97.143620527536072, 78.43459781117835, 82.214144782667248, 67.246072805959614, 66.578937262360313, 74.100307226086798, 64.861423011415653, 59.167561212002269, 62.479712687304911, 63.568362396107467, 55.906096471453267, 42.790802900362839, 55.693923524361097, 50.776364877715011, 41.196111220671298, 51.062413666348945, 58.493563858289065, 53.081835042922769, 73.060663128159547, 68.216252021223 61, 66.7701034934517, 59.766251249152 02, 35.413635503802389, 22.705615809958832, 16.458048045346381, 44.910670465379937, 59.282513769840705, 69.241393677323856, 81.778634874076346, 88.409923803546008, 94.688033733251245, 96.640867526244051, 91.806226496828543, 94.570526932206619, 99.250924315589074, 97.899164767741183, 75.176507616277235, 80.947474423758905, 71.859103451990862, 93.863684037461738, 96.757146539348298, 96.528614354976241, 99.366456533638413, 102.18717608176904, 102.06596663023235, 101.78493139911082, 103.7883358299547, 99.915220403870748, 107.43478470929935, 104.46449552620618, 105.70789868195298, 101.10596541338749, 100.75737831526195, 91.742897073196886, 88.307278943069033, 90.936627732905492, 71.134275744339803, 72.504304977841457, 76.233185506299705, 63.281284410272761, 45.380164336858961, 43.018963766250437, 49.133789791276826, 53.507751009532953, 48.586423555688746, -4.4730776113028883, 50.833000650183408, 51.003802143009629, 39.577356593427531, 47.096919248906332, 55.442197175664383, 56.967128095484341, 49.383247263177985]
```


Ma è molto più facile da capire quando lo rappresentiamo in modo grafico:

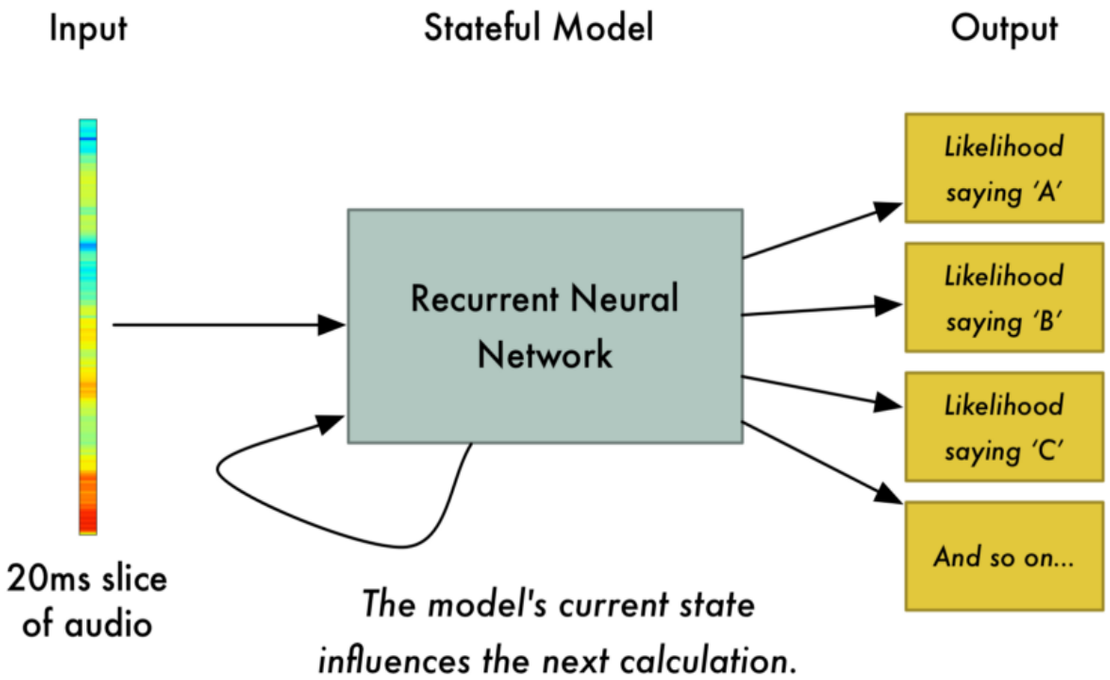




Se ripetiamo l’operazione su tutti i blocchi da 20 millisecondi del nostro file audio, otteniamo uno spettrogramma , come vediamo qui a sinistra.

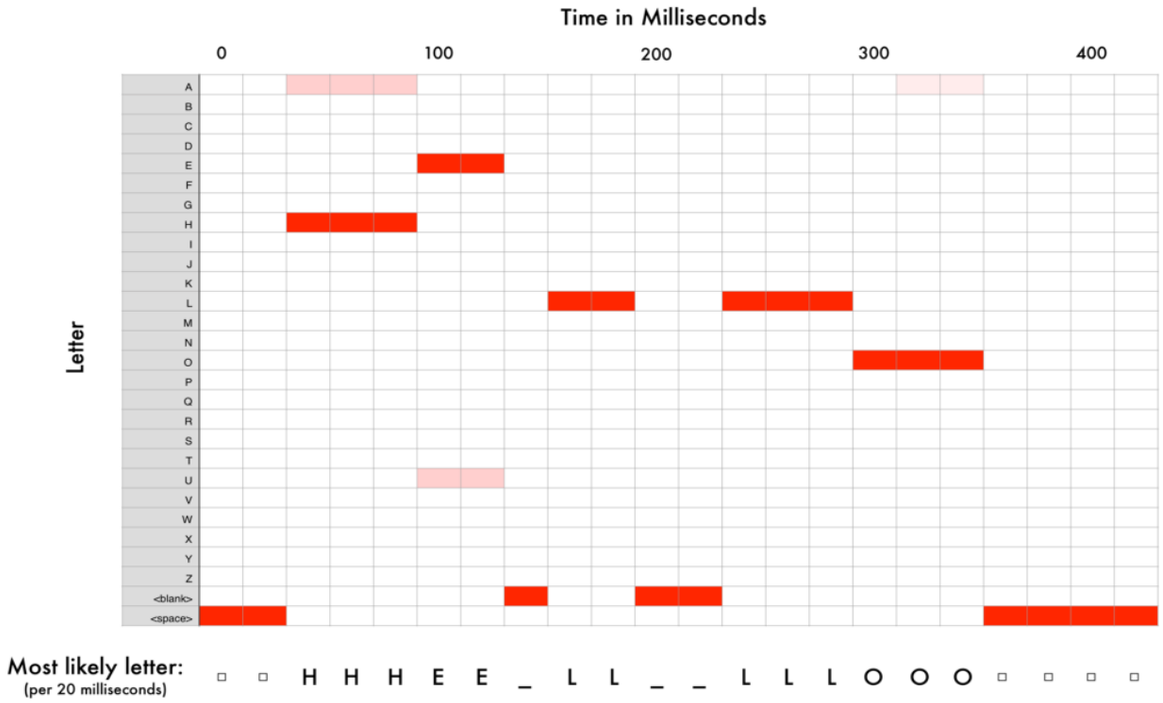
Riconoscimento vocale

- Rete neurale ricorrente



Useremo una Rete Neurale Ricorrente, cioè una Rete Neurale i cui risultati sono influenzati dai risultati precedenti. Questo perché ogni lettera individuata influisce sulle possibilità che una specifica lettera venga dopo.

- Dal dataset di parole al risultato finale

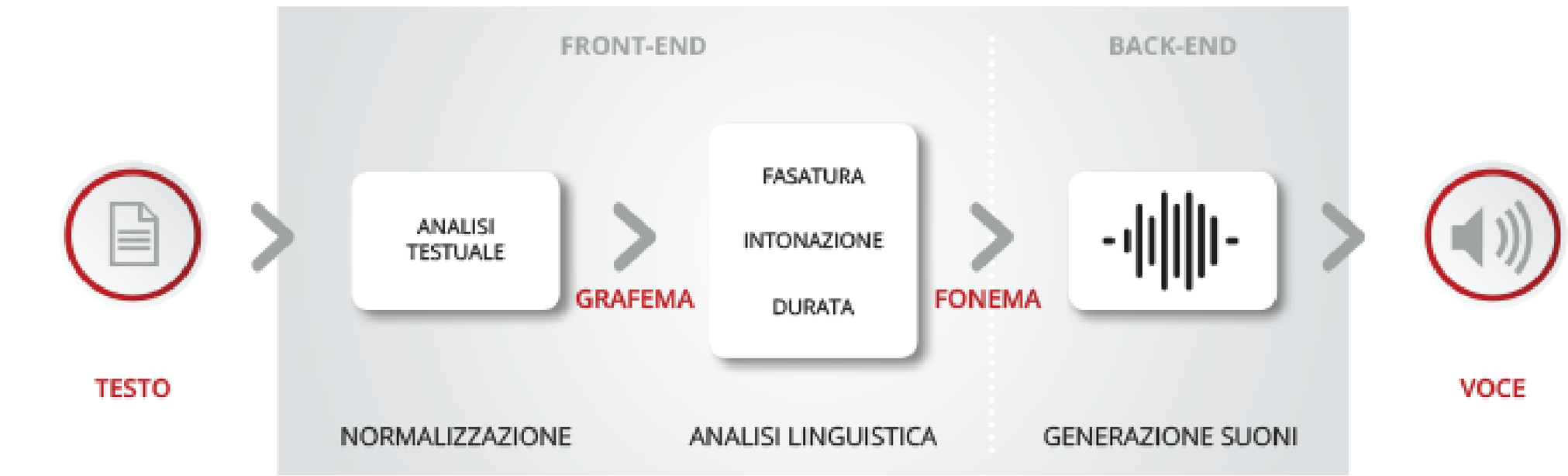


In primo luogo, sostituiremo tutte le lettere ripetute eccessivamente con lettere singole. Poi andremo a rimuovere eventuali spazi vuoti. Questo ci lascia con alcune possibili trascrizioni che possono tutte somigliare alla parola pronunciata. Ma la macchina deve capire quale delle possibili trascrizioni è quella che abbiamo effettivamente pronunciato.

Sintesi Vocale

Per sintesi vocale si intende la riproduzione artificiale della voce umana.

Essa è costituita da due fasi: front-end e back-end



Sintesi vocale

La sintesi vocale può avvenire grazie a diverse tecnologie, per esempio:

Sintesi concatenante

Essa viene utilizzata nella sintesi vocale per generare sequenze di suoni specificate dall'utente da un database costruito dalle registrazioni di altre sequenze

Sintesi formante

Consiste nella generazione di forme d'onda di cui si modulano alcuni parametri acustici come la frequenza fondamentale, i toni e i livelli di rumore

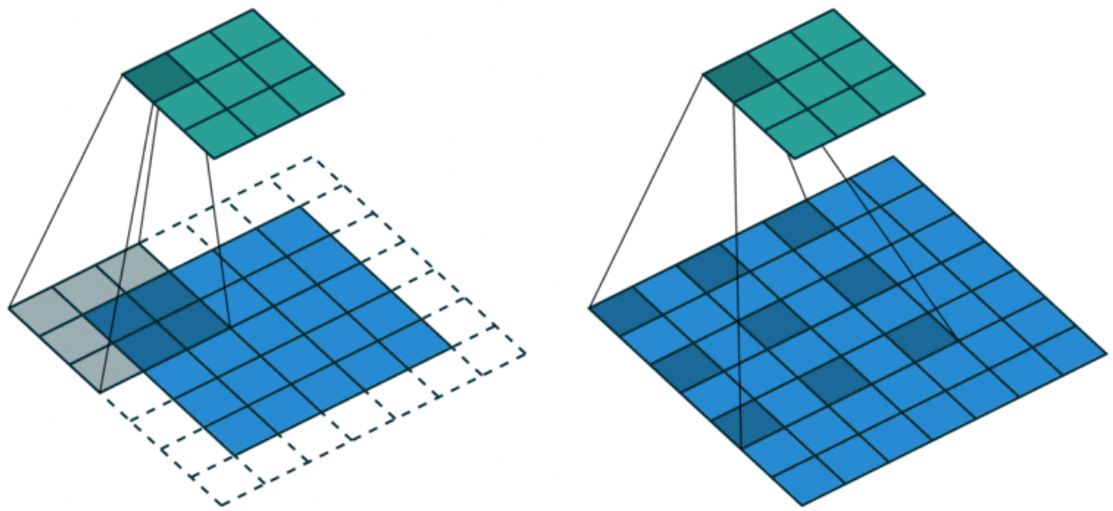
Sintesi basata sull'apprendimento profondo

La sintesi basata sull'apprendimento profondo avviene attraverso l'utilizzo delle reti neurali. Ad utilizzare quest'ultimo tipo di sintesi è **WaveNet**.

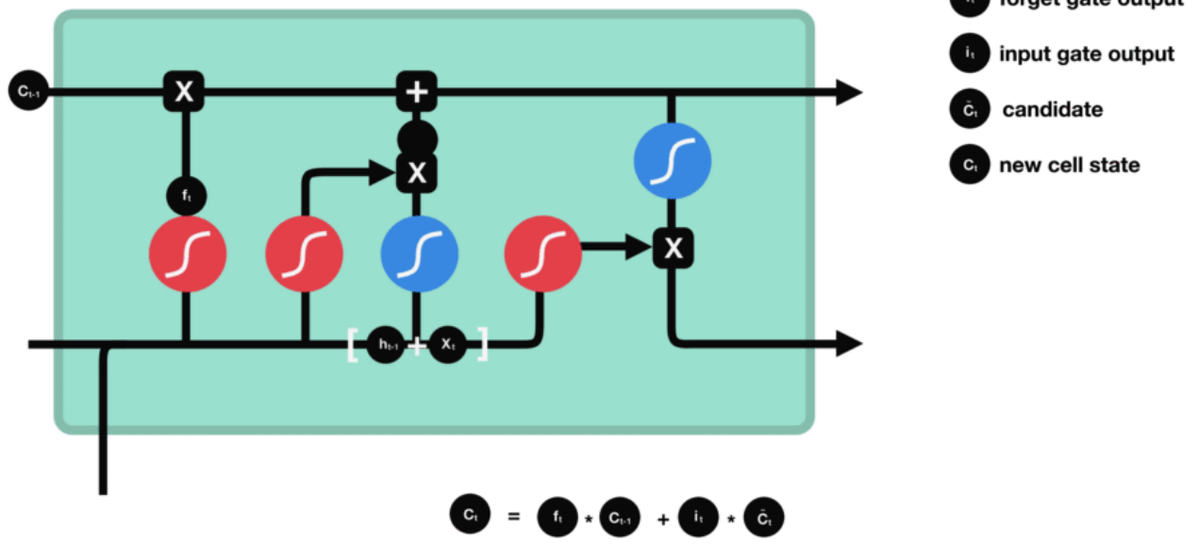
Analisi del caso WaveNet

Nel cuore di WaveNet troviamo due importanti processi:

- Convoluzioni Dilatate

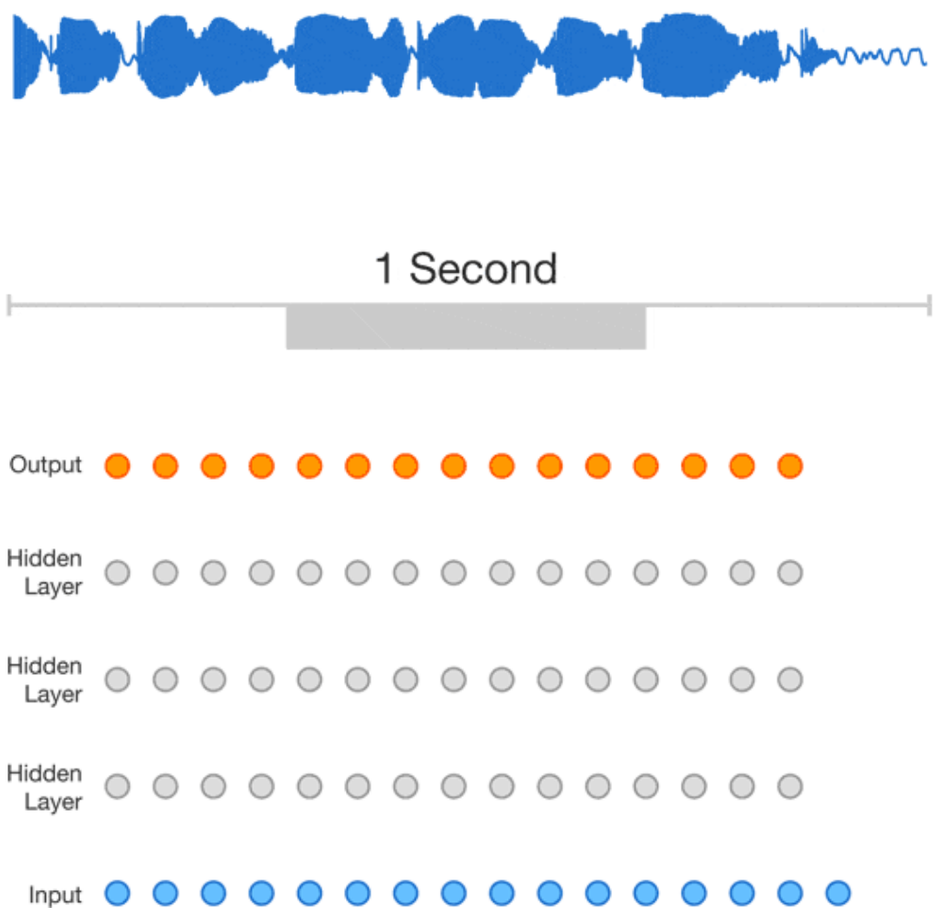


- Long-Short Term Memory



Analisi del caso WaveNet

- Produzione dei campioni



Si generano 16.000 campioni al secondo con 256 possibili valori l'uno grazie all'uso della μ-law. L'utilizzo delle convoluzioni dilatate permette di avere un input più grande senza aumentare il numero di livelli e di conseguenza il costo computazionale. L'input fornito è lo spettrogramma prodotto dall'apparato di Text-To-Speech che può essere unito a variabili globali o locali per inserire particolarità della voce e/o della lingua parlata. Poiché si fa utilizzo di una LSTM possiamo inserire anche valori ottenuti precedentemente come input per ottenere un output più preciso. Infine, seguendo una distribuzione probabilistica, il singolo campione ogni ciclo assume uno di quei 256 possibili valori

Applicazioni

Fonte: The Verge

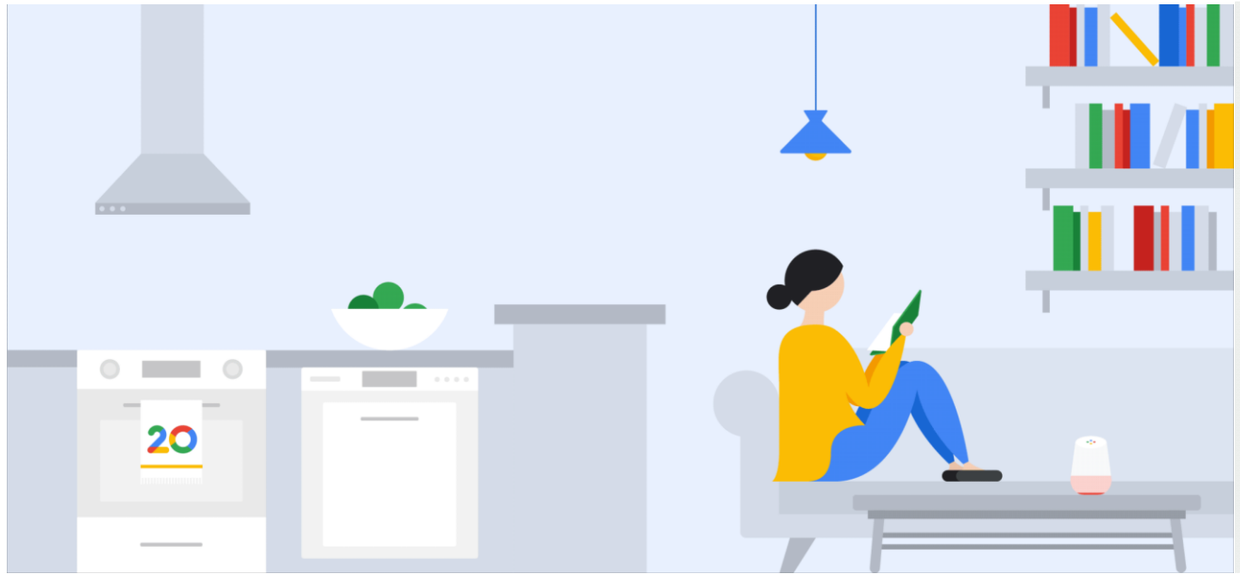
GOOGLE \ TECH \ ARTIFICIAL INTELLIGENCE

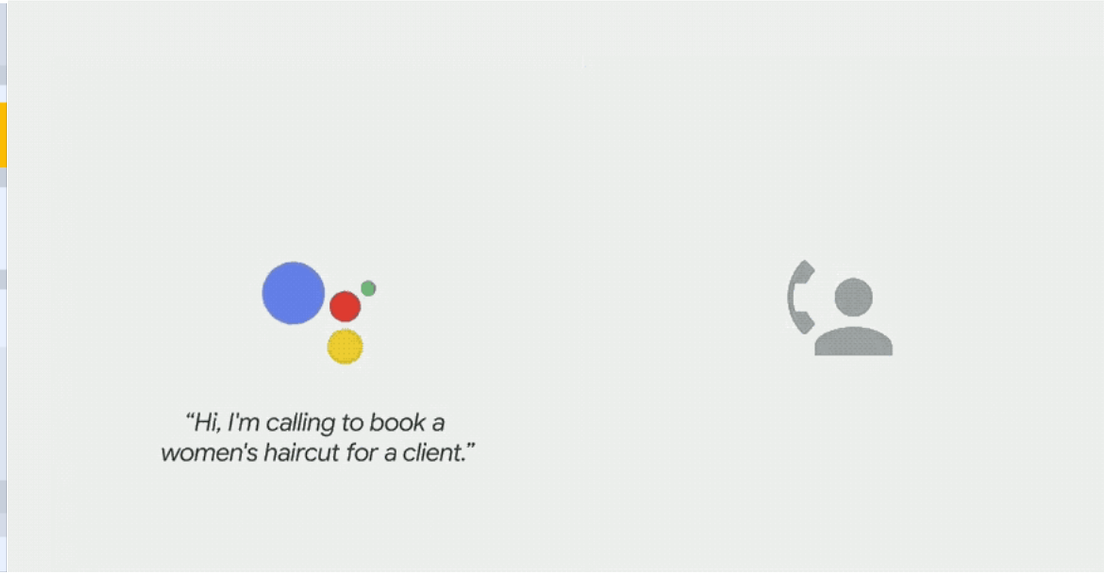
Google Assistant will alert people that Duplex calls are being recorded

The convincingly human-sounding AI technology will identify itself as Google Assistant when calls begin

By Chris Welch | @chriswelch | May 18, 2018, 6:02pm EDT

Tra le applicazioni più importanti di WaveNet vi sono sicuramente Google Assistant ed il servizio di segreteria Google Duplex





Conclusioni

- Come abbiamo visto, l'interfaccia uomo-macchina inizia a diventare sempre più diretta e di facile utilizzo, se non immediato, grazie all'utilizzo di sistemi e processi innovativi. Se da un lato ciò comporti stupore e meraviglia dall'altro l'intera tecnologia ha raggiunto livelli raccapricianti. *«Ok google! Fai partire il prossimo episodio di Black Mirror!»* **«Con piacere!»**

GRAZIE PER L'ATTENZIONE