AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Omówienie różnych metod rozpoznawania mowy

Na podstawie artykułu: „Comparative study of automatic speech recognition techniques”

Beniamin Sawicki
Wydział Inżynierii Mechanicznej i Robotyki
Inżynieria Akustyczna
Kraków, 12.01.2015

Mowa jako najbardziej intuicyjny kontroler najbliższego otoczenia człowieka

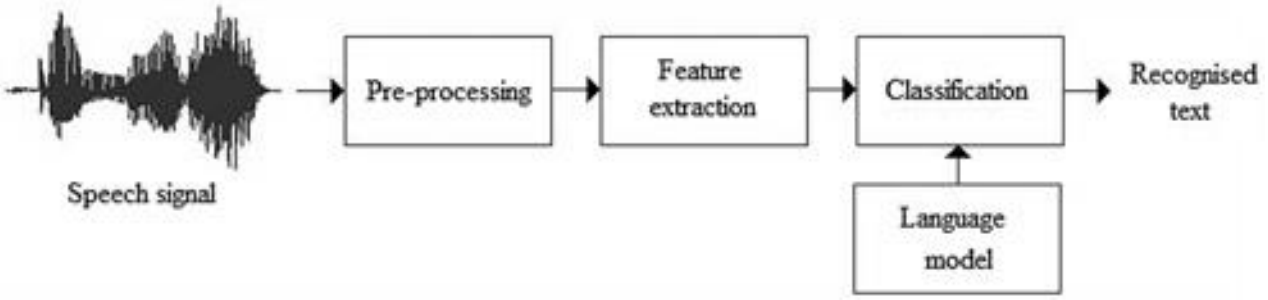
- Rozpoznanie i przetworzenie mowy ciągłej w czasie rzeczywistym
- Odporność na wpływ otoczenia i wszelkich zakłóceń
- Rozróżnianie konkretnych mówców i ich emocji

Problemy:

- Powtarzalność mowy
- Brak wyraźnych granic pomiędzy fonemami i słowami
- Hałas
- Płeć, styl i prędkość mówienia, dialekty

Proces rozpoznawania mowy

- Stacjonarność sygnału generowanego przez trakt głosowy w czasie 10-20ms
- Fonemy, formowane w słowa i zdania
- Zbiór fonemów, charakterystyczny dla danego języka
- Sygnał mowy jako elektryczna reprezentacja fali akustycznej



Etap ekstrakcji cech sygnału

Mel-frequency cepstral coefficients

- Skala melowa, określająca subiektywny odbiór wysokości dźwięku przez ludzkie ucho względem skali w hercach

$$F_{mel} = 1127 log_e \left(1 + \frac{f}{700}\right)$$

- Podział sygnału na ramki o długości 25-30ms z zakładką 10ms
- Zastosowanie okna Hamminga na każdej z ramek
- Transformata Fouriera na każdej z ramek
- Filtracja danych bankiem filtrów i obliczenie logarytmu energii
- Bezpośrednia transformata kosinusowa (DCT), której wynikiem są współczynniki MFC

Etap ekstrakcji cech sygnału

Mel-frequency cepstral coefficients

- Mając na uwadze koartykulację fonemów, trwającą dłużej niż ramka (30ms), analizuje się również korelacje czasowe pomiędzy ramkami.

- Wektor cech MFCC zawiera:
  - Cechy statyczne – analiza poszczególnych ramek
  - Cechy dynamiczne – różnice między cechami statycznymi kolejnych ramek
  - Cechy przyspieszenia – różnice między cechami dynamicznymi
  - Znormalizowany logarytm energii*

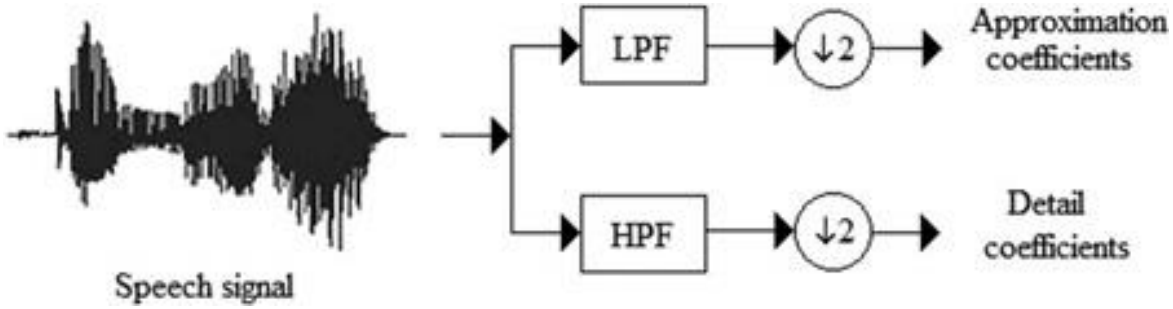
Etap ekstrakcji cech sygnału

Dyskretna transformata falkowa (DWT)

- Uzyskiwanie informacji o przebiegu czasowym sygnału niestacjonarnego

- Model adekwatny dla ucha ludzkiego:
  - Wąskie okno czasowe stosowane dla wysokich częstotliwości
  - Szerokie okno czasowe stosowane dla niskich częstotliwości

- Podział sygnału: aproksymacja (LF) i detale (HF)



Etap ekstrakcji cech sygnału

Dyskretna transformata falkowa (DWT)

- Usunięcie składowych wysokich częstotliwości – niska częstotliwość sygnału mowy
- W porównaniu do MFCC, dobra rozdzielczość częstotliwościowa dla niskich częstotliwości; lepsza lokalizacja zjawisk przejściowych w dziedzinie czasu
- Stosowanie falek ortogonalnych – Daubechies, Haar'a oraz różnej ilości poziomów dekompozycji
- *Wavelet packet transform* – dalsza dekompozycja

Etap ekstrakcji cech sygnału

Liniowe kodowanie predykcyjne (LPC)

- Analiza w dziedzinie czasu, odwzorowująca rezonansową strukturę traktu głosowego
- Ramkowanie, okienkowanie i autokorelacja między ramkami sygnału wejściowego
- Aproksymacja każdej kolejnej próbki jako liniowa kombinacja N poprzednich próbek

$$\hat{s}[n] = \sum_{k=1}^{P} a_k s(n-k)$$

- Używane w kombinacji z DWT, rozwinięcie metody: *Linear Predictive Cepstral Coefficients (LPCC)*

Etap ekstrakcji cech sygnału

Percepcyjna predykcja liniowa (PLP)

- Metoda PLP oparta na trzech charakterystykach:
  - Rozdzielczość spektralna pasma krytycznego
  - Regulacja krzywej jednakowej głośności
  - Zastosowanie *intensity-loudness power law*

- Transformata Fouriera na okienkowanej ramce sygnału

- Filtracja skalą Barka (1 bark = 100 melów)

Skala Barka obejmuje cały zakres częstotliwości z obszaru 24 pasm krytycznych, w których odbiór jednego dźwięku zależy od obecności innego dźwięku.

Etap ekstrakcji cech sygnału

Percepcyjna predykcja liniowa (PLP)

- Po filtracji, sygnał ważony jest krzywą jednakowej głośności
- Kompresja sygnału przy użyciu *intensity-loudness power law*
- Odwrotna transformata Fouriera => analiza predykcji liniowej => analiza cepstralna
- Wyższa skuteczność algorytmu niż LPCC w przypadku środowiska o dużych zakłóceniach

Etap ekstrakcji cech sygnału

RASTA-PLP

RelAtive SpecTrA-preceptual linear prediction

- Scalenie techniki RASTA z metodą PLP, aby zmniejszyć podatność metody PLP na zakłócenia
- Czasowe właściwości przebiegu zakłóceń są odróżnialne od właściwości sygnału mowy
- Filtracja pasmowo-przepustowa energii obecnej w podpasmach wygładza zakłócenia
- Najskuteczniejsza metoda w przypadku mocno zniekształconych sygnałów

Etap ekstrakcji cech sygnału

Algorytmy dodatkowe:

- Vector quantisation
- Principal component analysis
- Linear discriminant analysis

Etap klasyfikacji uzyskanych segmentów mowy

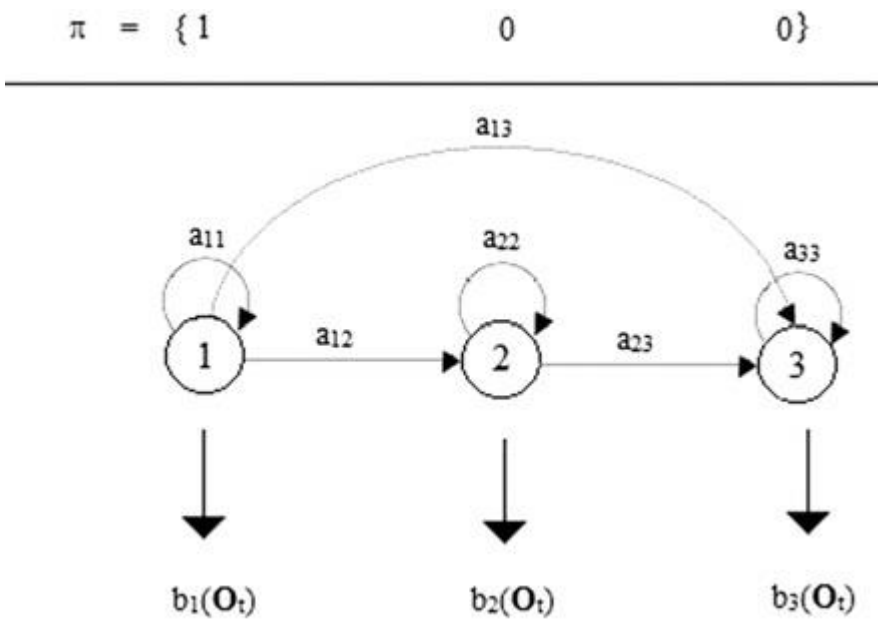
Hidden Markov models

- Prawdopodobieństwo wygenerowania wypowiedzi poprzez wymowę konkretnego fonemu lub słowa

- Możliwe zmiany stanu, a_{ij}
- Możliwe obserwacje, reprezentujące możliwy dźwięk, powstały podczas każdego ze stanów, $b_j(\boldsymbol{O}_t)$
- Dystrybucja prawdopodobieństwa stanu początkowego π

$$\lambda = (A, b, \pi)$$

A=$\{a_{ij}\}$, B=$\{b_j(\boldsymbol{O}_t)\}$, 1≤i, j≤N,1≤k≤M
N – liczba stanów
M – liczba obserwacji



First-order three-state left-to-right HMM

Etap klasyfikacji uzyskanych segmentów mowy

Hidden Markov models

- Ocena prawdopodobieństwa sekwencji wypowiedzi dla danego HMM
- Wybór najlepszej sekwencji modeli stanów
- Modyfikacja odpowiednich parametrów modeli dla lepszej reprezentacji wypowiedzi

Słowa zawsze bazują na powiązaniach konkretnych fonemów. Stąd, dobre rozpoznanie konkretnych fonemów wiąże się z dobrym rozpoznaniem słów.

Etap klasyfikacji uzyskanych segmentów mowy

Sztuczne sieci neuronowe

- Rozpoznawanie wzorców
- Faza treningu, umożliwiająca naukę systemu
- Zdolność do klasyfikacji nowych, nieznanych danych
- Niezdolne do uwydatnienia zmienności w czasie sygnału mowy
- Stosowane najczęściej w hybrydzie z HMM

Etap klasyfikacji uzyskanych segmentów mowy

Sztuczne sieci neuronowe

Perceptrony wielowarstwowe

- Sieci złożone z co najmniej trzech warstw: wejściowa, ukryta i wejściowa. Rezultat klasyfikacji odnosi się do neuronu o najwyższym uaktywnieniu.

Samoorganizujące mapy

- Grupowanie danych w topograficzne mapy, od wielkowymiarowych przestrzeni wejściowych po niskowymiarowe przestrzenie cech

SOM posiadają zdolność do rozróżniania głównych cech wprowadzonych do sieci danych, dzięki procesowi nauki.

Etap klasyfikacji uzyskanych segmentów mowy

Sztuczne sieci neuronowe

Radial basis functions

- Sieci złożone z trzech warstw: wejściowa, ukryta i wejściowa
- Tworzenie klastrów opartych na wprowadzonych wzorcach
- Funkcja Gaussa stosowana do obliczenia powiązania danych wejściowych z utworzonymi klastrami

Rekurencyjne sieci neuronowe

- Sieci złożone z trzech warstw: wejściowa, ukryta i wejściowa
- Wyniki z odpowiednich węzłów są mnożone przez odpowiadające wagi i podane z powrotem do węzła

Fuzzy neural network

- Powiązanie rozmytych systemów z sieciami neuronowymi
- Element jest powiązany w sieci z odpowiednim stopniem członkostwa, dzięki funkcji członkostwa

Analiza modeli językowych

Poziomy analizy lingwistycznej języka:

1. Fonologia – brzmienie, różnice w wymowie
2. Morfologia – znaczenie składowych słowa
3. Poziom leksykalny – interpretacja pojedynczych słów
4. Poziom syntaktyczny – analiza słów w kontekście zdania
5. Poziom semantyczny – znaczenia zdań
6. Rozmowa – znaczenie całego tekstu
7. Poziom pragmatyczny – analiza intencji, planów, celów. Analiza tematu wypowiedzi podczas interpretacji wieloznaczeniowego słowa

Analiza modeli językowych

- Implementacja poziomów w ASR poprzez zastosowanie *Natural language processing*
  - Przetworzenie sygnału mowy w sekwencję fonemów; próba zrozumienia słów przez NLP
  - ASR zwraca więcej niż jedno rozpoznanie danego słowa. Przy pomocy NLP można wybrać pasujące najlepiej do kontekstu
  - Kombinacja NLP i ASR

- Należy brać pod uwagę nieprzestrzeganie zasad gramatyki, dialekty, styl mówienia

Analiza modeli językowych

- Implementacja poziomów w ASR poprzez zastosowanie *Natural language processing*
  - Przetworzenie sygnału mowy w sekwencję fonemów; próba zrozumienia słów przez NLP
  - ASR zwraca więcej niż jedno rozpoznanie danego słowa. Przy pomocy NLP można wybrać pasujące najlepiej do kontekstu
  - Kombinacja NLP i ASR

- Należy brać pod uwagę nieprzestrzeganie zasad gramatyki, dialekty, styl mówienia

Porównanie wybranych algorytmów

Feature extraction technique	Advantages	Disadvantages
MFCC	• provides good discrimination • low correlation between coefficients • not based on linear characteristics; hence, similar to the human auditory perception system • important phonetic characteristics can be captured	• low robustness to noise • in a continuous speech environment, a frame may not contain information of only one phoneme, but of two consecutive phonemes • limited representation of speech signals since only the power spectrum is considered, ignoring the phase spectrum of speech signals
DWT	• considers also temporal information present in speech signals, apart from the frequency information • able to perform efficient time and frequency localisations • successfully used for de-noising tasks • capable of compressing a signal without major degradation	• not flexible since the same basic wavelets have to be used for all speech signals
WPT	• same as DWT, but WPT shows also further detail present in the high frequency bands	• not flexible since the same basic wavelets have to be used for all speech signals
LPC	• spectral envelope is represented with low dimension feature vectors • good source-to-vocal tract separation is obtained • LPC method is simple to implement and mathematically precise	• linear scales are not adequate for the representation of speech production or perception • Feature components are highly correlated • cannot include a priori information on the speech signal under test
LPCC	• same as LPC, but thanks to the cepstral analysis, the feature components are decorrelated • increase in robustness when compared to LPC	• linear scales are not adequate for the representation of speech production or perception • cannot include a priori information on the speech signal under test
PLP	• reduction in the discrepancy between voiced and unvoiced speech • PLP peaks are reasonably independent to the length of the vocal tract • resultant feature vector is low-dimensional • based on short term spectrum of the speech signals	• resultant feature vectors are dependent on the whole spectral balance of the formant amplitudes • spectral balance is easily altered by the communication channel, noise, and the equipment used

Classification technique	Advantages	Disadvantages
HMM	• able to model time distribution of speech signals • simple to adapt • capable to model a sequence of discrete or continuous symbols • inputs can be of variable length	• based on the assumption that the probability of being in a particular state is dependent only on its preceding state, ignoring any long-term dependencies • emission probabilities are arbitrarily chosen; hence, these might not even represent properly the output probabilities of the corresponding state
ANN (in general)	• good classifiers • highly adequate for pattern recognition applications • self-organising • self-learning • self-adaptive in new environments • robust	• based on ERM; hence, prone to over training a local minima problems
MLP	• good discriminating ability	• unable to model time distribution of speech signals • inputs have to be of fixed length • able to deal with small vocabularies only
SOM	• no a priori information is required for training a SOM • can easily adapt if a new sample is presented to it • capable of parallel computation	• SOM algorithm is not well defined mathematically; hence, values for the network parameters need to be found by trial-and-error • ordered mapping obtained after the training phase may be lost when applied in real environments due to frequent adaptations

Przykłady

Year	Research work	Speaker in/ dependent (SI/SD)	Language	Feature extraction technique	Classification technique	Language model	Accuracy, %
2009	isolated word recognition	SI	Malayalam	DWT WPT	MLP	N/A	89.00 61.00
2010	context-independent phoneme recognition	SI	TIMIT Corpus – 39 classes	MFCC	CDHMM	Bigram	63.07
2011	continuous phoneme recognition	SI	TIMIT Corpus – 39 classes	MFCC	HMM-MLP	Bigram	77.83
2003	isolated word recognition	SI	English SD2 Corpus	MFCC WPT	HMM	N/A	38.77 56.90
2009	isolated word recognition	SI	50 English words	Subband MFCC	CDHMM-FNN	N/A	89.50
2011	isolated word recognition	SI	Indian	LPCC MFCC	Modified-SOM	N/A	88.05 89.27
2011	continuous phoneme recognition	SI	TIMIT Corpus – 39 classes	PLP	SMLP	N/A	78.90
2002	isolated spoken digits	SD	Urdu	MFCC	MLP	N/A	94.00
2009	isolated spoken digits	SI	Persian	MFCC & DWT	MLP	N/A	98.00
2011	isolated word recognition	SI	six English words	LPCC	RBF MLP	N/A	98.69 96.00
2009	continuous word recognition	SI	ten English words	cepstrum analysis	HMM-RBF	N/A	80.00
1999	continuous phoneme recognition	SI	TIMIT Corpus – 39 classes	MFCC	SVM	N/A	77.60
2005	word recognition	SI	DARPA RM1 Corpus	MFCC	HMM-SVM	RM word-pair grammar	94.10

Dziękuję za uwagę.