Podstawy automatycznego rozpoznawania mowy

Prezentuje
Józef Kotus

Rys historyczny

- 1930-1950 – pierwsze systemy Automatycznego rozpoznawania mowy (ang. *Automatic Speech Recognition* – ASR), metody holistyczne; „ad-hoc”; izolowane słowa; małe słowniki; Bell Laboratories
- 1950-1960 – pierwsze systemy ASR oparte na zależnościach fonetycznych; małe słowniki;

Rys historyczny

- 1960-1980 – systemy oparte o rozpoznawanie wzorca (ang. *pattern recognition*); wykorzystanie parametrów kodowania predykcyjnego (LPC); sekwencje izolowanych lub połączonych słów; małe i średnie słowniki
- 1980-2000 – wprowadzenie statystycznego modelowania zależności dynamicznych i statycznych w **mowie ciągłej**; zastosowanie ukrytych modeli Markowa (ang. *Hidden Markov Models* - HMM)

Rys historyczny

- 2000-teraz – kombinacje modeli HMM oraz zależności akustyczno fonetycznych w celu znajdowania i korekcji nieregularności językowych, **deep learning**, systemy pracujące w chmurze; zwiększanie odporności systemu na pracę w środowisku szumowym; rozpoznawanie wielomodalne

Wprowadzenie

- Rozpoznawanie mowy (ang. speech recognition, speech-to-text – STT)
- Biorąc pod uwagę rodzaj rozpoznawanej mowy, wyróżnia się:
  - rozpoznawanie izolowanych słów (ang. isolated words),
  - rozpoznawanie słów łączonych (ang. connected words),
  - rozpoznawanie mowy ciągłej (ang. continuous speech),
  - rozpoznawanie mowy spontanicznej (ang. spontaneous speech).
- W przypadku sposobu obsługi mówcy (ang. speaker dependence) systemy dzielą się:
  - na te, które potrafią rozpoznać tylko konkretnego mówcę (ang. Speaker dependent system),
  - te, które potrafią rozpoznać dowolnego mówcę (ang. Speaker independent system)
  - oraz wreszcie na te, które adaptują się do konkretnego mówcy (ang. speaker adaptable system)

Karolina Kuligowska, Paweł Kisielewicz, Aleksandra Włodarz, Wady i ograniczenia systemów rozpoznawania mowy, http://bazekon.icm.edu.pl/bazekon/element/bwmeta1.element.ekon-element-000171530952

Wprowadzenie

- **Słownik** jest zbiorem wyrazów, które mają być rozpoznane, dlatego tak ważny jest jego rozmiar.
- **Jeżeli liczba słów w słowniku jest bardzo mała**, ale słowa te znacząco różnią się między sobą pod względem akustycznym, **system może osiągnąć bardzo wysoki poziom dokładności rozpoznawania mowy**.
- Im większy słownik, tym więcej niejasności wynikających z liczby możliwych alternatywnych sposobów wymowy danego słowa.
- **Małe słowniki mogą zwierać poniżej 30 słów**, z kolei większość dużych systemów rozpoznawania mowy zawiera słowniki o rozmiarach kilku tysięcy słów.
- **Słowniki systemów przeznaczonych do dyktowania i transkrypcji mogą zawierać 10.000 słów i więcej**.
- Nawet tak duży rozmiar słownika może nie być wystarczający, gdyż w świecie rzeczywistym **nigdy nie da się przewidzieć, jakie słowa wypowie użytkownik**

Skuteczność rozpoznawania

- Do oceny skuteczności systemu ASR stosowana jest miara „**wyrazowej stopy błędu**”
  (ang. ***Word Error Rate*** – ***WER***)

$$WER = \frac{\quad}{\quad} \cdot 100\%$$

Skuteczność rozpoznawania

- Do oceny skuteczności systemu ASR stosowana jest miara „**wyrazowej stopy błędu**” (ang. ***Word Error Rate*** – ***WER***)

$$\boldsymbol{WER} = \frac{D+S+I}{\underbrace{H+D+S}_{N}} \cdot 100\%$$

H – liczba poprawnie rozpoznanych słów
D – liczba nie rozpoznanych słów (ang. *deletions*)
S – liczba błędnie rozpoznanych słów (ang. *substitutions*)
I – liczba wstawionych słów (ang. *insertions*)
N – liczba nadanych słów = *H*+*D*+*S*

Speech-to-text

- Automatic Speech Recognition (ASR)
- Ideally we want to have a system that deals with: spontaneous speech, multi-speakers, unlimited output vocabulary, any acoustic condition
- But performances differ greatly for different contexts (read vs spontaneous speech ; small vs large vocabulary ; quiet vs noisy)

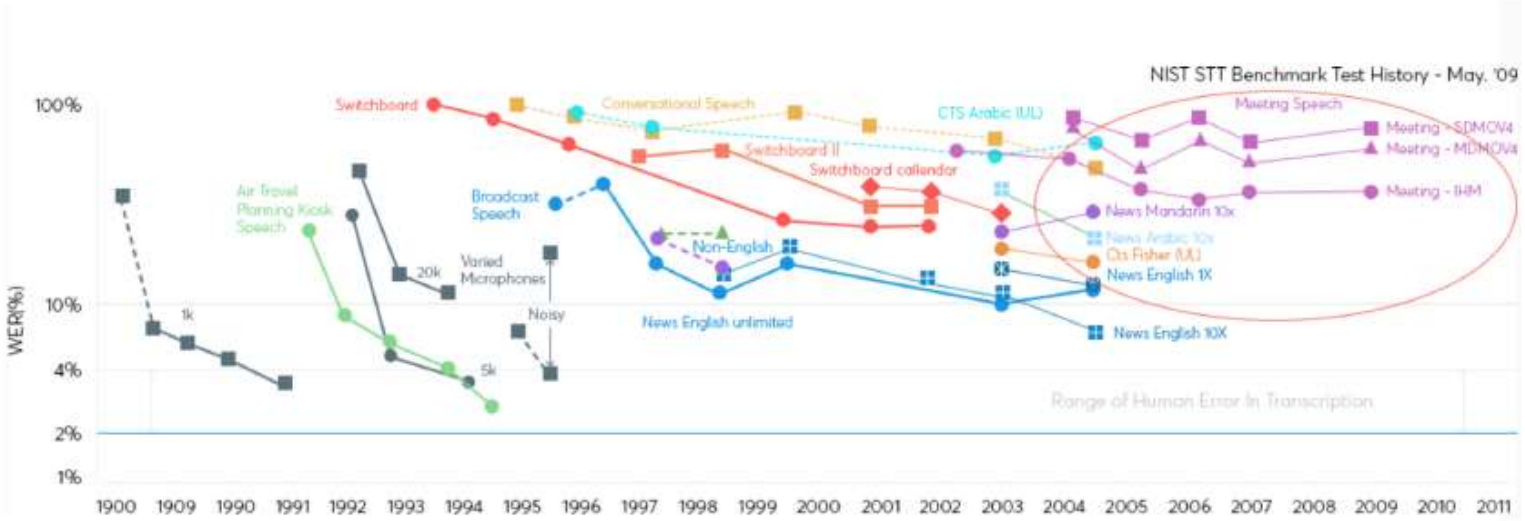
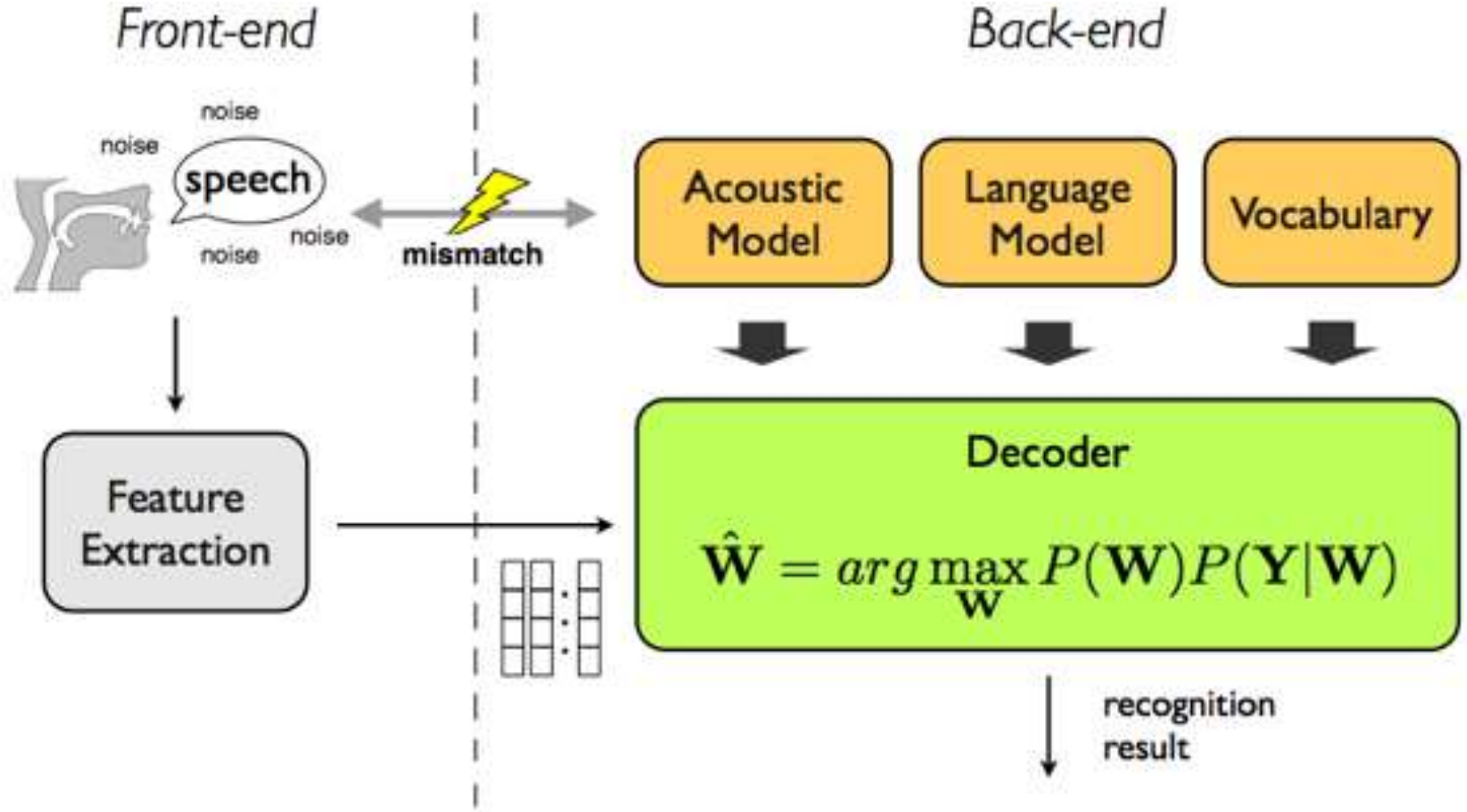


Figure: NIST ASR benchmark tests history ($<$ 2015)

Podejścia i techniki wykorzystywane w rozpoznawaniu mowy

- **Podejście akustyczno-fonetyczne (ang. acoustic-phonetic approach)**, które zakłada, że jednostki fonetyczne są charakteryzowane przez szereg cech, takich jak na przykład częstotliwość, tonacja, barwa dźwięku. Cechy te są wydobywane z sygnału mowy i wykorzystywane między innymi przy segmentacji mowy.
- **Podejście wykorzystujące rozpoznawanie wzorców (ang. *pattern recognition approach*)**, które obejmuje dwa niezbędne etapy: trenowanie wzorców (ang. *Pattern training*) oraz porównanie wzorców (ang. *pattern comparison*). Ważną cechą takiego podejścia jest to, że wykorzystuje ono dobrze sformułowany aparat matematyczny oraz ustanawia spójne reprezentacje wzorców mowy dla wiarygodnego porównywania wzorców, od zestawu oznakowanych próbek treningowych po formalny algorytm treningowy.
- **Podejście wykorzystujące wiedzę (ang. *knowledge based approach*)**, nazywane również podejściem bazującym na sztucznej inteligencji (ang. *Artificial Intelligence approach*). Polega ono na zmechanizowaniu procedury rozpoznania] mowy w sposób zbliżony do tego, jak dokonuje tego człowiek, wykorzystując posiadaną wiedzę dotyczącą między innymi cech akustycznych. Podejście to jest połączeniem podejścia akustyczno-fonetycznego i podejścia wykorzystującego rozpoznawanie wzorców.

Typowy schemat systemu ASR



https://www.esat.kuleuven.be/psi/spraak/theses/08-09-en/MDT.php

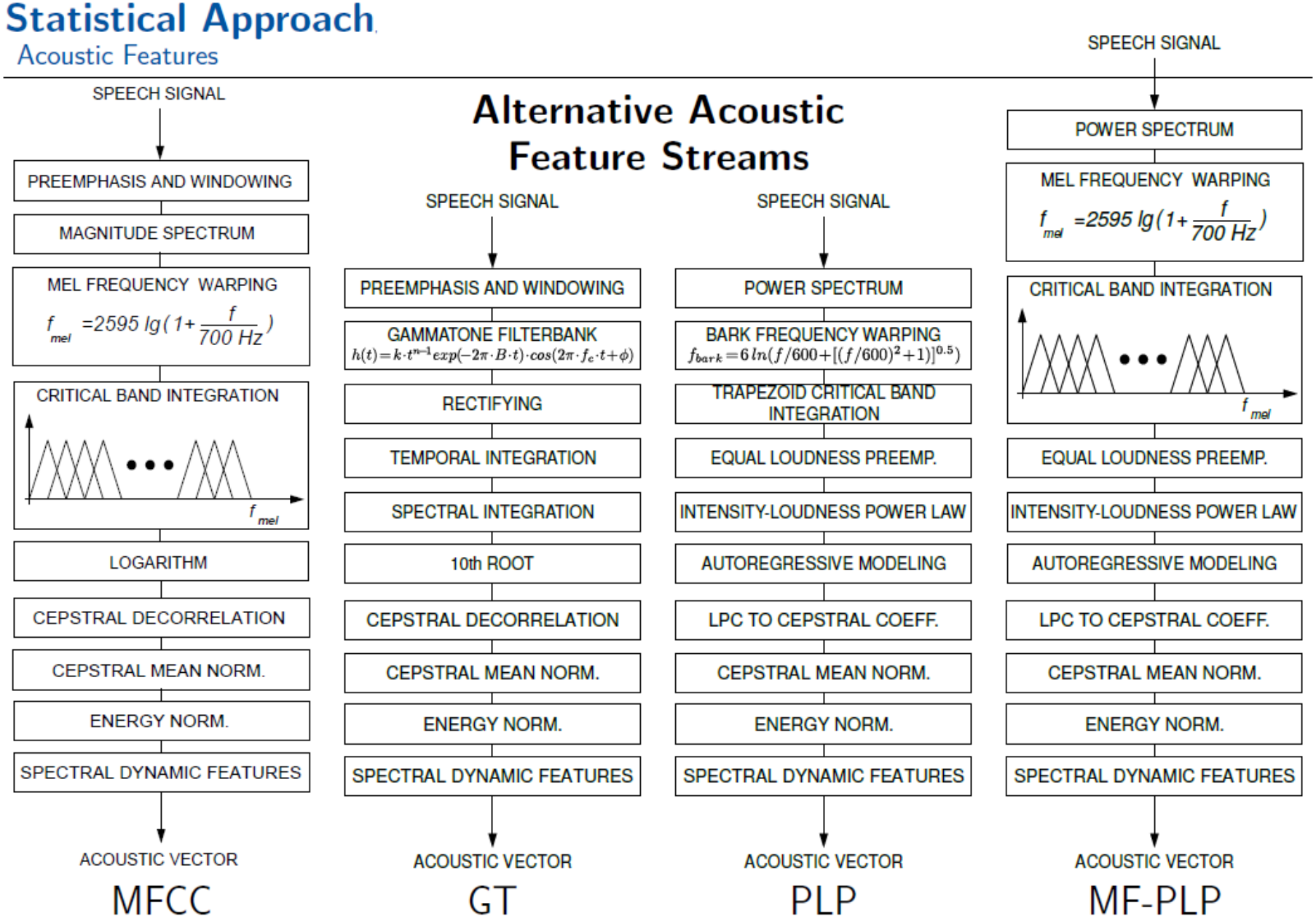
Model akustyczny

- Przy tworzeniu modelu akustycznego korzysta się najczęściej z parametrów mel-cepstralnych (MFCC) lub parametrów LPC
- W celu zamodelowania najkrócej trwających fonemów (głoski wybuchowe – ang. *plosive phonemes: -p; -t; -k*) należy dobrać odpowiednie okno analizy – typowo o długości 10 ms
- Typowo stosuje się modele **trifonowe**
- **Istotne – osobny model ciszy**

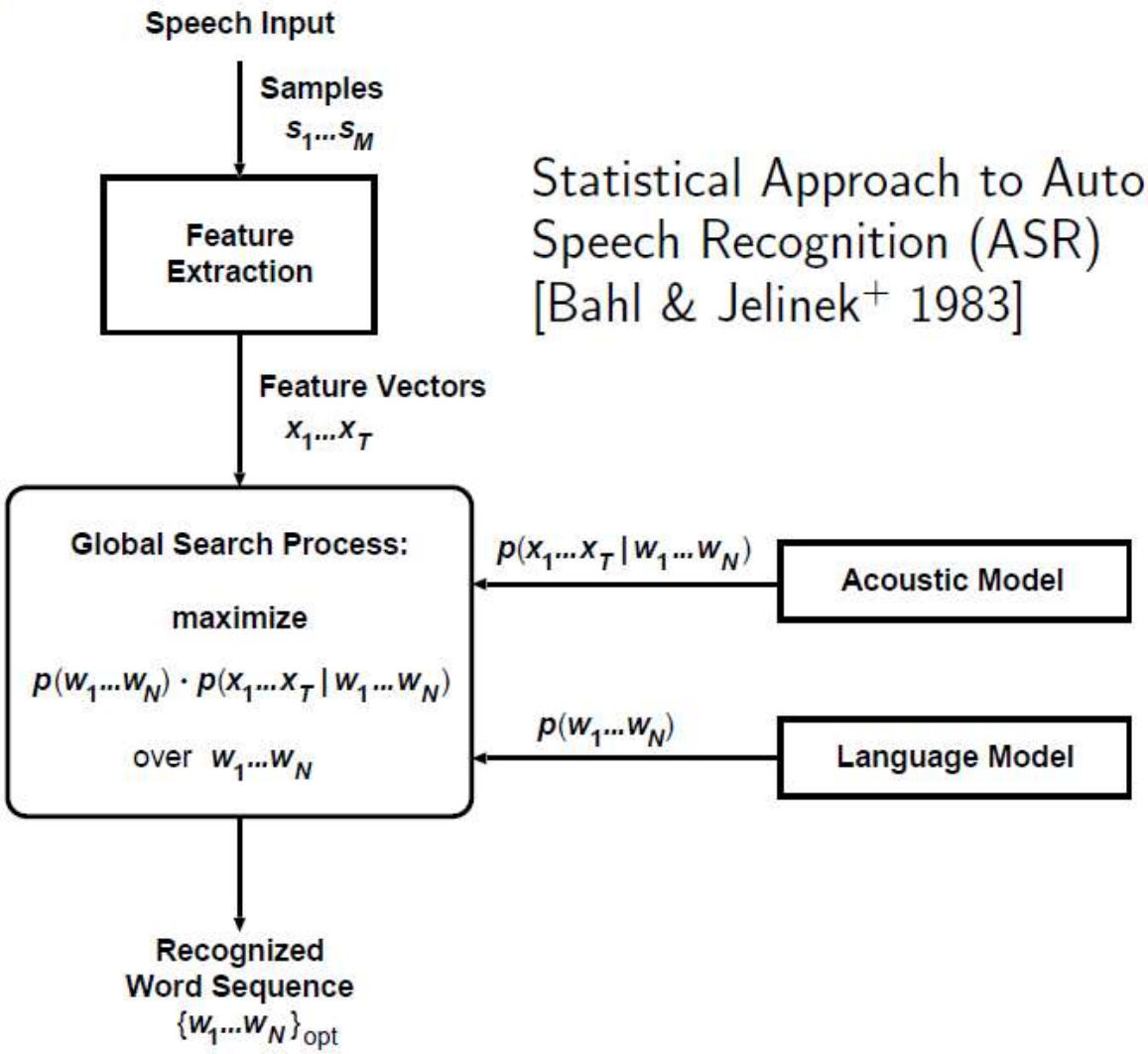
Ilustracja podziału na trifony

Zdjęcie pochodzi z filmu: Uczenie maszynowe – system rozpoznawania mowy
https://pionier.tv/wideo/czas-nauki/uczenie-maszynowe-system-rozpoznawania-mowy/

Alternative Acoustic Feature Streams

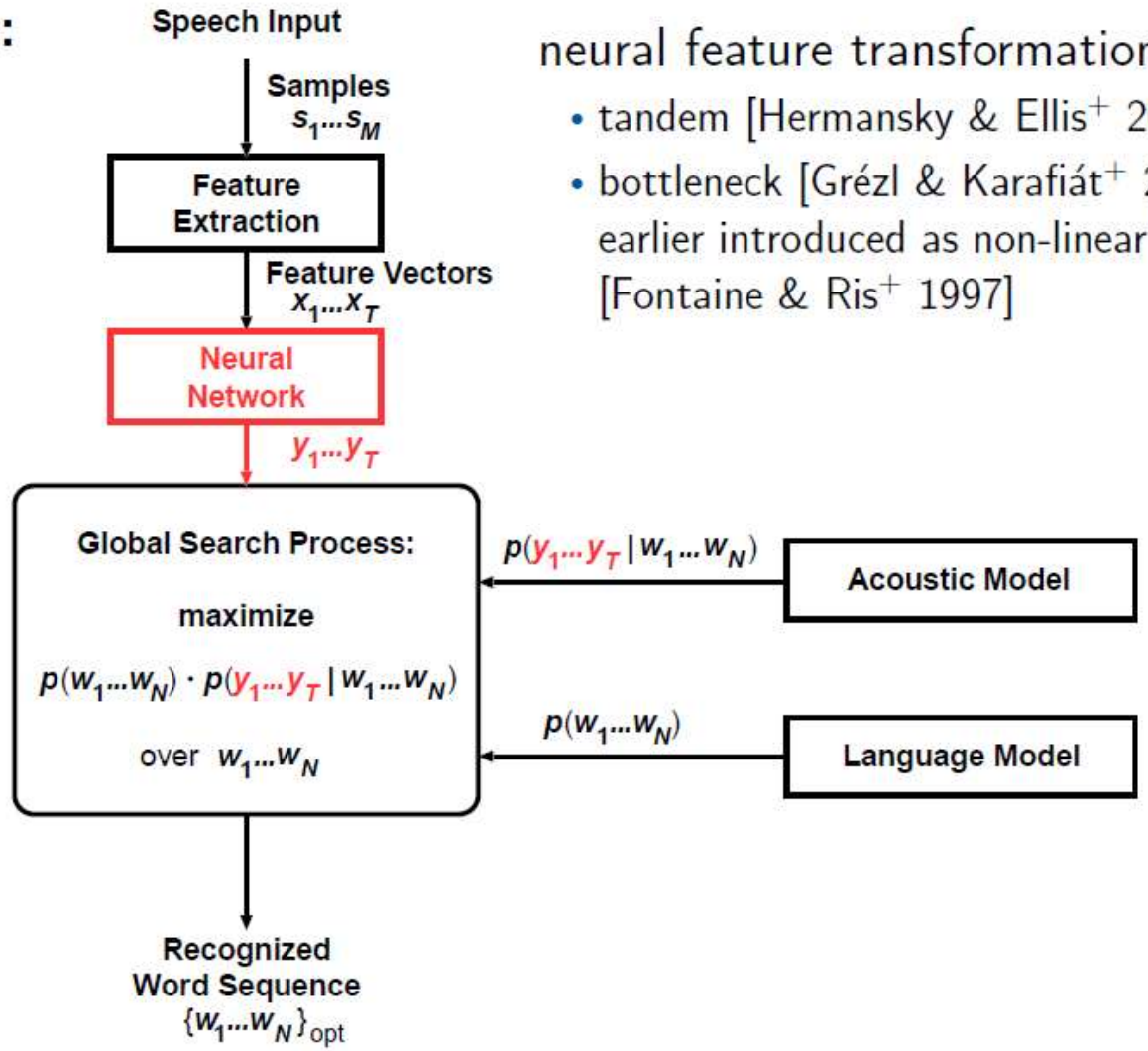


ASR Architecture



Statistical Approach to Automatic Speech Recognition (ASR)
[Bahl & Jelinek+ 1983]

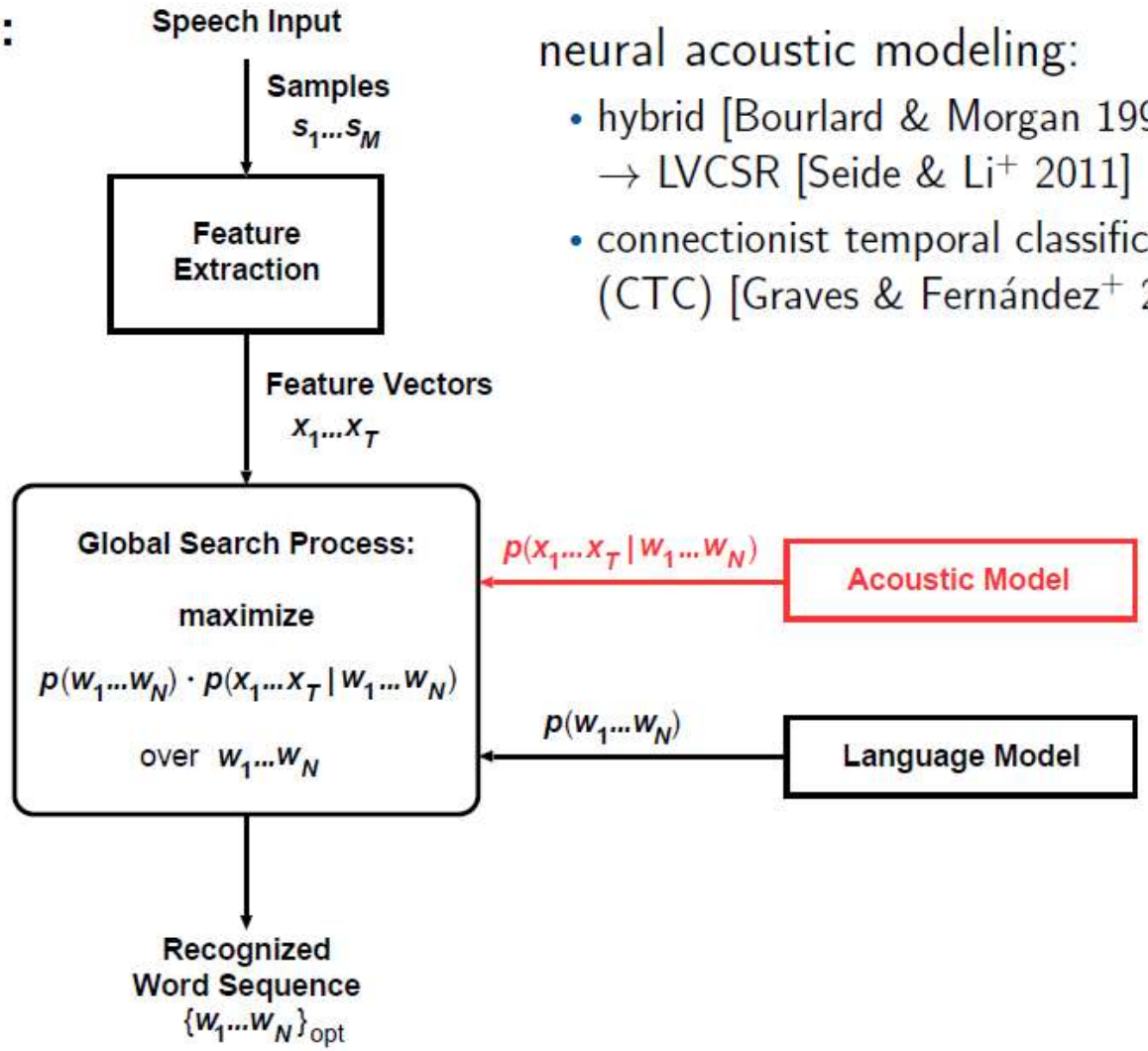
ASR Architecture: Neural Networks



neural feature transformation:

- tandem [Hermansky & Ellis$^+$ 2000]
- bottleneck [Grézl & Karafiát$^+$ 2007] earlier introduced as non-linear LDA [Fontaine & Ris$^+$ 1997]

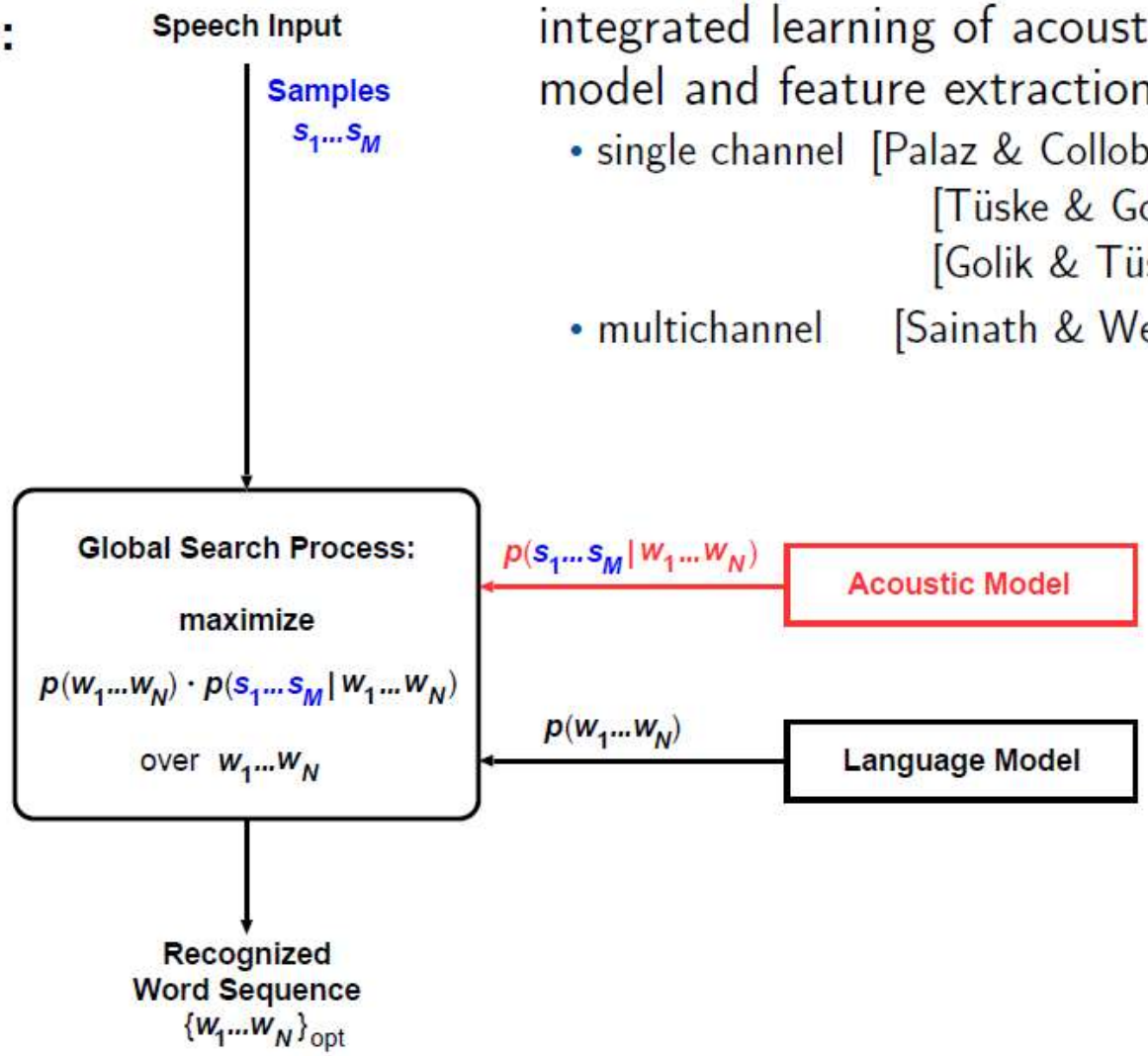
ASR Architecture: Neural Networks



neural acoustic modeling:

- hybrid [Bourlard & Morgan 1993] → LVCSR [Seide & Li+ 2011]
- connectionist temporal classification (CTC) [Graves & Fernández+ 2006]

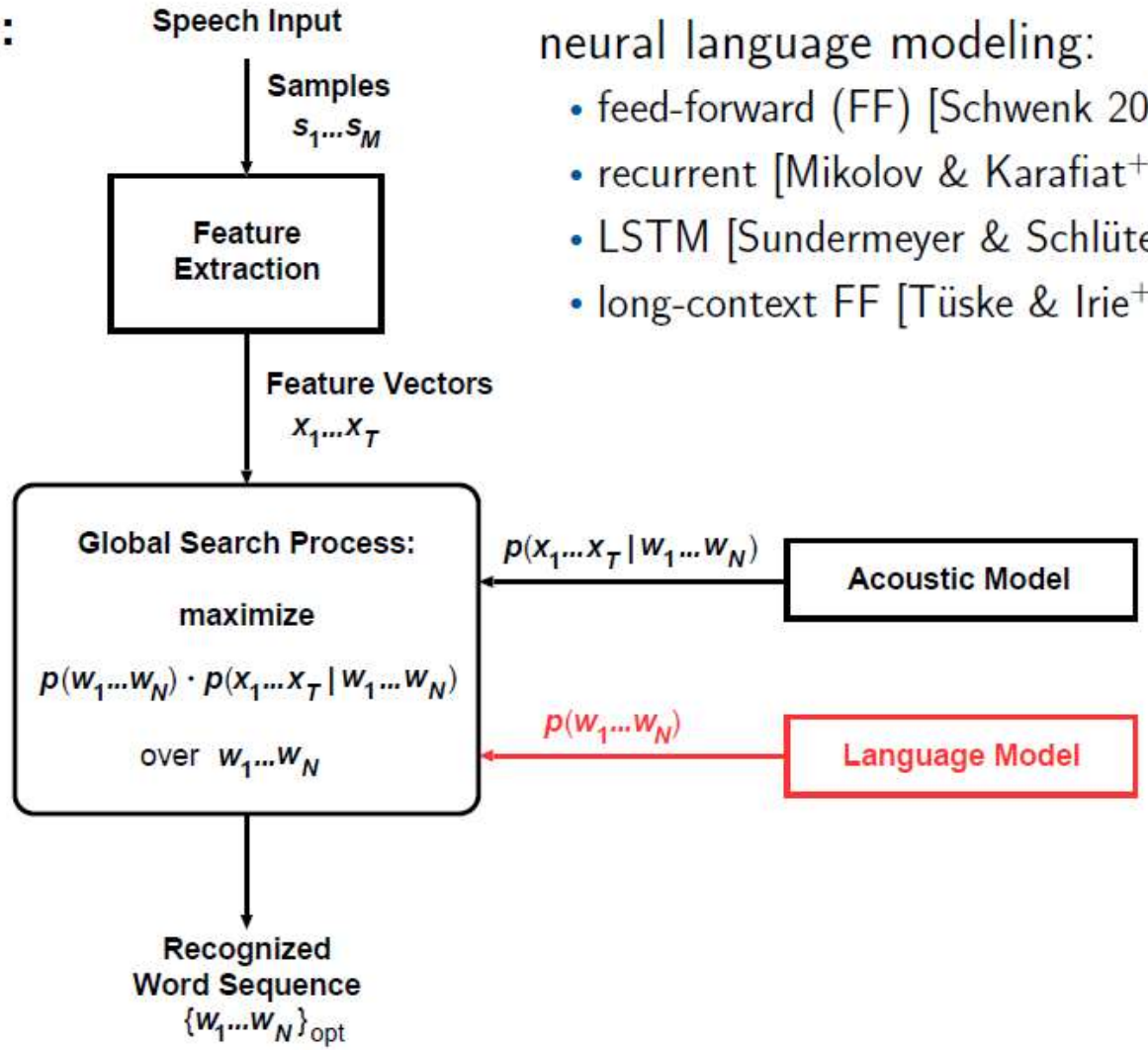
ASR Architecture: Neural Networks



integrated learning of acoustic model and feature extraction

- single channel [Palaz & Collobert+ 2013] [Tüske & Golik+ 2014] [Golik & Tüske+ 2015]
- multichannel [Sainath & Weiss+ 2015]

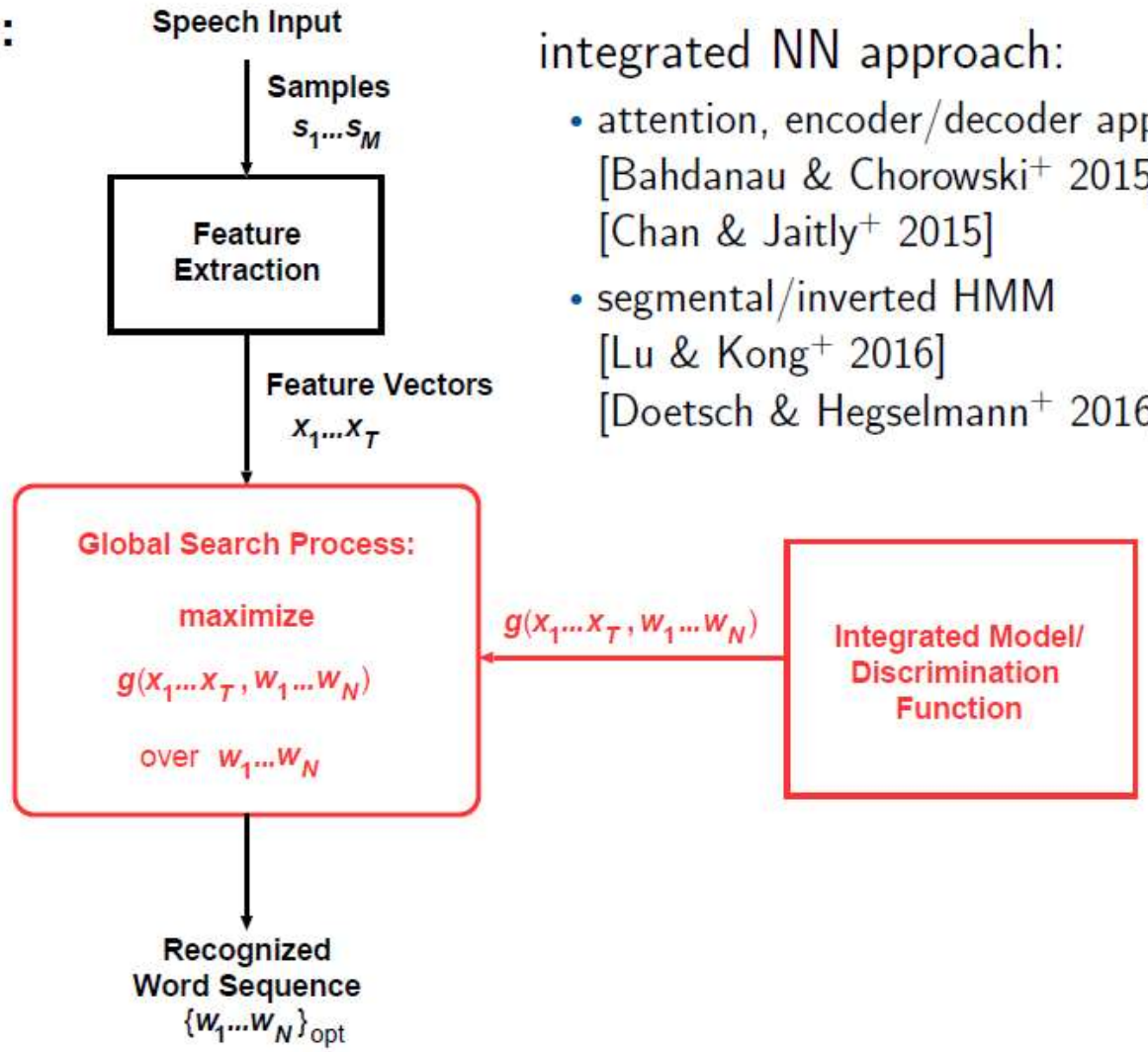
ASR Architecture: Neural Networks



neural language modeling:

- feed-forward (FF) [Schwenk 2007]
- recurrent [Mikolov & Karafiat+ 2010]
- LSTM [Sundermeyer & Schlüter+ 2012]
- long-context FF [Tüske & Irie+ 2016]

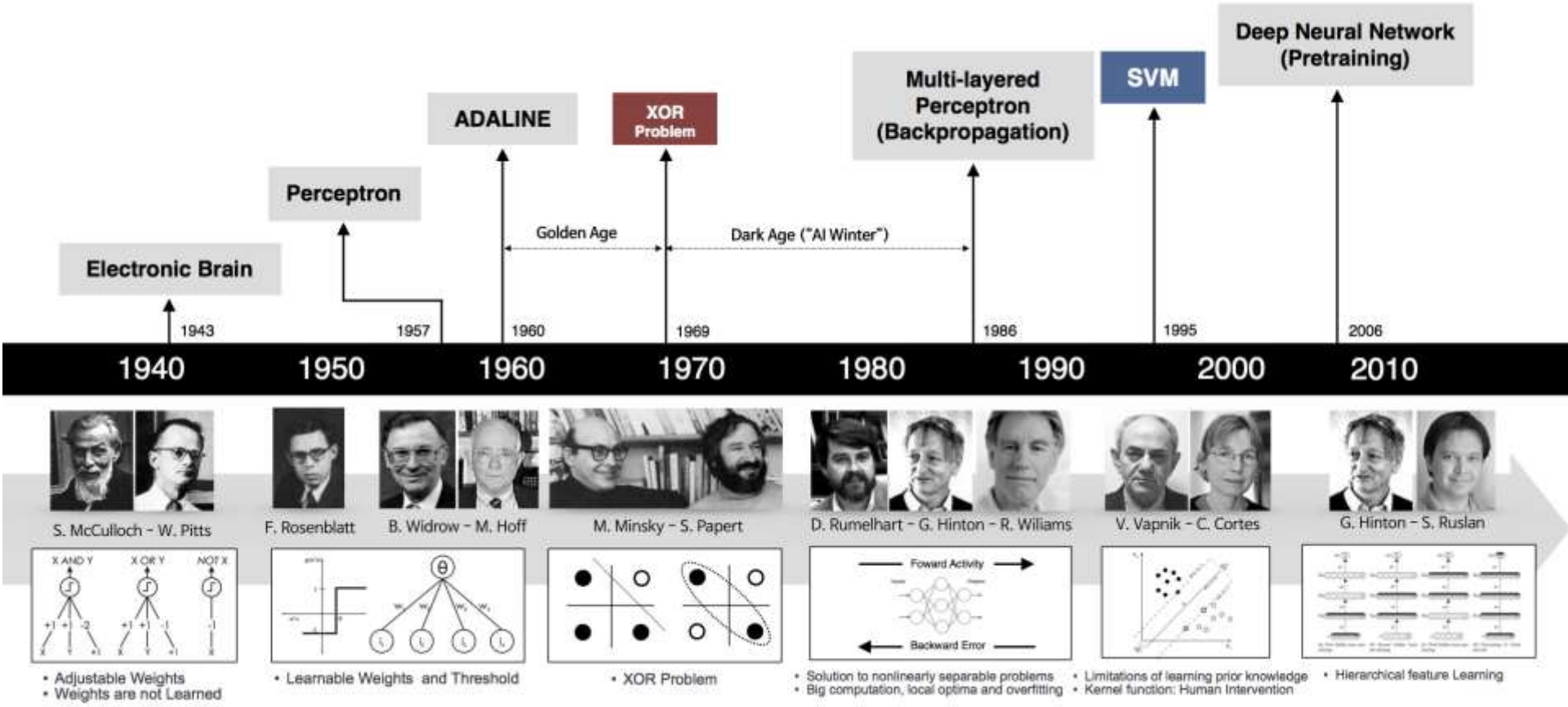
ASR Architecture: Neural Networks



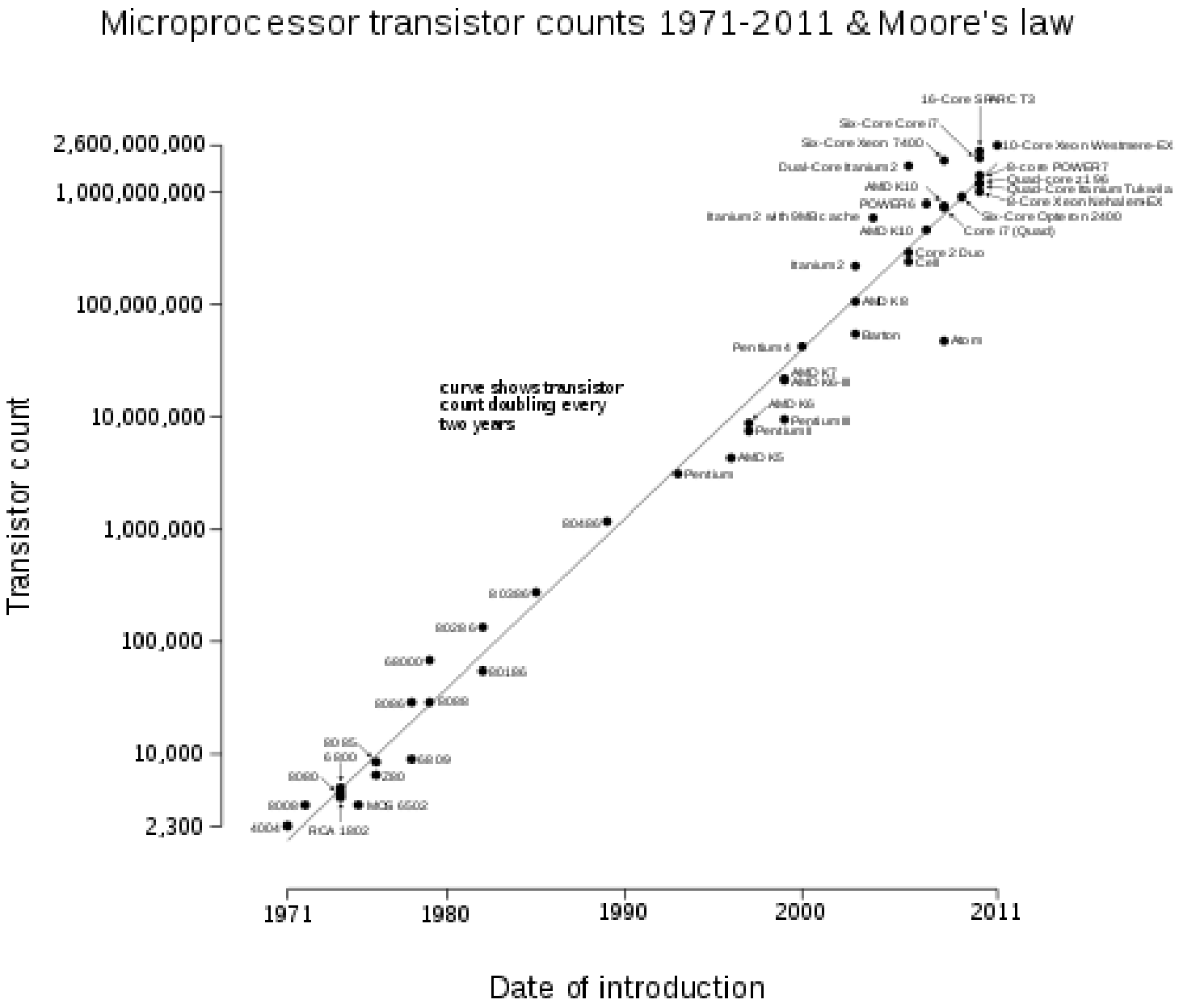
integrated NN approach:

- attention, encoder/decoder approach
  [Bahdanau & Chorowski+ 2015]
  [Chan & Jaitly+ 2015]
- segmental/inverted HMM
  [Lu & Kong+ 2016]
  [Doetsch & Hegselmann+ 2016]

Kamienie milowe w rozwoju sztucznych sieci neuronowych



http://beamlab.org/deeplearning/2017/02/23/deep_learning_101_part1.html



https://commons.wikimedia.org/wiki/User:Wgsimon

Progresses over the years

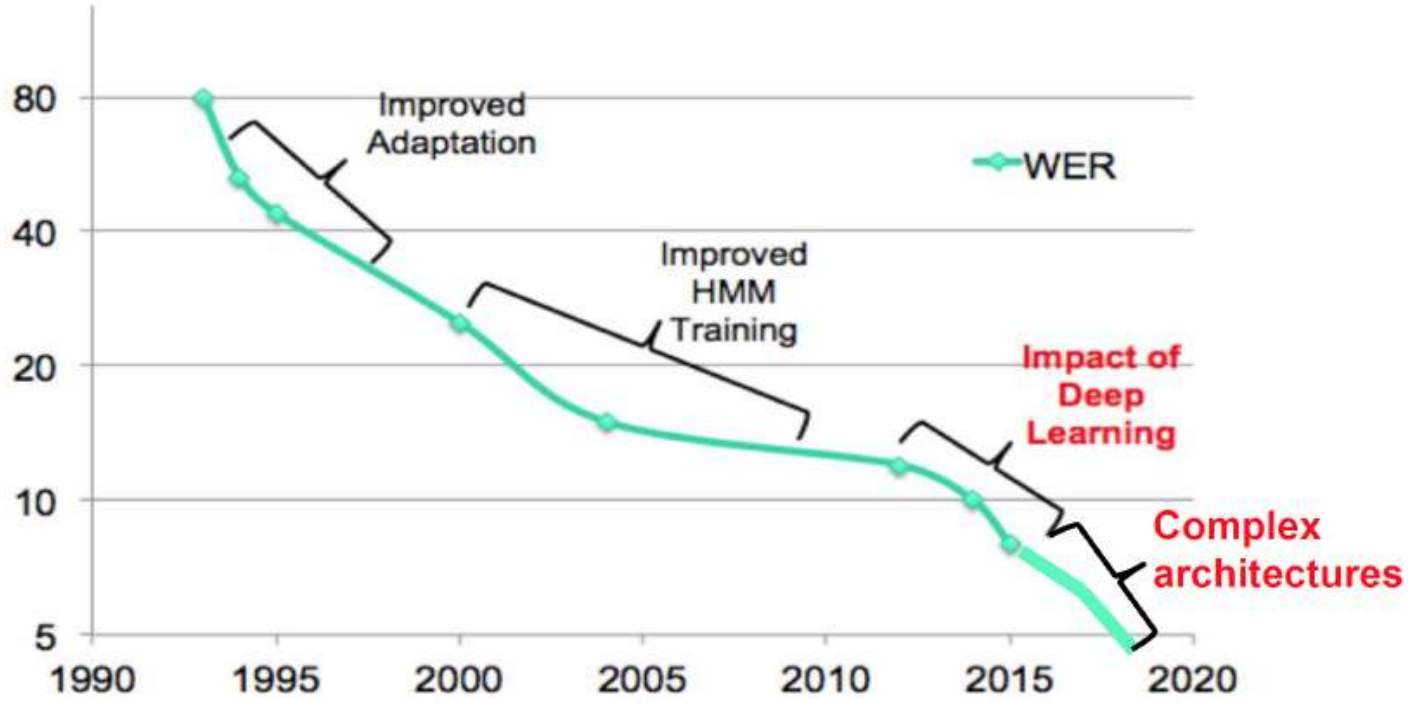


Figure: ASR Performance[1] on English Conversational Telephony (Switchboard)

[1] Image from Bhuvana Ramabhadran' s presentation at Interspeech 2018

Complexity: Computation Times

Training times (without GPUs!) for training corpus of 50 Million words:

Models	PPL	CPU Time (Order)
Count model	163.7	30 min
MLP	136.5	1 week
LSTM-RNN	107.8	3 weeks

- problem: high computation times
- remedy: two types of language models:
  - count model: trained on a huge corpus: 3.1 Billion words
  - NN models: trained on a small corpus: 50 Million words
- resulting language model:
  linear interpolation of *two* models

Deep learning breakthrough

Like in vision, due to

- More data
  - ex: (2015) Librispeech (en) 1.000h (Panayotov et al., 2015)
  - ex: (2016) Baidu Deep Speech 2 (en) 12.000h (Amodei et al., 2016)
  - ex: (2017) Google Home (en) 18.000h (from a Google presentation)
  - ex: (2018) Google wav2words (en) >100.000h?[3] (informal discussion)
- Computation (ex: GPU)
- Better optimization algorithms and training objectives
- ASR Toolkits (ex: Kaldi (Povey et al., 2011) and DL frameworks (Tensorflow and the like)

[3] >11 years of speech !

On par with human transcription ?

Read Speech			
Test set	DS1	DS2	Human
WSJ eval'92	4.94	3.60	5.03
WSJ eval'93	6.94	4.98	8.08
LibriSpeech test-clean	7.89	5.33	5.83
LibriSpeech test-other	21.74	13.25	12.69

Figure: Comparison of WER for two speech systems and human level performance on **read** speech (from (Amodei et al., 2016)

Accented Speech			
Test set	DS1	DS2	Human
VoxForge American-Canadian	15.01	7.55	4.85
VoxForge Commonwealth	28.46	13.56	8.15
VoxForge European	31.20	17.55	12.76
VoxForge Indian	45.35	22.44	22.15

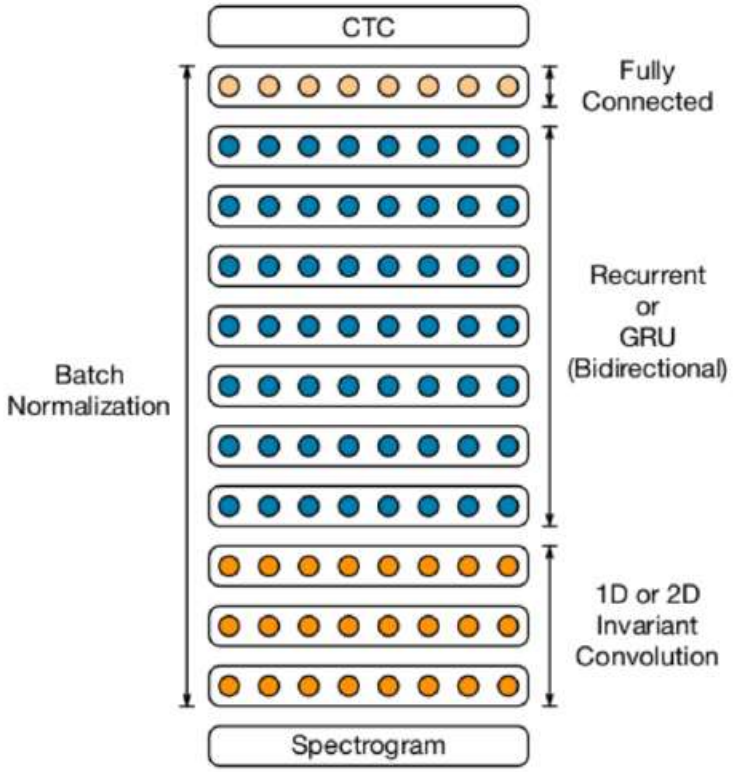
Figure: Comparison of WER for two speech systems and human level performance on **accented** speech (from (Amodei et al., 2016)

On par with human transcription ?

Noisy Speech			
Test set	DS1	DS2	Human
CHiME eval clean	6.30	3.34	3.46
CHiME eval real	67.94	21.79	11.84
CHiME eval sim	80.27	45.05	31.33

Figure: Comparison of WER for two speech systems and human level performance on **noisy** speech (from (Amodei et al., 2016)

Deep Speach 2



Read Speech			
Test set	DS1	DS2	Human
WSJ eval'92	4.94	3.60	5.03
WSJ eval'93	6.94	4.98	8.08
LibriSpeech test-clean	7.89	5.33	5.83
LibriSpeech test-other	21.74	13.25	12.69

Amodei, Dario, et al. "Deep speech 2: End-to-end speech recognition in english and mandarin."(2015).

https://www.is.umk.pl/~grochu/wiki/lib/exe/fetch.php?media=zajecia:nn_2018_1:nn-wyklad.pdf

Wady i ograniczenia systemów rozpoznawania mowy

- Środowisko rozpoznawania mowy
- Urządzenie rejestrujące głos
- Prozodia
- Zmienność mowy
- Styl mowy
- Mowa dzieci
- Ograniczony słownik
- Kontekst i homonimy
- Różnorodność języków
- Języki słowiańskie

Karolina Kuligowska, Paweł Kisielewicz, Aleksandra Włodarz, Wady i ograniczenia systemów rozpoznawania mowy, Roczniki Kolegium Analiz Ekonomicznych / Szkoła Główna Handlowa, 2018 | nr 49 Społeczno-ekonomiczne aspekty rozwoju gospodarki cyfrowej : koncepcje zarządzania i bezpieczeństwa | 307—317

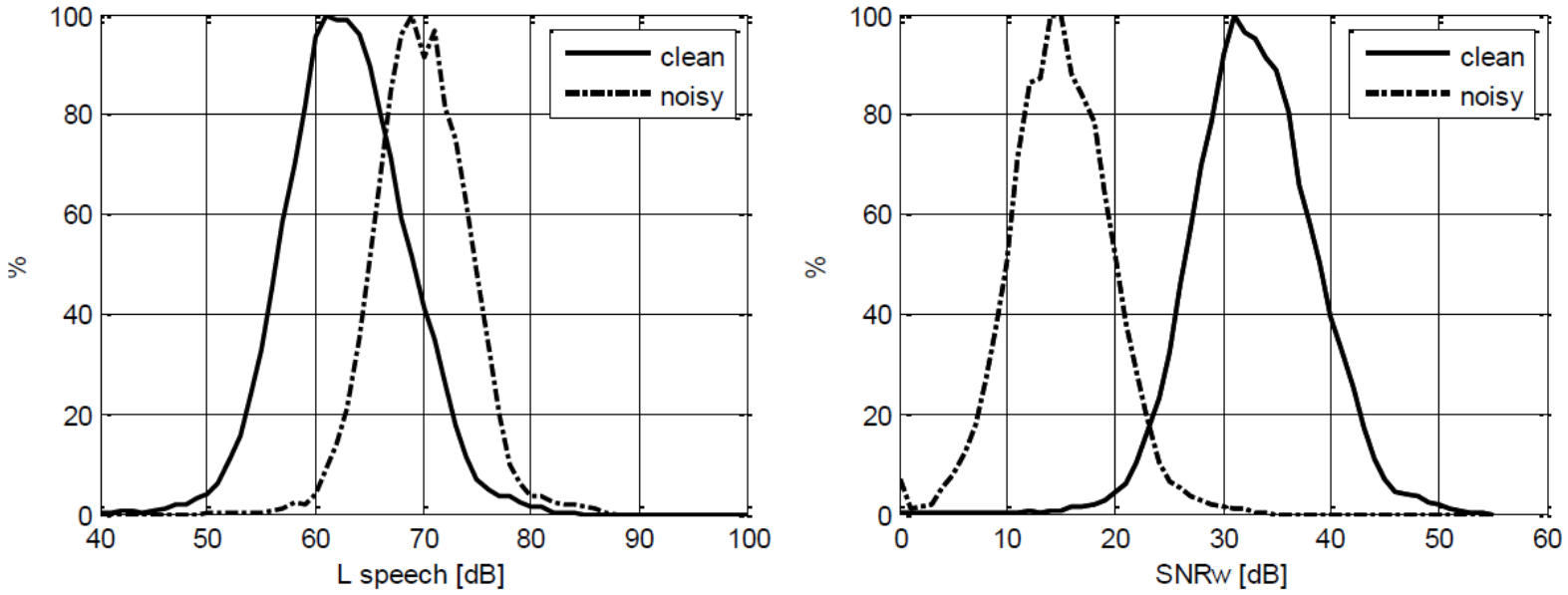
http://bazekon.icm.edu.pl/bazekon/element/bwmeta1.element.ekon-element-000171530952

Środowisko rozpoznawania mowy

- Wydajność rozpoznawania mowy drastycznie spada w hałaśliwym otoczeniu
  - zakłócenia w przestrzeni powodują rozbieżności pomiędzy warunkami treningowymi (czystymi) oraz warunkami, w jakich odbywa się rozpoznawanie (hałaśliwymi),
  - zakłócenia zniekształcają i zanieczyszczają sygnał mowy oraz zmieniają wektory danych reprezentujących mowę.
- Badania prowadzone nad problemem odporności na zakłócenia skupiają się na dwóch kierunkach:
  - usuwanie szumu z zakłóconego hałasem sygnału poprzez filtr szumów, odejmowanie widmowe, filtr Wienera, filtr RASTA, mapowanie wektorów stochastycznych;
  - kompensowanie efektu szumu w przestrzeni modelu akustycznego, dopasowując środowisko treningowe do warunków realizacyjnych, w jakich odbywa się rozpoznawanie mowy poprzez mapowanie wektorów stochastycznych oraz sekwencyjne szacowanie hałasu.

Środowisko rozpoznawania mowy

- **Efekt Lombarda** lub refleks Lombarda – niezamierzona tendencja mówiącego do zwiększenia natężenia głosu w celu poprawienia słyszalności przy mówieniu w głośnym otoczeniu.



https://pl.wikipedia.org/wiki/Efekt_Lombarda

Czyzewski, A., Kostek, B., Bratoszewski, P. et al. An audio-visual corpus for multimodal automatic speech recognition. J Intell Inf Syst ***49,*** *167–192 (2017). https://doi.org/10.1007/s10844-016-0438-z*

Urządzenie rejestrujące głos

- Gdy mikrofon nie jest wystarczająco czuły albo zbyt wrażliwy, może wygenerować informację audio, która będzie trudna do rozszyfrowania.
- Do izolowania głosów od szumów często stosuje się zestaw mikrofonów, gdzie czysty sygnał mowy przechwycony przez kilka mikrofonów jest oddzielony od hałaśliwego sygnału.
- Zestaw mikrofonów można kierować w najbardziej dogodną stronę, co poprawia wydajność rozpoznawania, jednak zakłada to synchroniczną i ciągłą obserwację sygnału, co nie zawsze jest możliwe do uzyskania.

Prozodia

- Prozodia jest istotna w zrozumieniu języka mówionego: ułatwia rozpoznać wypowiadane słowa, globalne i lokalne dwuznaczności oraz analizować strukturę dyskursu.
- Jednakże cechy prozodyczne nie są wykorzystywane w większości współczesnych systemów rozpoznawania mowy.
- Informacje prozodyczne są trudne do modelowania i nadal szuka się rozwiązań, które pomogłyby przezwyciężyć problem prozodii w kontekście systemów automatycznego rozpoznawania mowy.

Zmienność mowy

- Każdy człowiek ma swój indywidualny sposób mówienia, inny ton i barwę głosu, mówi w innym tempie oraz rytmie, inaczej artykułuje wyrazy, używa innego języka.
- Zmienność mowy determinują także wszelkie wady wymowy i problemy z dykcją, różnice demograficzne, kulturowe i geograficzne oraz wiek, przynależność klasowa i akcent.

Styl mowy

- Systemy rozpoznawania izolowanych słów wymagają krótkich pauz pomiędzy wypowiadanymi słowami.
- Taki styl mówienia nie jest naturalny, dlatego systemy te tracą na swej popularności na rzecz systemów rozpoznawania mowy ciągłej, na których obecnie skupiona jest większość badań.
- **W spontanicznej swobodnej wypowiedzi lub pod presją czasu bardzo często dochodzi do redukcji wymowy niektórych fonemów lub sylab**, co może doprowadzić do utraty części informacji i niesie za sobą wyższy wskaźnik błędów podczas rozpoznawania.

Mowa dzieci

- Dzieci, w porównaniu z dorosłymi, mają krótszą krtań oraz fałdy głosowe.
- Skutkuje to słabą rozdzielczością widmową dźwięków głosu oraz nieliniowym wzrostem formantowych częstotliwości.
- Sporym problemem jest też nieprawidłowa wymowa dzieci. Bardzo często nie znają one poprawnych form fleksyjnych określonych słów, szczególnie tych, które są wyjątkami do ogólnie przyjętych zasad.
- Pomimo tego, iż zaproponowano kilka technik, które miały za zadanie poprawę dokładności systemów rozpoznawania w przypadku dziecięcych głosów, wydajność takich systemów jest dużo niższa niż w przypadku rozpoznawania mowy dorosłego człowieka.

Ograniczony słownik

- **Dużo pracy wymaga stworzenie i rozwinięcie dobrego słownika.**
- Większość systemów automatycznego rozpoznawania mowy działa z dużym, lecz zwykle ograniczonym słownikiem, znajdującym najlepiej pasujące słowa dla danego sygnału akustycznego.
- Podczas gdy systemy rozpoznawania mowy ciągłej tworzą wysokiej jakości transkrypcję, nie radzą sobie z rozpoznawaniem słów spoza słownika.
- **Jeśli słowa w języku wykazują zmienność morfologiczną, konieczne może okazać się rozszerzenie słownika nawet do setek tysięcy słów.**
- Niektóre języki posiadają tak bogatą morfologię, że modelowanie języka wymagałoby słownika, który wykracza rozmiarem ponad rozsądną wielkość.
- W takich przypadkach, aby skonstruować słownik, najlepiej zrezygnować z modelowania opartego na słowach, a wykorzystać podsłowa, na przykład morfemy.

Kontekst i homonimy

- Jednym z problemów rozpoznawania mowy jest określenie kontekstu, w jakim słowa zostały wymówione.
- Niektóre słowa, które brzmią bardzo podobnie, mogą zostać dobrze rozpoznane tylko wtedy, gdy znany jest ich kontekst.
- Dodatkowo kontekst wpływa na dokładność rozpoznania homonimów – wyrazów o takim samym brzmieniu, lecz różnym znaczeniu.
- Systemy rozpoznawania mowy nie mają możliwości odróżnienia ich na podstawie samego dźwięku.
- Określenie kontekstu danego słowa wpływa pozytywnie na dokładność rozpoznania i wydajność systemów pod względem rozróżniania homonimów.

Kontekst i homonimy

Jerzy Ficowski
DZIWNA RYMOWANKA

Pewien żarłok nie nażarty
raz wygłodniał nie na żarty.
I wywiesił szyld na płocie
że ochotę ma na płocie.
Tutaj na brak ryb narzeka,
bo daleko rybna rzeka.
Więc się zgłosił pewien żebrak
i rzekł żarłokowi, że brak
płoci, karpi oraz śledzi,
ale rzeki pilnie śledzi,
i gdy tylko będzie w stanie
to o świcie z łóżka wstanie,
po czym ruszy na Pomorze
i w zdobyciu ryb pomoże…
Odtąd żarłok nasz jedynie,
zamiast smacznych ryb je dynie.

http://zecernia.com/html/?p=2108

Różnorodność języków

- **Różnorodność języków stawia wyzwania przed rozpoznawaniem mowy.**
- Języki aglutynacyjne mają bogatszy zasób słownictwa (a tym samym większe słowniki) ze względu na tworzenie słów, polegające na łączeniu ze sobą wielu morfemów.
- Z kolei języki fleksyjne charakteryzują się stosunkowo swobodnym szykiem zdania oraz bardzo bogatym systemem morfologicznym i derywacyjnym.
- **Dokładna analiza cech dystynktywnych danego języka ułatwia wybór metody rozpoznawania mowy.**
- Rodzaj fonemów występujących w danym języku, warianty alofoniczne, wzory sylab oraz cechy fleksyjne decydują o tym, jaką technikę zastosować dla rozpoznawania mowy w danym języku.

Language coverage

- Google addresses (only) 100 languages (ASR)
- Language technology issues: 300 languages (95 % population)
- Language coverage / revitalisation / documentation issues: $>$ 6000 languages !

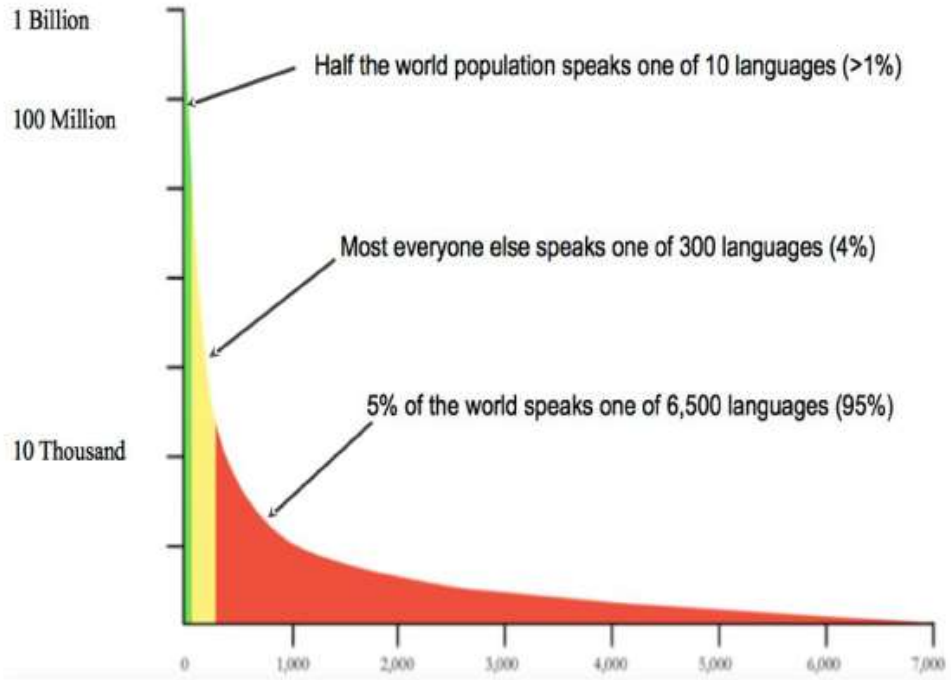


Figure: from Laura Welcher - Big Data for Small Languages The Rosetta Project

Języki słowiańskie

- **Większość systemów rozpoznawania mowy operuje na najbardziej rozpowszechnionych w świecie językach, takich jak angielski, francuski, niemiecki czy japoński.**
- Języki słowiańskie w dalszym ciągu czekają na intensywny rozwój technologii mowy pod ich kątem.
  - **Jednym z wyzwań jest fleksyjna natura języków słowiańskich, która modyfikuje podstawową formę elementów leksykalnych zgodnie z relacjami gramatycznymi, morfologicznymi i kontekstowymi.**
  - **Liczba powstawania wielu form wyrazowych bardzo często przekracza milion odrębnych pozycji, które muszą zostać uwzględnione i właściwie zarządzane.**
  - Różnica ta jest bardzo duża w porównaniu z systemami rozpoznawania mowy zaprojektowanymi na przykład dla języka angielskiego, w którym słownik 50 tys. najczęściej używanych słów daje wskaźnik pokrycia 99%.
  - Języki słowiańskie wymagają na ogół słowników, które są od 10 do 20 razy większe20

Audiowizualne rozpoznawanie mowy

Autor:
Piotr Bratoszewski

Bimodalne rozpoznawanie mowy

- Dołączenie do wektora parametrów akustycznego parametrów wizyjnych
- Fonemy = wizemy
- Widowiskowe podejście – czytanie z ruchu warg
- Teoretycznie wzrost skuteczności w warunkach szumowych
- Wiele problemów do rozwiązania (detekcja ust, framerate, cechy osobnicze)

Wprowadzenie

- Największym obecnie wyzwaniem w systemach automatycznego rozpoznawania mowy jest stworzenie rozwiązania pozwalającego na skuteczne rozpoznawanie mowy w trudnych warunkach akustycznych

Wprowadzenie

- W celu poprawy skuteczności rozpoznawania mowy w warunkach szumowych podjęto badania nad dodawaniem dodatkowej modalności do systemów ASR – modalności wizyjnej

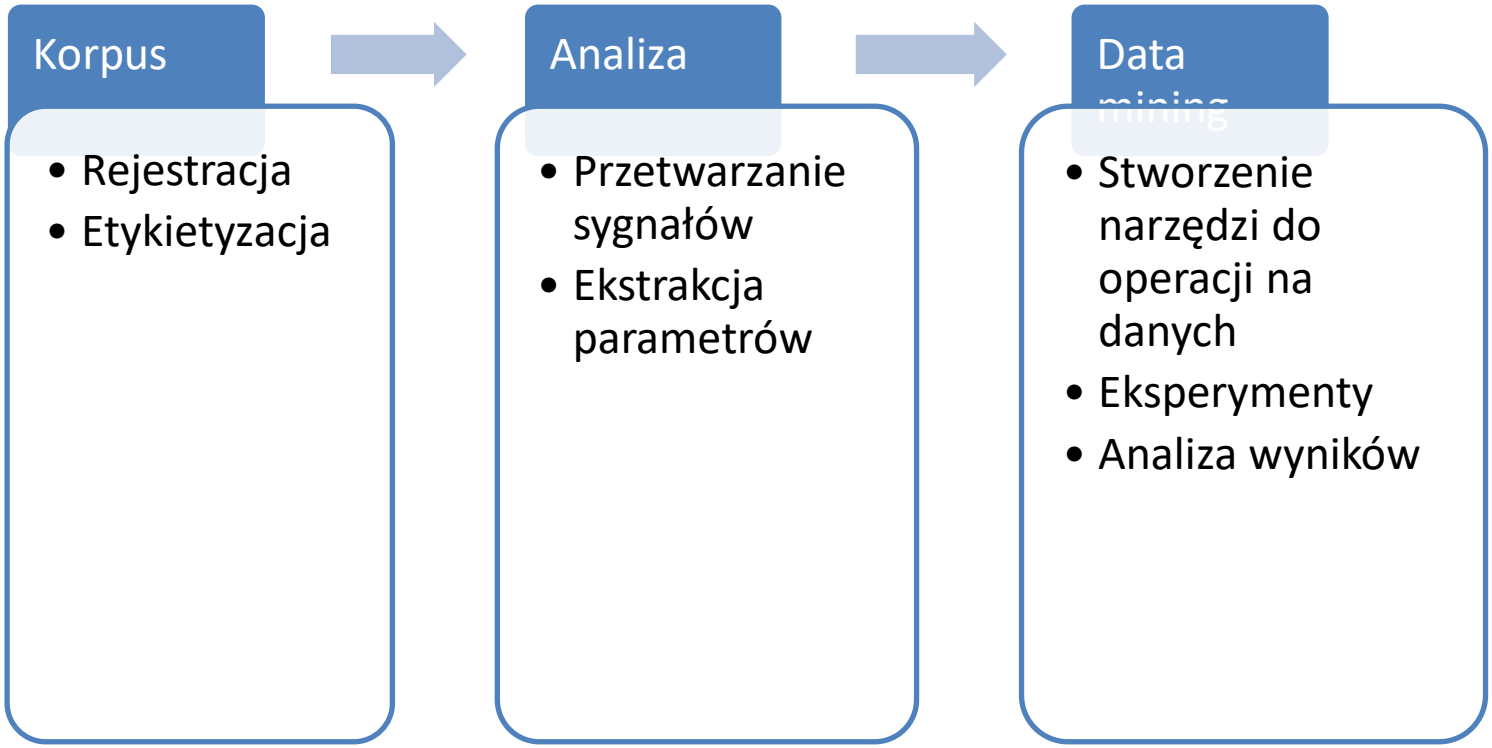
Systemy AVSR

- AVSR – Audio-Visual Speech Recognition
- Pionierskie prace dotyczące AVSR zostały zapoczątkowane przez Petajana w latach 1984 [1].
- Jedne z najnowszych badań dotyczą zastosowania sensora Kinect (kamera RGB, kamera głębi, macierz 4 mikrofonów) do zagadnienia AVSR. Prace prowadzone przez Galatasa et. al [2].

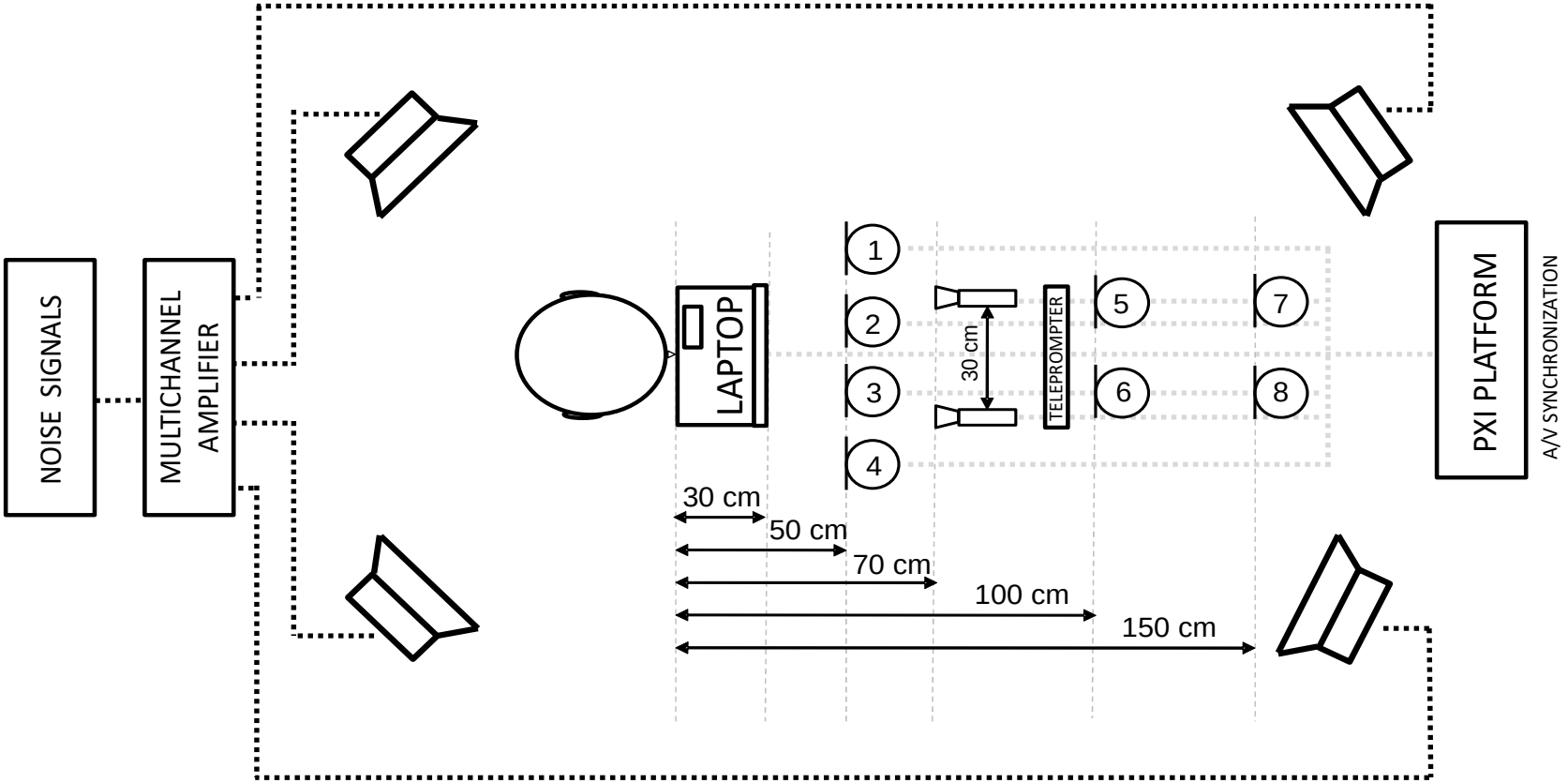
1. Petajan E., Automatic lipreading to enhance speech recognition, in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 1985, pp.40–47.
2. G. Galatas, G. Potamianos and F. Makedon, "Audio-visual speech recognition incorporating facial depth information captured by the Kinect," *(EUSIPCO), 2012,* pp. 2714-2717.

Tworzenie systemu AVSR

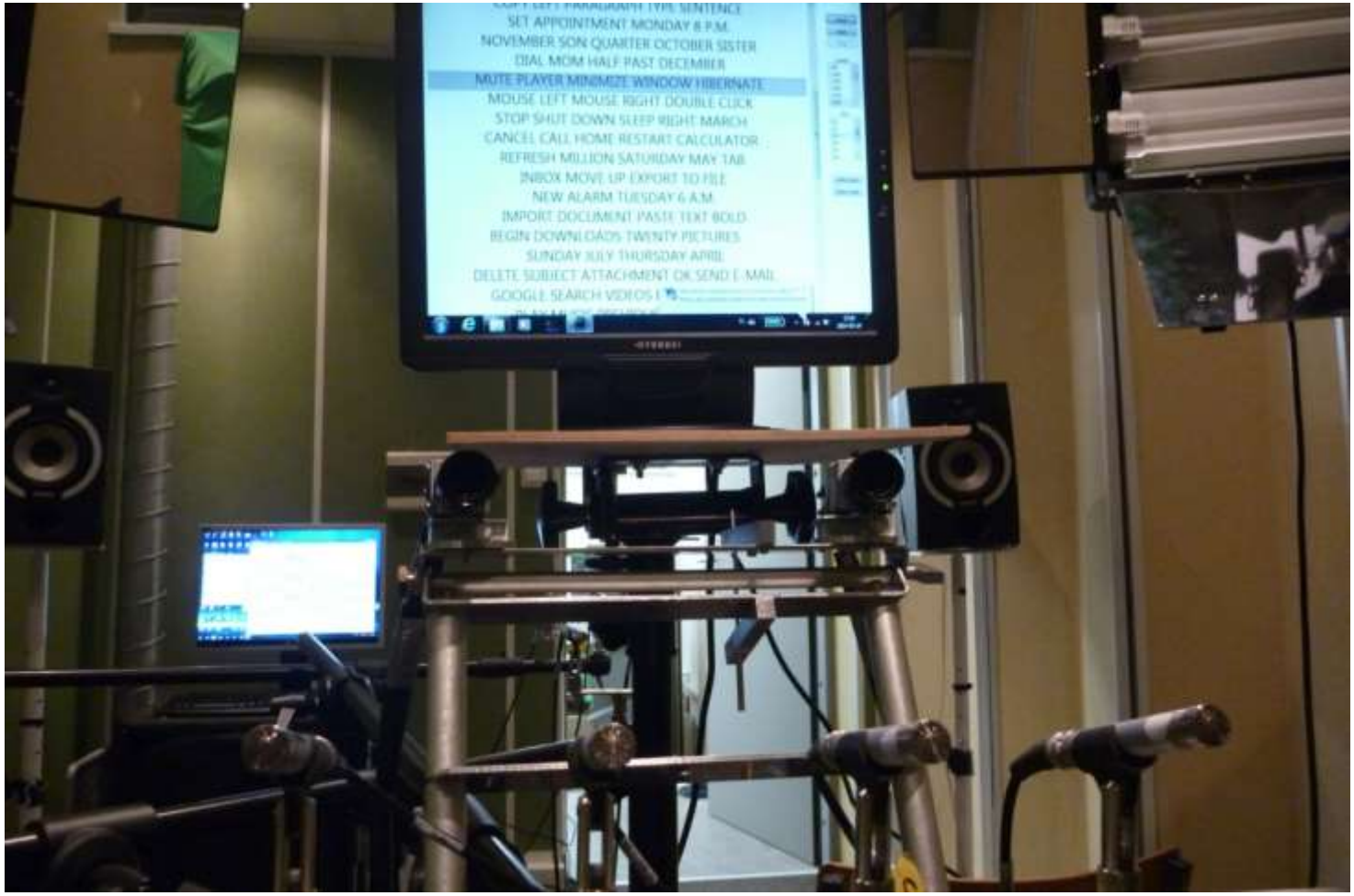
Kroki które należy wykonać:



Rejestracja korpusu



Rejestracja korpusu



Rejestracja korpusu

W sumie udostępnionych 35 mówców

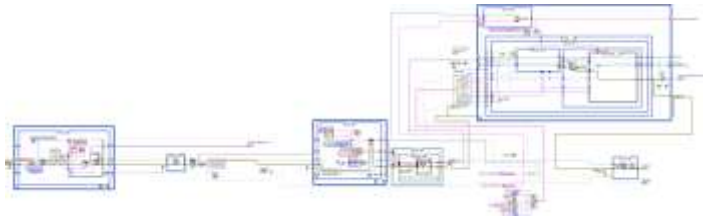
- **37 godzin** oznakowanego materiału audio-wizualnego
- 33 mężczyzn, 9 kobiet
- Równy podział na mówców „natywnych/nienatywnych”
- **3,75 TB danych**

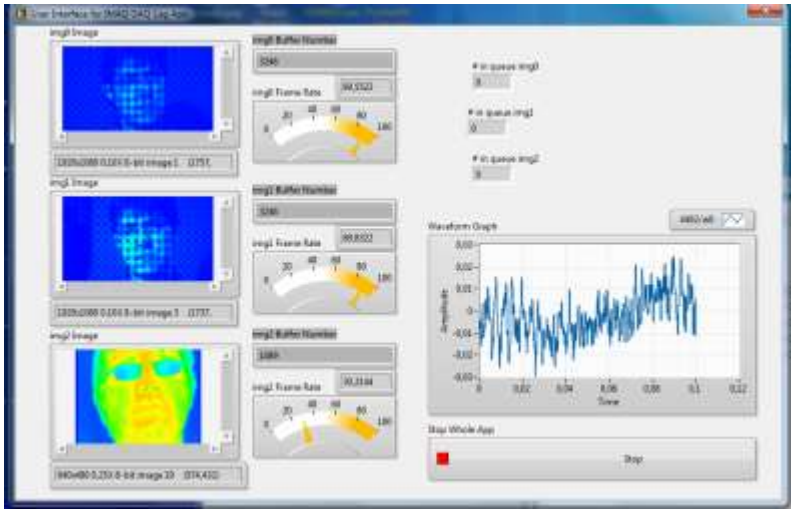
Inne korpusy

Database	Year	Spk.	Resolution	Framerate	Language material	Additional Features
TULIPS1	1995	12	100x75	30 fps	digits 1-4	no
DAVID	1996	123	640x480	30 fps	digits, alphabet, nonsense utterances	varying background
XM2VTS	1999	295	720x576	25 fps	3 sentences (digits and words)	head rotations, glasses, hats
BANCA	2003	52	720x576	25 fps	digits, name, date of birth and address	controlled, degraded and adverse conditions, impostor recordings
GRID	2005	34	720x576	25 fps	1000 command-like sentences	no
VIDTIMIT	2008	43	512x384	25 fps	10 TIMIT sentences	office noise and zoom
WAPUSK20	2010	20	640x480	48 fps	100 GRID sentences	stereoscopic camera, office noise
UNMC-VIER	2011	123	708x640 max	29 fps	12 XM2VTS sentences	varying speech pace, expressions, illumination, head poses and quality
KSM	2015	Max 42	1920x1080	100 fps	168 commands (isolated, sentences)	stereo camera, varying noise, word SNR, supplied with labels

Tools

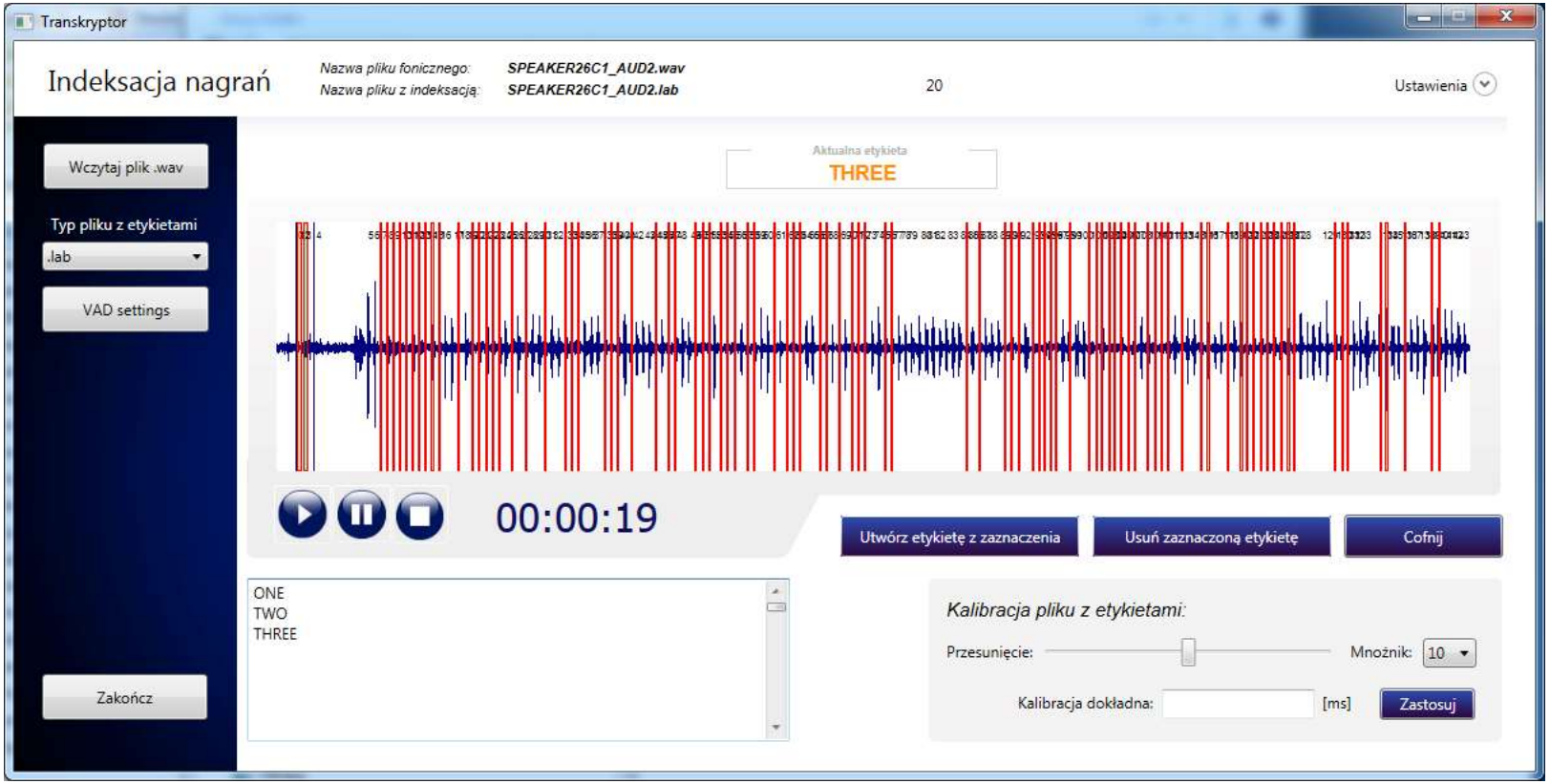
PXI Recording App – Block program developed in LabVIEW environment in order to register A/V streams at full bandwidth





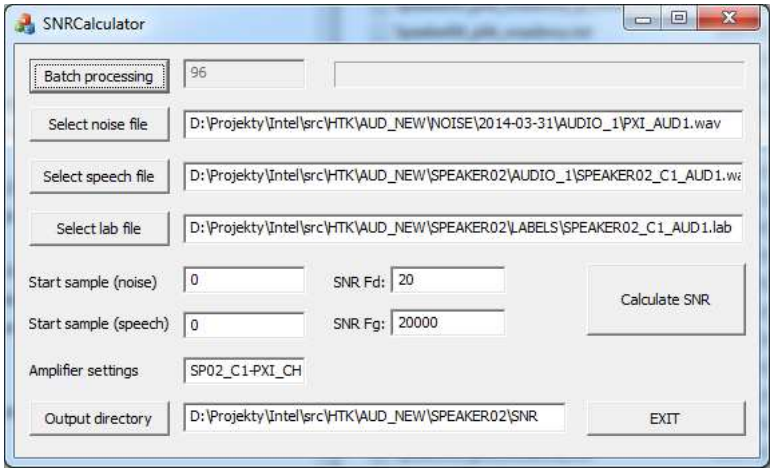
Training aspects

Transkryptor – c# program for labeling the material in database



Tools

Snr_calculator – c++ program for calculating SNR values given speech recordings, noise recordings and labels



Training aspects

Speaker dependence	Modalities	Fusion	Vocabulary	SNR
• Single speaker • Native-only • All speakers • Leave-one-out	• Audio only • Video only • A + V	• Concatena-tion • PCA	• Numerals • Commands • Sentences	• Clean • Noisy (three types) • Distant • Close

All possibilities - more than 200 combinations (4*3*3*4*4)

AVSR Results

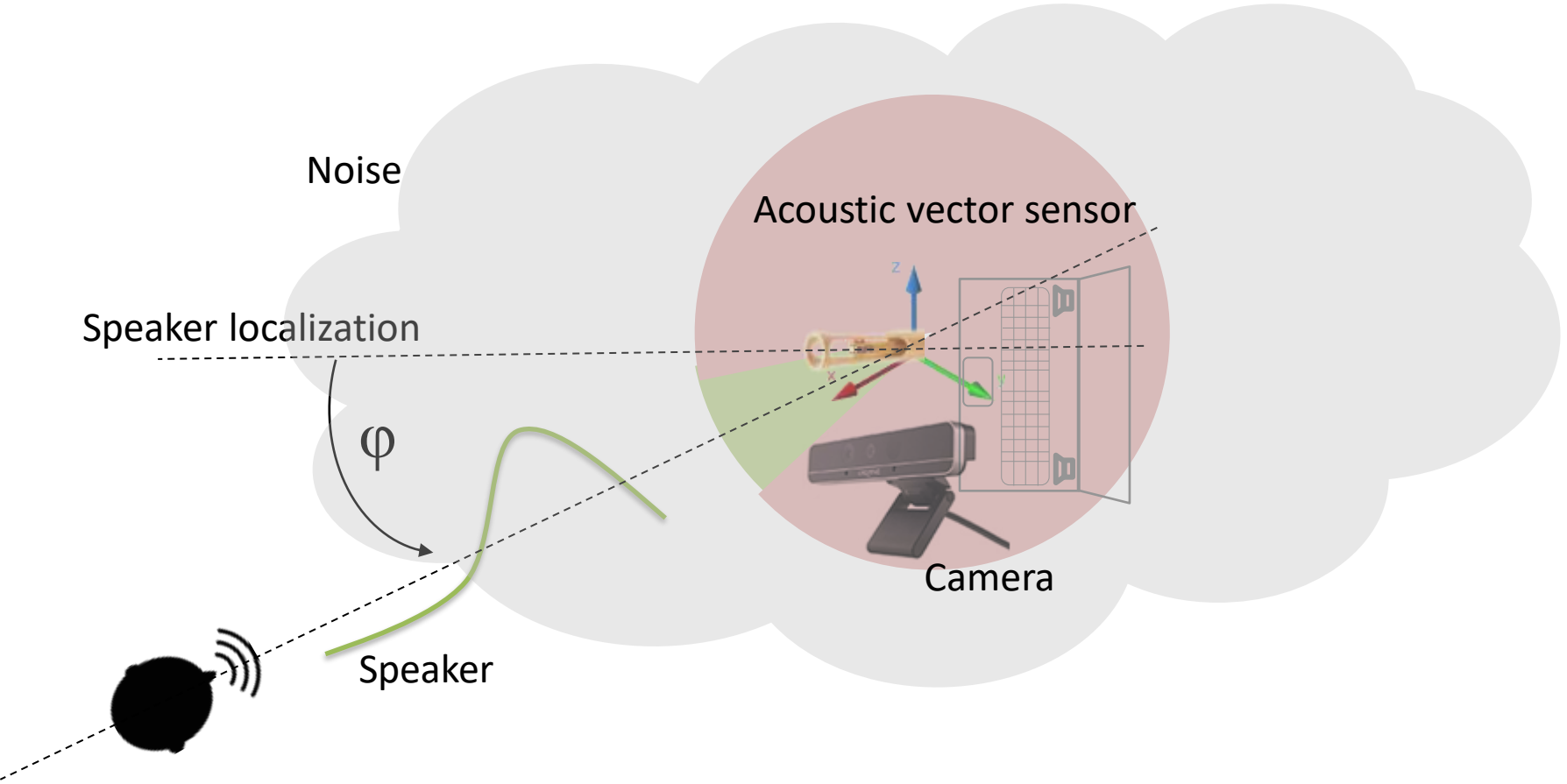
Self-developed AVSR system Accuracy in averaged WER:

Noise	Acoustic Features	Visual Features	WER [%]
none	MFCC (39)	none	21
babble	MFCC (39)	none	51
babble	MFCC (39)	AAM-Shape (10)	46

Comparison to the state-of-the-art ASR System created by Nuance and Intel

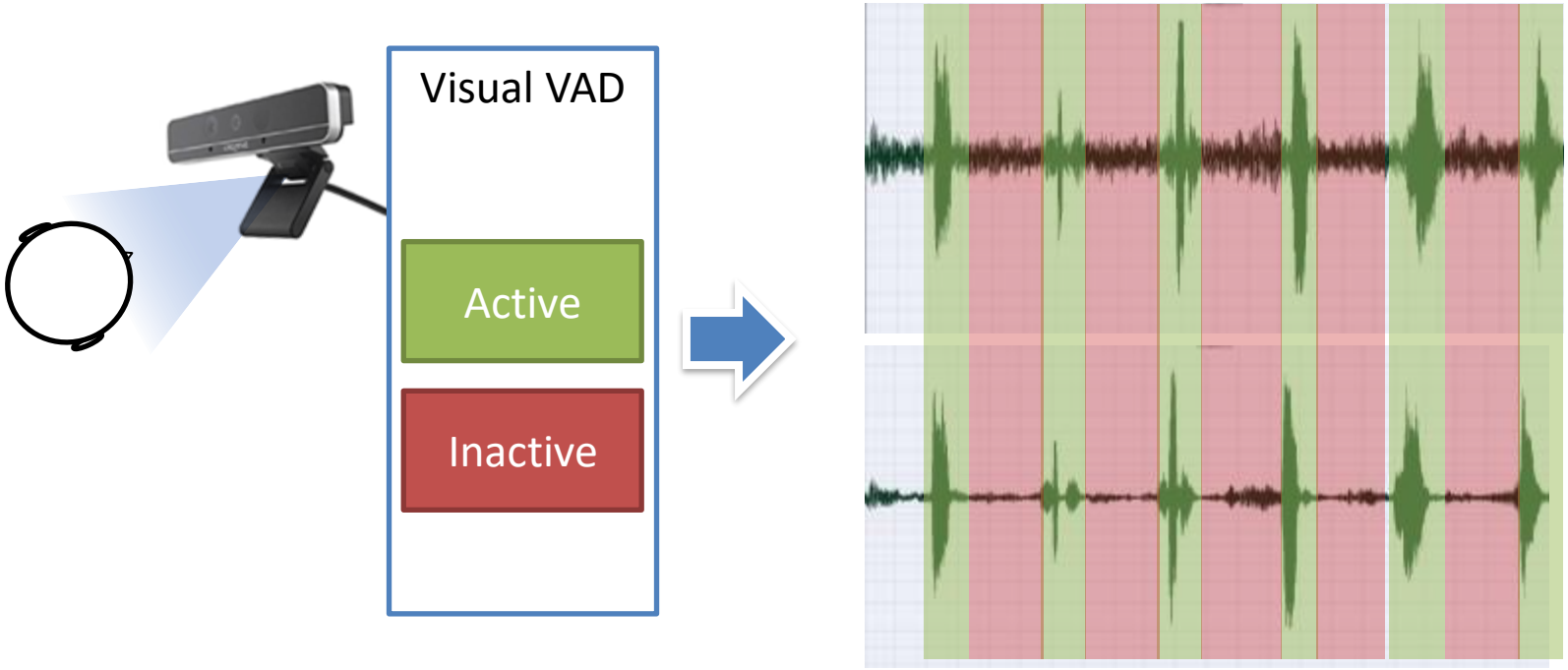
Conditions	MODALITY	RS-Unconstr.	RS-Constr.	RS-Unconstr.	RS-Constr.
clean	21	41,8	21,9	33,9	17,8
noisy	51	61,5	49,2	54,5	41,3
avsr	46	-	-	-	-
dictionary	-	EN-US		EN-GB	

New Approach

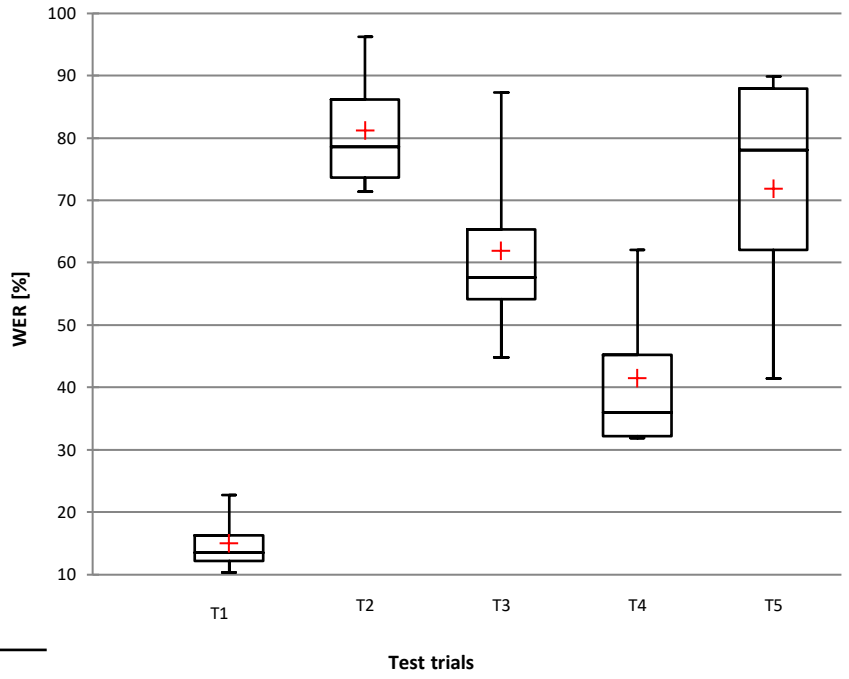


Czyżewski A., Bratoszewski P., Łopatka K., ***Kotus J.****, Szwoch G.,* ***Sposób i układ do poprawy jakości sygnału mowy w systemach rozpoznawania mowy i komunikacyjnych****, P.414116, data zgłoszenia: 23.09.2015, data przyznania patentu 20.11.2017*

New Approach



New Approach - Results



T	**Test conditions**
T1:	Clean environment, no signal processing methods
T2:	noisy environment, no signal processing methods
T3:	noisy environment, spatial filtration employed
T4:	noisy environment, spatial filtration and visual VAD
T5:	noisy environment, visual VAD

Bibliografia

- HTK Book: speech.ee.ntu.edu.tw/homework/DSP_HW2-1/htkbook.pdf
- Rabiner L., A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition
- Benesty, Springer Handbook of Speech Processing
- https://lidilem.univ-grenoble-alpes.fr/sites/lidilem/files/Mediatheque/Actualites/Com_doc/asr2019-lig-lidilem.pdf
- https://pionier.tv/wideo/czas-nauki/uczenie-maszynowe-system-rozpoznawania-mowy/
- https://www.is.umk.pl/~grochu/wiki/lib/exe/fetch.php?media=zajecia:nn_2018_1:nn-wyklad.pdf
- https://web.stanford.edu/~jurafsky/
- http://bazekon.icm.edu.pl/bazekon/element/bwmeta1.element.ekon-element-000171530952
- https://online.kitp.ucsb.edu/online/hearing17/schlueter/pdf/Schlueter_Hearing17_KITP.pdf