

# Technologie zpracování řeči pro indexování záznamů přednášek

**Honza Černocký**

**Speech@FIT, Ústav počítačové grafiky  
Fakulta informačních technologií VUT v Brně**



**CESNET Praha**

**14. 11. 2006**

# Plán

- Speech@FIT
- Technologie
  - Rozpoznávání s velkým slovníkem LVCSR
  - Detekce klíčových slov
  - Identifikace a verifikace mluvčího
- eLearning@Speech@FIT

# Speech@FIT – historie a lidé

- Vznikla 1997 na FEI VUT.
- Od roku 2002 na FIT VUT
- Honza Černocký – vedoucí
- Hynek Heřmanský – guru
- 4 učitelé FIT
- 7 doktorandů a zaměstnanců na projektech
- 2 studenti – vědecké síly.
- 4 lidé technická podpora



## HW a SW

- 2 IBM Blade centers s 26 „žiletkami“ po 2 CPU, další na cestě
- Dalších cca 100 počítačů v učebnách
- 12 TB diskového prostoru
- Profesionální správa
- HTK, Matlab, SGE
- STK, SNet.



# Speech@FIT – projekty

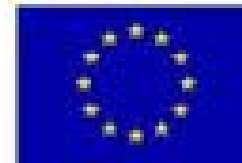
- 7 let zkušeností v projektech financovaných EU
  - Minulé: SpeechDat-E, SpeeCon, Multi-modal meeting manager (M4)
  - Současné: Augmented Multiparty interaction (AMI), CareTaker, AMIDA
- Účast v českých projektech (GAČR, FRVŠ, CESNET, ESF, atd.)
- Spolupráce s průmyslem (Siemens R&D, TEMIC, Lingea, CAMEA, ...) a státními institucemi (Ministerstvo obrany)

**Celkové prostředky pro výzkum na rok 2006: ~6 mil. Kč**



Ministerstvo obrany  
České republiky

**SIEMENS**



Information Society  
Technologies



**CAMEA**

**TEMIC**  
Speech Dialog Systems

LINGEA



# Evaluace – NIST

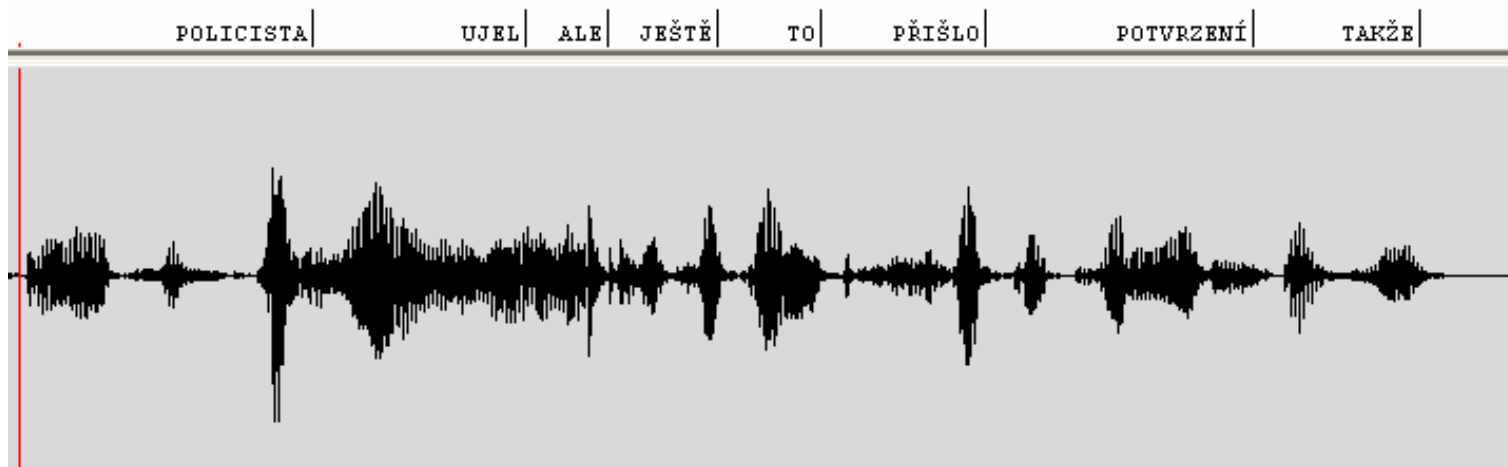
- „*Já jsem lepší než ti druzí*“ – irelevantní, pokud nejsou stejná data a evaluační metriky
- **NIST** – agentura vlády USA, <http://www.nist.gov/speech>
- Její řečová skupina organizuje **pravidelné evaluace řečových technologií** (rozpoznávání řeči, detekce řečníka, rozpoznávání jazyka).
- Všechny participující laboratoři obdrží **stejná** data a mají omezený čas na jejich analýzu a odeslání výsledků NISTu – **objektivní srovnání** výsledků.
- Výsledky a detaily jednotlivých systémů se diskutují na následném workshopu (většinou sponsorován NSA)
- Speech@FIT se evaluací účastní (Meeting recognition 2005, 2006, Language ID 2003, 2005, SpkVer 1998, 1999, 2006), STD 2006.

# Plán

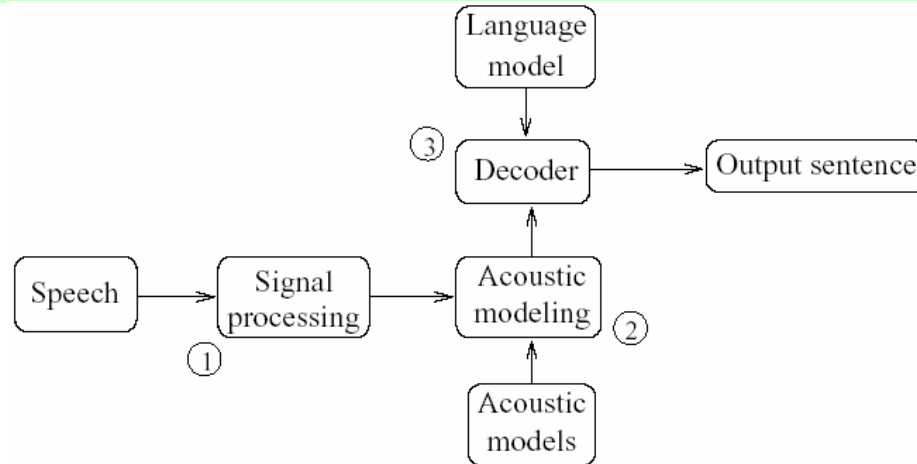
- Speech@FIT
- Technologie
  - Rozpoznávání s velkým slovníkem LVCSR
  - Detekce klíčových slov
  - Identifikace a verifikace mluvčího
- eLearning@Speech@FIT

# LVCSR

**Úkol LVCSR:** Celkový přepis – rozpoznávání plynulé řeči s velkým slovníkem (large vocabulary continuous speech recognition LVCSR)



# LVCSR – technické řešení



**Akustické modely:** určují pravděpodobnost, že vstupní signál náleží k akustickým jednotkám (fonémům) – trénování na velkých řečových databázích (in house, LDC)

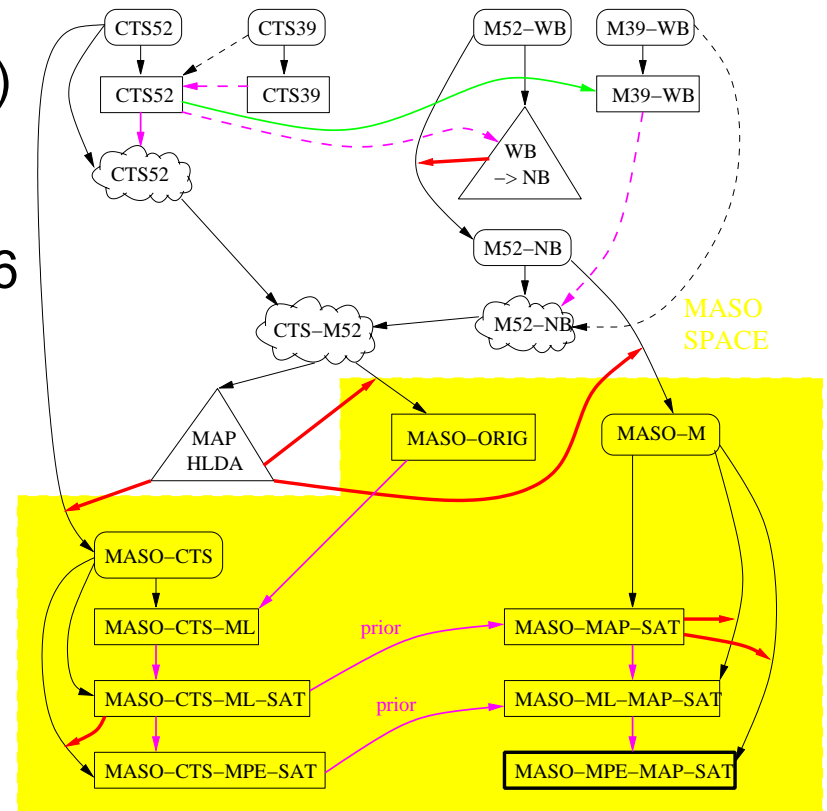
**Jazykový model:** určuje pravděpodobnost sekvencí slov „prezident Václav Klaus“ vs. „prezident Václav aplaus“ – trénování na velkých korpusech textu.

**Dekodér:** vyhledává optimální sekvenci slov.

# LVCSR ve Speech@FIT

- Zkušenost s vývojem systému pro rozpoznávání spontánní řeči v angličtině v rámci projektu **AMI**)
- Moderní techniky akustického modelování.
- 2. místo v NIST Spring RT 2006
- Vstup (front-end) pro další techniky (KWS, SpkID)

System	WER [%]
1.	24.1
<b>2. AMI</b>	<b>24.2</b>
3.	30.2



# LVCSR – jak dále

- Výzkum parametrů, transformací a modelů pro rozpoznávání (posterior features, HLDA, MLLT, SAT, MMI, MPE ...)
- Více práce na češtině

**Výstup LVCSR se nedá dobře číst, ale dá se v něm hledat**

**Zásadní nedostatek spontánních dat (řeč, text):**

- US-English: Switchboard-1: 2400 konverzací, 1993(!), , Fisher: 11699 konverzací, 2003, 2005, >1000 hodin.
- Čeština – téměř **NIC** ☹
- **S.O.S. česká spontánní data** – řeč, přepisy

# Plán

- Speech@FIT
- Technologie
  - Rozpoznávání s velkým slovníkem LVCSR
  - Detekce klíčových slov
  - Identifikace a verifikace mluvčího
- eLearning@Speech@FIT

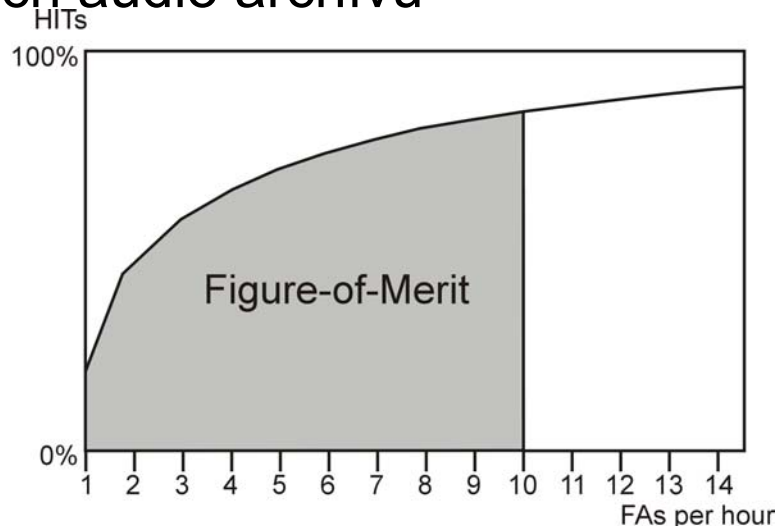
# Detekce klíčových slov - KWS

**Úkol KWS:** detekce klíčových slov nebo frází:

- On-line pro sledování např. odposlechů v reálném čase.
- Off-line pro prohledávání velkých audio archivů (dolování informace v audio).

## Používané techniky:

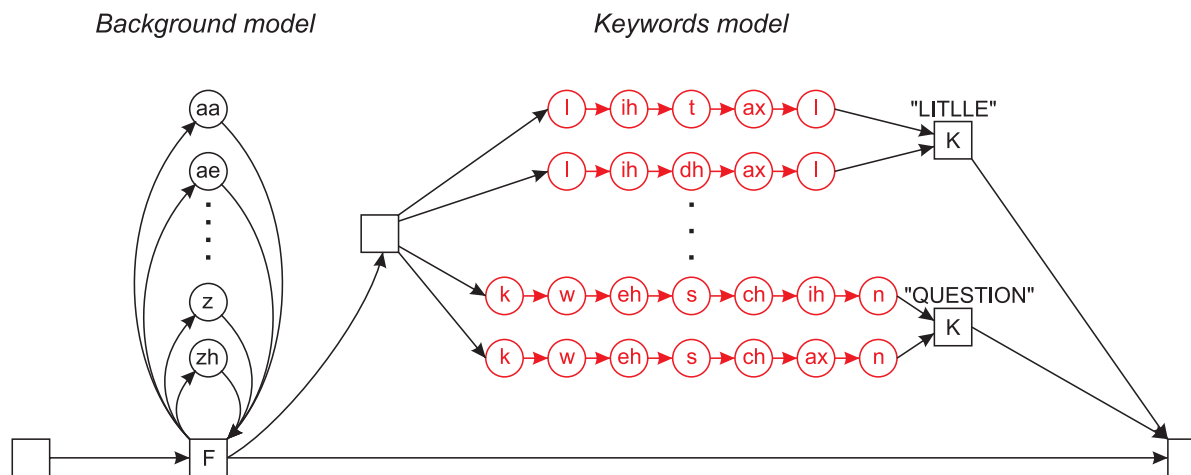
- Prohledávání výstupu LVCSR
- Akustický KWS
- KWS ve fonémových grafech.



**Problém není detekovat,  
ale odmítat falešné  
záchyty**



# Přístup 2: Akustický KWS



☺ řeší problém OOV

☺ cizí slova

☺ neznámý jazyk

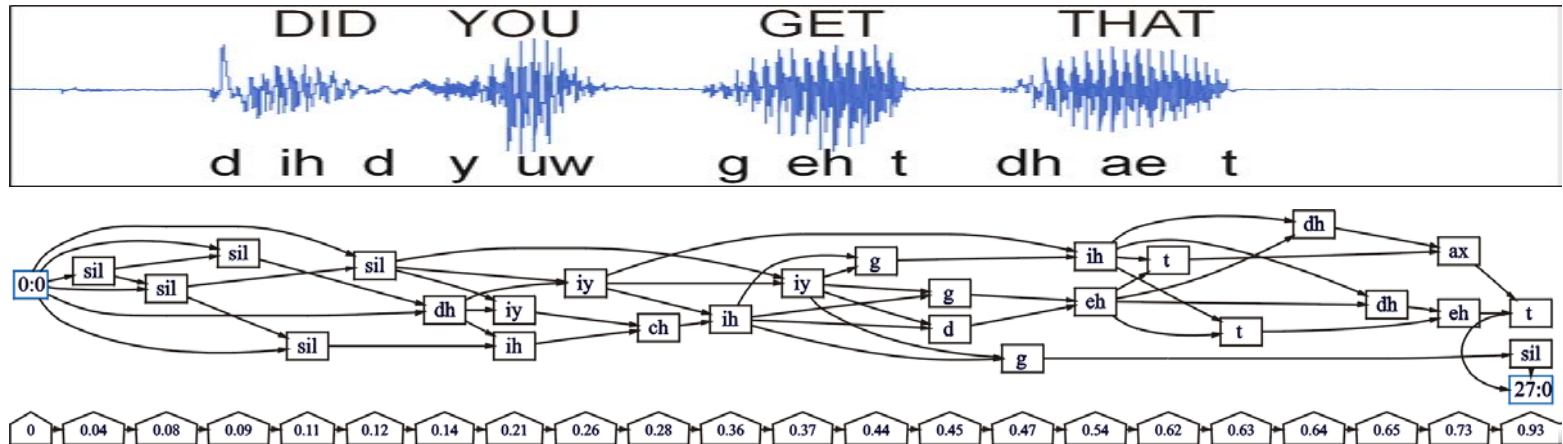
☺ on-line

☹ Nemožná indexace –  
nutnost projít vše

☹ Pouze cca 0.01xRT

☹ Nemá sílu LM

# Přístup 3: KWS na fonémových grafech



😊 řeší problém OOV a cizích slov

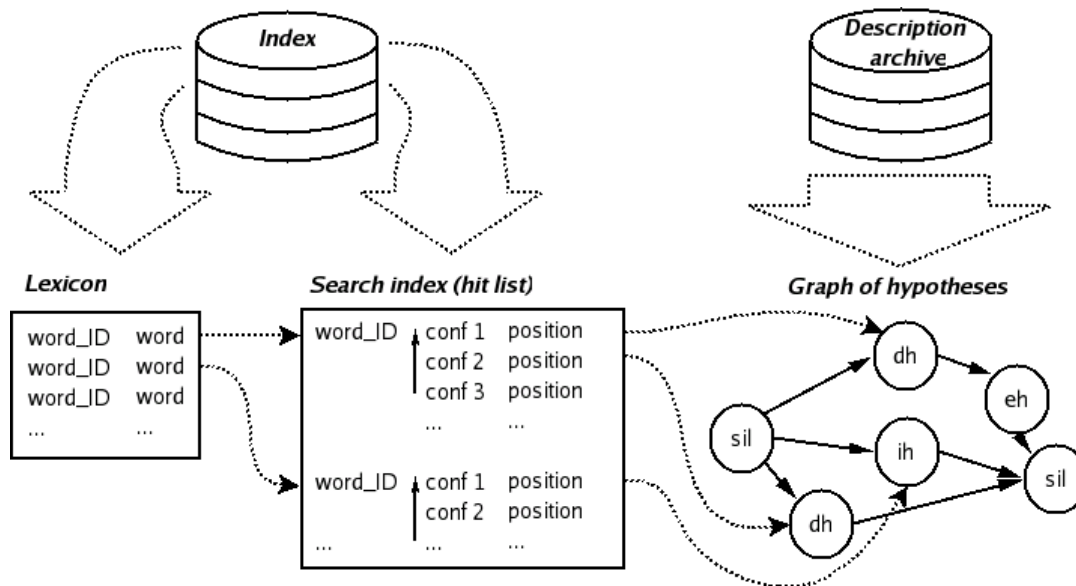
😊 možnost indexace sekvencí fonémů

😊 kombinace s LVCSR

☹ Pomalejší než LVCSR

☹ Složitější search engine

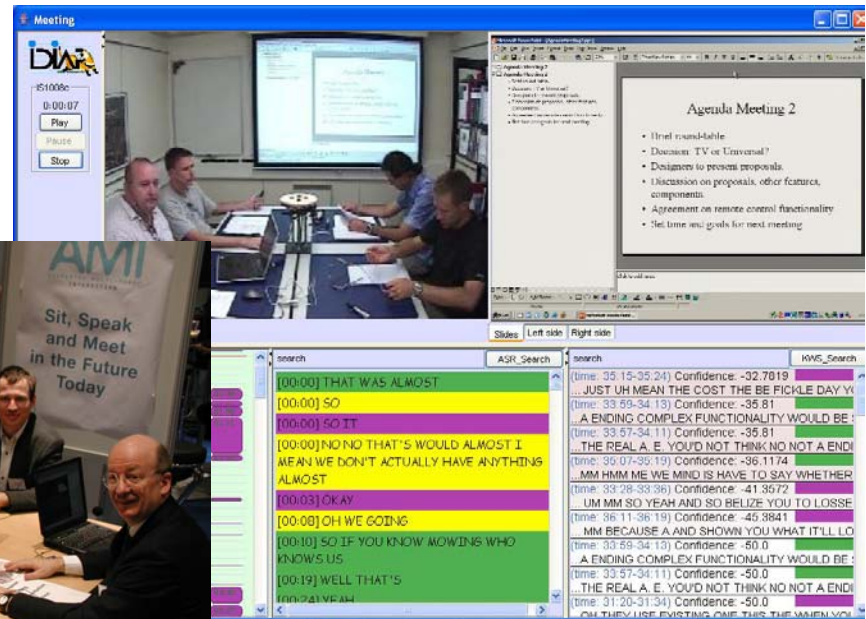
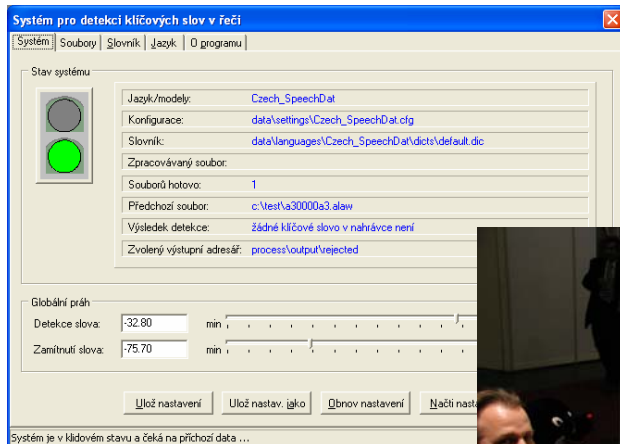
# Indexování a vyhledávání



- Indexace a vyhledávání v LVCSR včetně **složitých dotazů**
- Vyhledávání ve fonémových grafech

# KWS – implementace

- Akustický KWS (3 jazyky) – nasazen u složky MO ČR.
- Real-time zpracování pro meeting room – CeBIT.
- Integrace s multimodálním prohlížečem JFerret - **demo**



# KWS – jak dále

- Zlepšování základních technologií – LVCSR, rozpoznávání fonémů, atd.
- Sémantická podpora akustického KWS.
- NIST Spoken Term Detection (STD) evaluace **ted'**:
  - US angličtina: telefonní řeč, broadcast news, meetingy. LVCSR + fonetické vyhledávání.
  - Arabština: fonetické vyhledávání
  - Budou opět výsledky srovnatelné se světem, workshop ve Washingtonu D.C. 14-15. prosince.

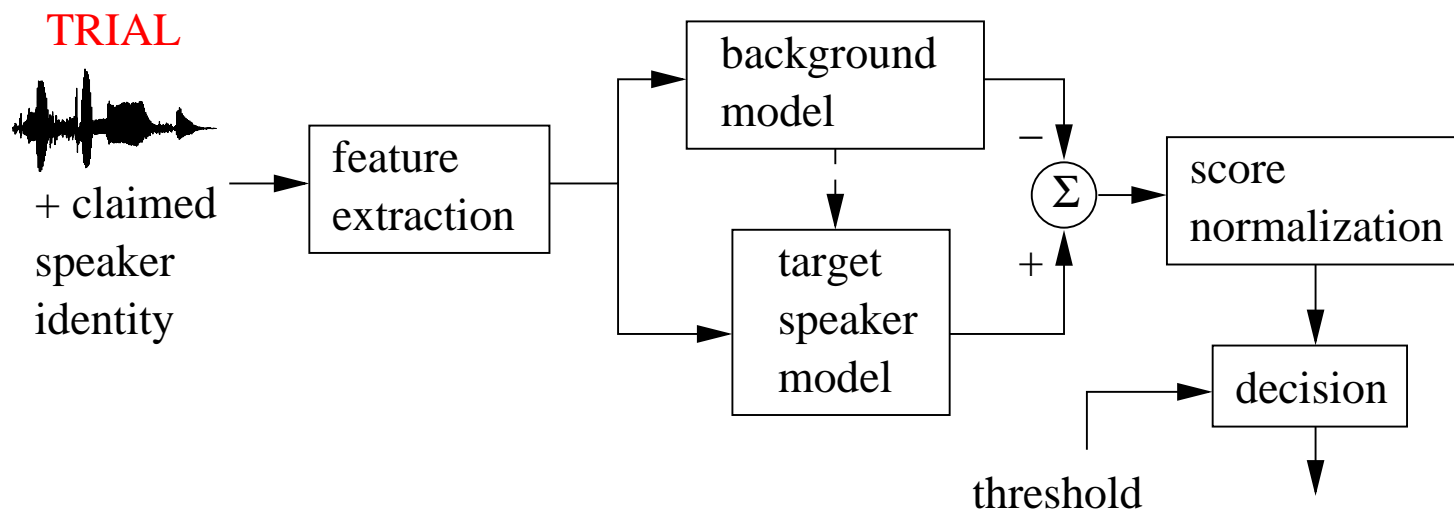
# Plán

- Speech@FIT
- Technologie
  - Rozpoznávání s velkým slovníkem LVCSR
  - Detekce klíčových slov
  - **Identifikace a verifikace mluvčího**
- eLearning@Speech@FIT

# Rozpoznávání mluvčího – SpkID, SpkVer

**Úkol SpkID:** přiřadit řečový segment k jednomu z  $N$  mluvčích nebo prohlásit, že to není žádný.

**Úkol SpkVer:** ověřit předpokládanou identitu „Je to opravdu pan Novák?“

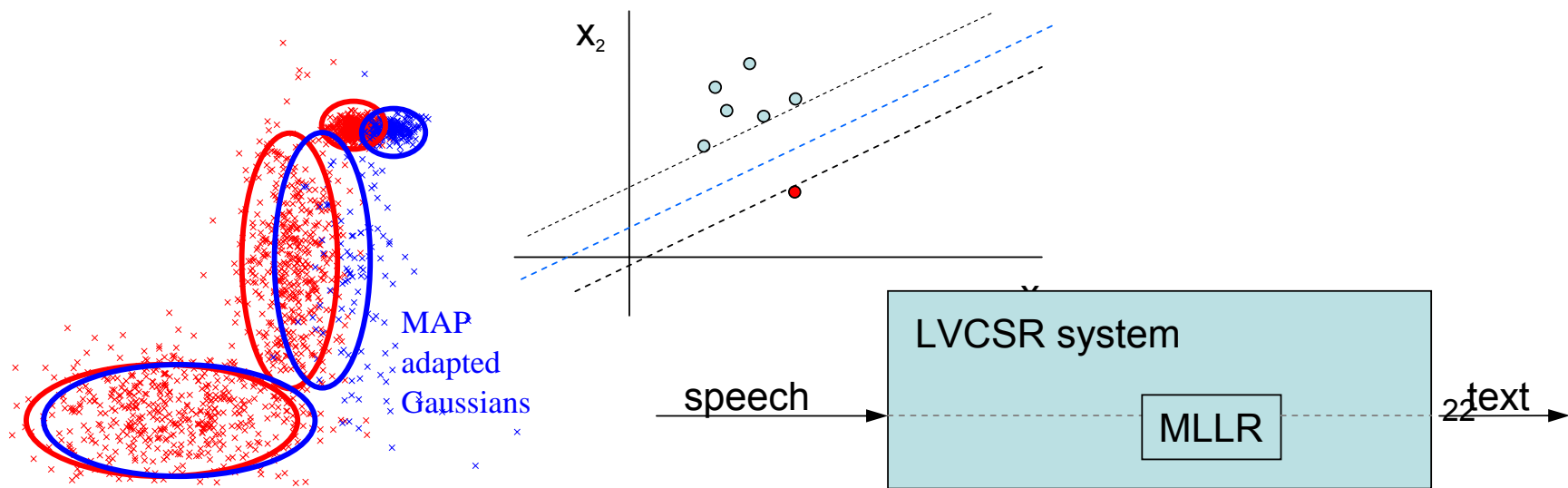


**TARGET/  
IMPOSTOR**

# SpkVer – technické řešení

## Kombinace tří dvojic extrakce příznaků/klasifikátor:

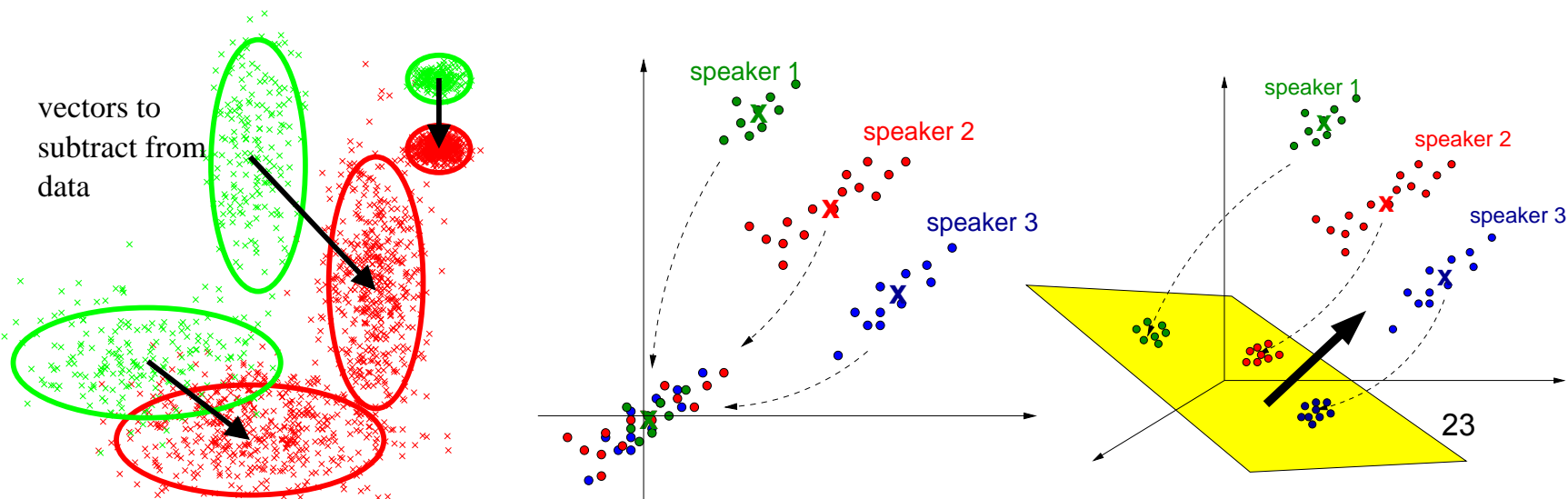
- Gaussian Mixture model (GMM) klasifikující Mel-frekvenční cepstrální koeficienty MFCC.
- Support Vector Machine (SVM) klasifikující supervektory středních hodnot z GMM
- SVM klasifikující adaptační matice z LVSCR



# SpkVer – adaptace na kanál

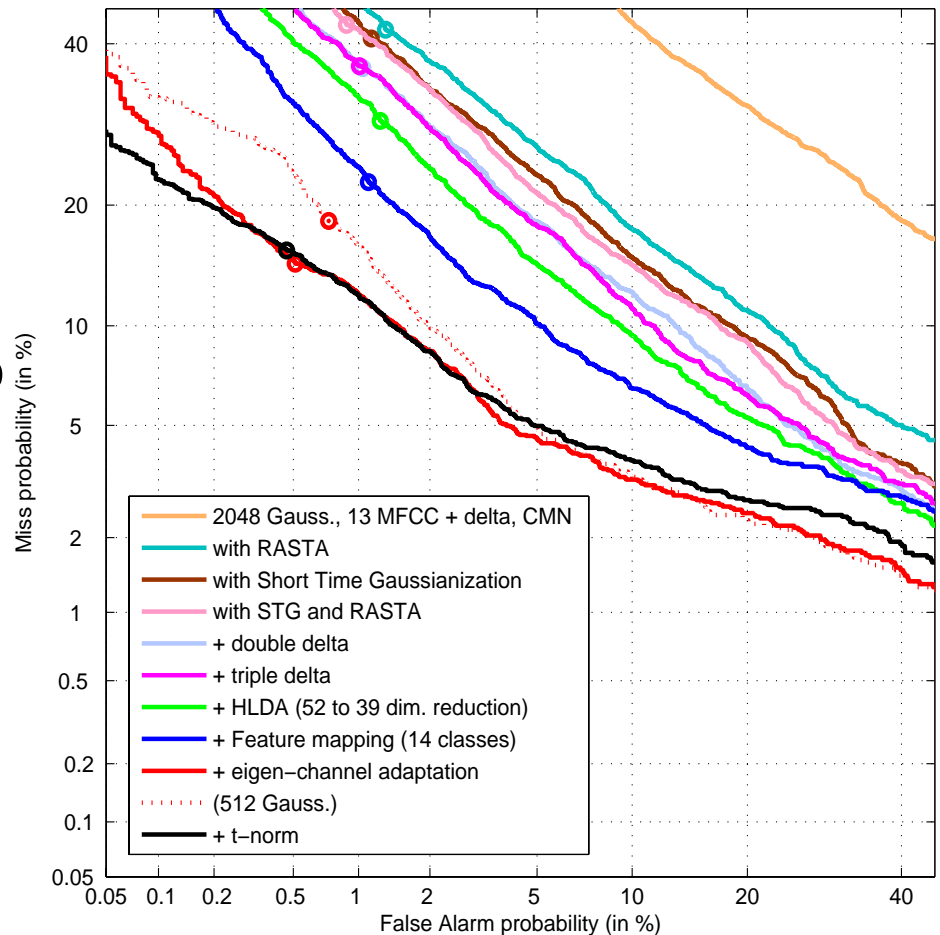
**Tři techniky pro odstranění závislosti na kanálu (pevná linka, mobil, satelitní telefon, IP) a na obsahu promluvy:**

- *Feature mapping* – trénování GMM pro jednotlivé typy kanálů, odečítání středních hodnot.
- *Eigen channel normalization* – posun modelů ve směru největší variability v závislosti na čemkoliv, co není informace o mluvčím.
- *Nuisance attribute projection (NAP)* – promítání dat do roviny, kde se neprojevuje variabilita kanálu.



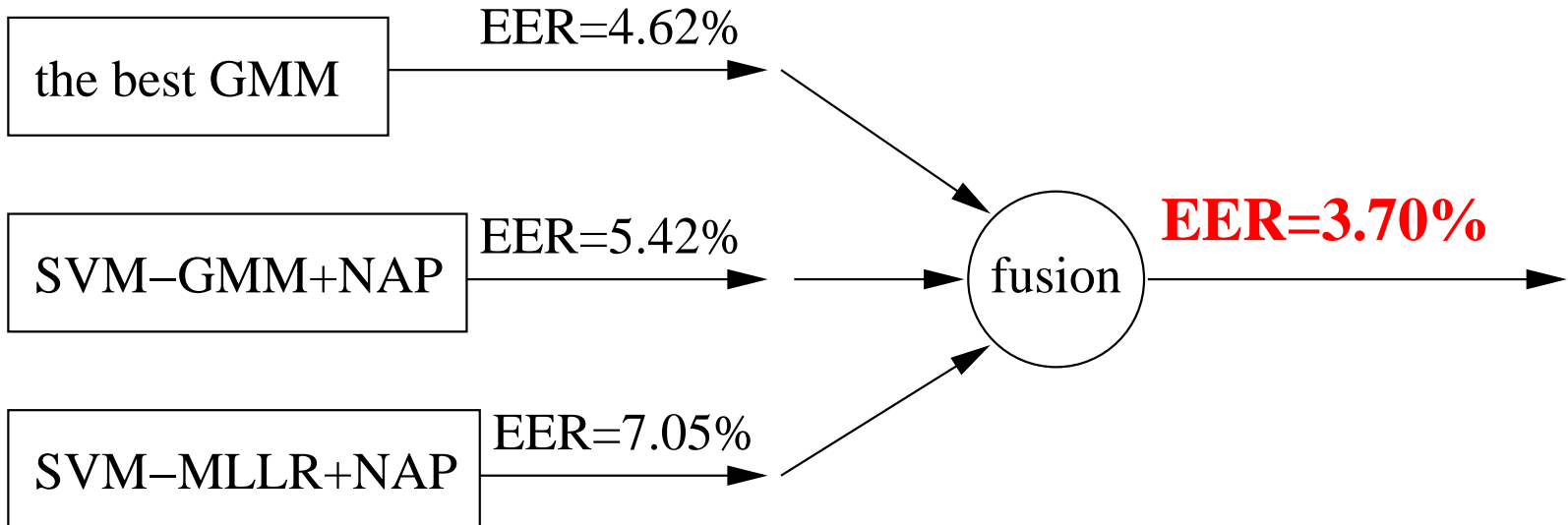
# Vyhodnocení – DET křivky

- Falešný záchyt (False alarm) – systém ověřil identitu imitátora jako cílového mluvčího.
- Odmítnutí (Miss) – systém odmítl pravého mluvčího.
- Equal Error Rate (EER) – pracovní bod, kde  $p(\text{FA})=p(\text{Miss})$ .
- Čím je křivka blíže počátku, tím je systém lepší



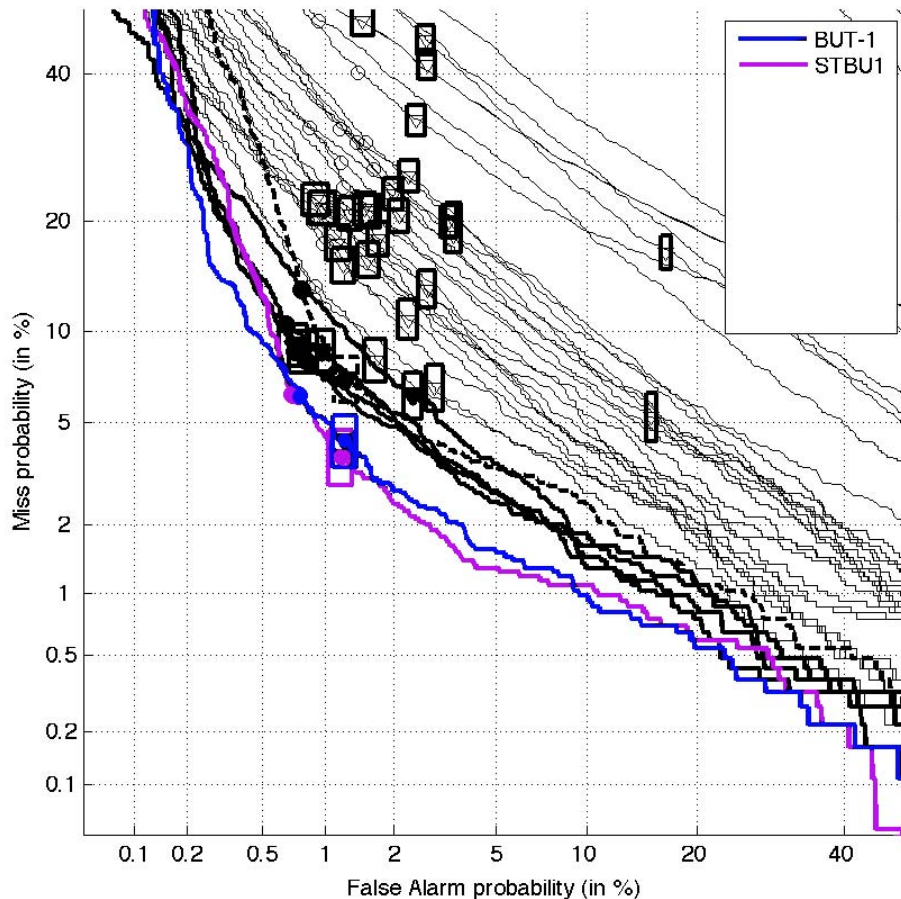
# Výsledky (NIST 2005)

- GMM systém má nejlepší výsledky ze 3 kombinací extrakce příznaků/klasifikátor
- Fúzí tří systémů se výsledky dále zlepší – komplementarita GMM a SVM !



# Výsledky NIST 2006

COMPOSITE 2006 (1conv4w-1conv4w): DET 1 All Trials (Common Test) Primary Systems



NIST data – **ukázka**

m M7004 NFIM A T  
9.5120265e+00  
m M7004 NJTZ B F  
-9.9731765e+001

# SpkID/SpkVer – jak dále

- Použití prosodických příznaků
- fonotaktika, neuronové sítě, atd.
- Další využití LVCSR
- Úvahy o back-ground modelech při nekoherentních trénovacích a testovacích signálech.



# Plán

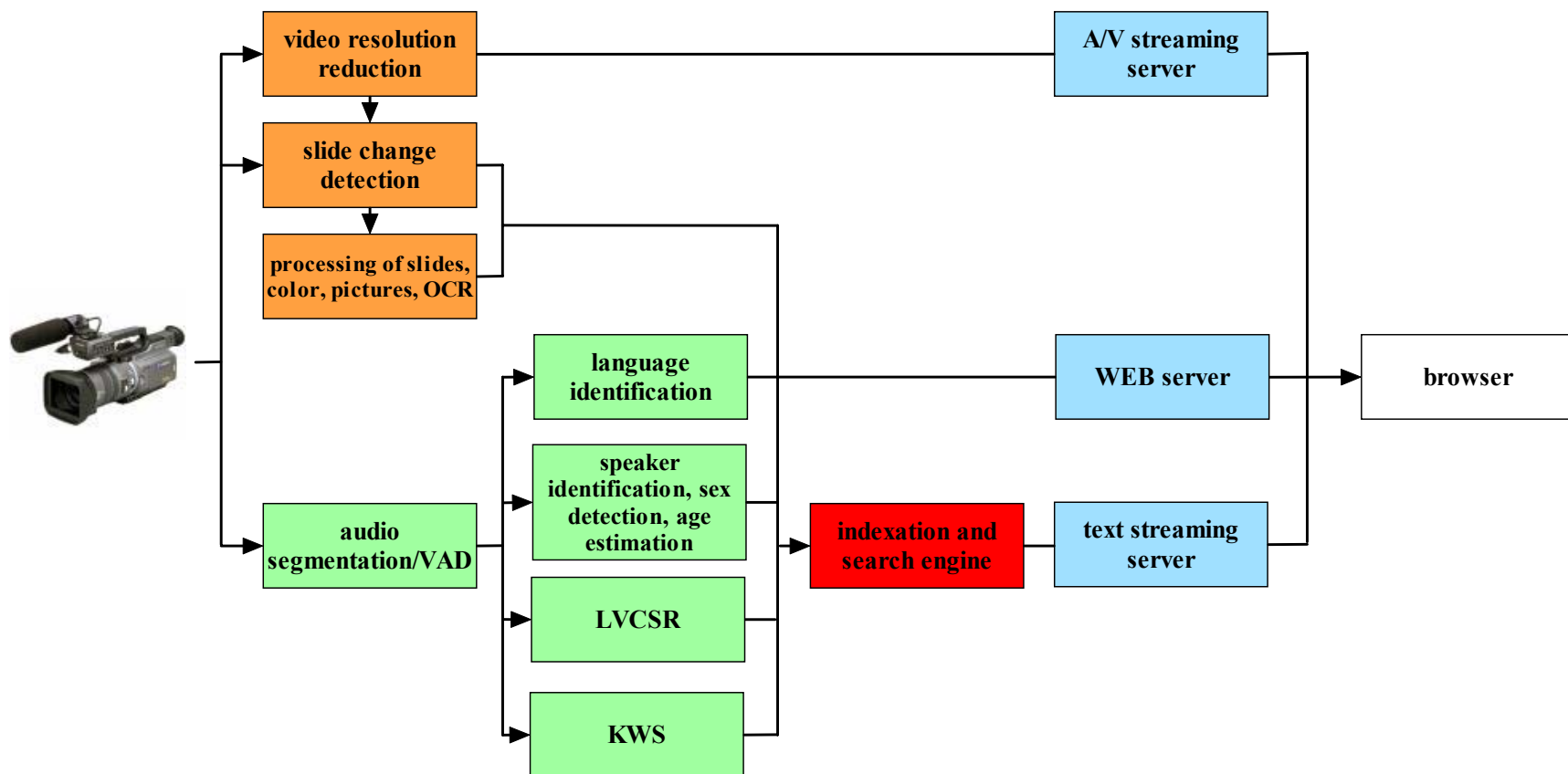
- Speech@FIT
- Technologie
  - Rozpoznávání s velkým slovníkem LVCSR
  - Detekce klíčových slov
  - Identifikace a verifikace mluvčího
- eLearning@Speech@FIT

# Řečové technologie v eLearningu

Usnadnění přístupu uživatelů k eLearningovým materiálům:

**Namísto poslouchání celé přednášky pro nalezení požadované pasáže bude možné tuto rychle a efektivně vyhledat v audiu podle obsahu a mluvčího.**

# The goal



# Předpokládané nasazení

- Fakulta informačních technologií VUT – cca **150 hodin** streamovaných a nahrávaných přednášek **týdně**.
- u dalších partnerů CESNET (jednání s FI MU).
- výhledové nasazení jako **služby** pro “počítačově méně zdatné” uživatele.

# TO DO's

- Výzkumná a vývojová práce na řečových technologiích.
- Presentační prostředí JFerret => **PRASE** (Presentation as Synchronized Experience)
- Integrace s videem (viz přednáška Stani Sumce)
- Lepší zpracování slajdů
- Ne toliko věda, ale hodně „černé“ programátorské práce !

# Závěry

Speech@FIT má co říci ke zpracování a vyhledávání v ukládaném a streamovaném audio/video pro:

- eLearning
- ale i další síťové aplikace zajímavé pro CESNET.

**Děkuji CESNET za pozvání na tento seminář a těším se na Vaše dotazy.**