

Počítačové zpracování řeči

Jan Nouza

Počítačové zpracování řeči

Vyučující: Prof. Ing. Jan Nouza, CSc.

Ústav: ITE

Rozsah: 2 + 2

Forma výuky:

- a) Přednášky
- b) cvičení – formou domácích úloh
- d) malý závěrečný projekt -> známka

Literatura:

- I. Nouza J., Koldovský Z., Vích R. (editoři): Řeč a počítač. TUL 2009.
- II. Psutka J., Müller L., Matoušek J., Radová V.: Mluvíme s počítačem česky. Academia Praha, 2006
- III. Huang X., Acero A., Hon H.-W.: Spoken Language Processing. A Guide to Theory, Algorithm and System Development. Prentice Hall. New Jersey 2001)

Podklady k přednáškám a cvičením - elearning

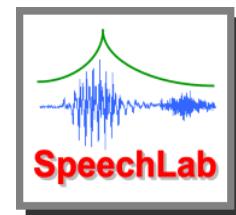
Proč speciální předmět o řeči?

Řečové technologie

- patří k nejdynamičtěji se rozvíjejícím oborům Umělé inteligence
- umožňují zautomatizovat řadu činností, které dosud vyžadovaly aktivní podíl člověka a jeho intelekt
- umožňují vytěžovat informace z mluvených dokumentů:
TVR pořady, záznamy jednání (parlament, soudy, zastupitelstva, firemní porady,)
- hlasová interakce je nejpřirozenější způsob mezilidské komunikace, hodí se i pro HCI (Human Computer Interaction)

SpeechLab TUL

patří mezi vedoucí
výzkumná pracoviště v ČR (známá i v Evropě)



Studijní linie v oboru IT - UI

Počítačové zpracování signálů	PZS	2+2	5	zk.	Málek J. ITE
Počítačové zpracování řeči	PZR	2+2	5	zk.	Nouza J. ITE
Pokročilé metody rozpoznávání řeči	PMR	2+2	5	zk.	Nouza J. ITE
Počítačové vidění	PVI	2+2	5	zk.	Chaloupka J. ITE
Zpracování obrazu	ZPO	2+2	5	zk.	
Interakce člověka s počítačem	ICP	2+2	5	zk.	Jeníček J. ITE
Počítačová lingvistika	PLI	2+2	5	zk.	Červa P. ITE
Biologické signály	BSI	2+2	5	zk.	Koldovský Z ITE

Co jsou moderní hlasové technologie?

Komunikace mezi lidmi

- **Telekomunikační služby** – rozmach mobilní telefonie

Komunikace mezi člověkem a počítačem

- **Počítačová syntéza řeči** – „počítač mluví“
hlášení na nádražích, čtení textových dokumentů, např. pro nevidomé, ...
- **Počítačové rozpoznávání řeči** – „počítač poslouchá“
glasové vytáčení, ovládání PC, glosové vyhledávání, diktát do počítače, přepis zpráv, rozhovorů, automatický přepis jednání
- **Dialog s počítačem** – „počítač naslouchá a mluví“
automatické informační systémy, rezervace po telefonu, SIRI, ...
- **Rozpoznávání řečníka** – “počítač zjišťuje kdo mluví“
identifikace osob podle hlasu, např. pro zabezpečení transakcí, při odhalování kriminálních činů, ...

Co umí a neumí hlasové technologie?

Hlasová syntéza (TTS – Text-to-Speech)

- **cíl:** přečíst libovolný text, včetně zkratek a číslic

Příklad:

První čtení proběhlo 17. 8. 2005 na 46. schůzi. Návrh zákona přikázán k projednání výborům (usnesení č. 1822). Petiční výbor návrh zákona neprojednal. Výbor pro veřejnou správu, regionální rozvoj a životní prostředí projednal návrh zákona a vydal 6. 12. 2005 usnesení doručené poslancům jako tisk 1056/1.

Ukázky: systém z roku 2005 výstup DP F. Kynycha (2020)



Co umí a neumí hlasové technologie?

Rozpoznávání řeči

- umí:
 - a) spolehlivě rozpoznávat slovní povely
 - b) téměř stoprocentně zvládat úlohu diktování
 - c) dobře přepisovat plynulou přirozenou řeč
- hlavní komplikace:
 - a) hlučné prostředí, nestacionární šum
 - b) spontánní řeč (často porušuje pravidla jazyka)
 - c) dialekty, nestandardní výslovnost,

Proč je rozpoznávání těžší než syntéza?

Hlasová syntéza

Řeč generovanou počítačem vnímá člověk vybavený intelektem a schopností překlenout případné chyby, domýšlet si souvislosti, apod.

Rozpoznávání řeči

Lidskou řeč, která je velmi složitá a variabilní, analyzuje méně dokonalý stroj bez vlastní inteligence.

Klíčové problémy rozpoznávání řeči:

- rozsáhlý slovník** (v češtině více než 1 milion slov a tvarů)
- přirozená řeč je plynulá** (mezi slovy nejsou pauzy)
zavolejmi prosím zítra ve čternácti hodinám
- každý člověk mluví jinak** (jiná výslovnost, výška a barva hlasu, intonace, volba slov, momentální stav,)
- řeč se nikdy neodehrává v úplném tichu** (mikrofon vždy snímá i okolní ruch, hluky, řeč jiné osoby, atd.)

Laboratoř zpracování řeči na TUL

Výzkumný tým:

- založen **1993**, nyní **8 pracovníků**
- účast v **národních a evropských** programech výzkumu
- výsledky: nejen publikace, ale i řada **praktických aplikací**

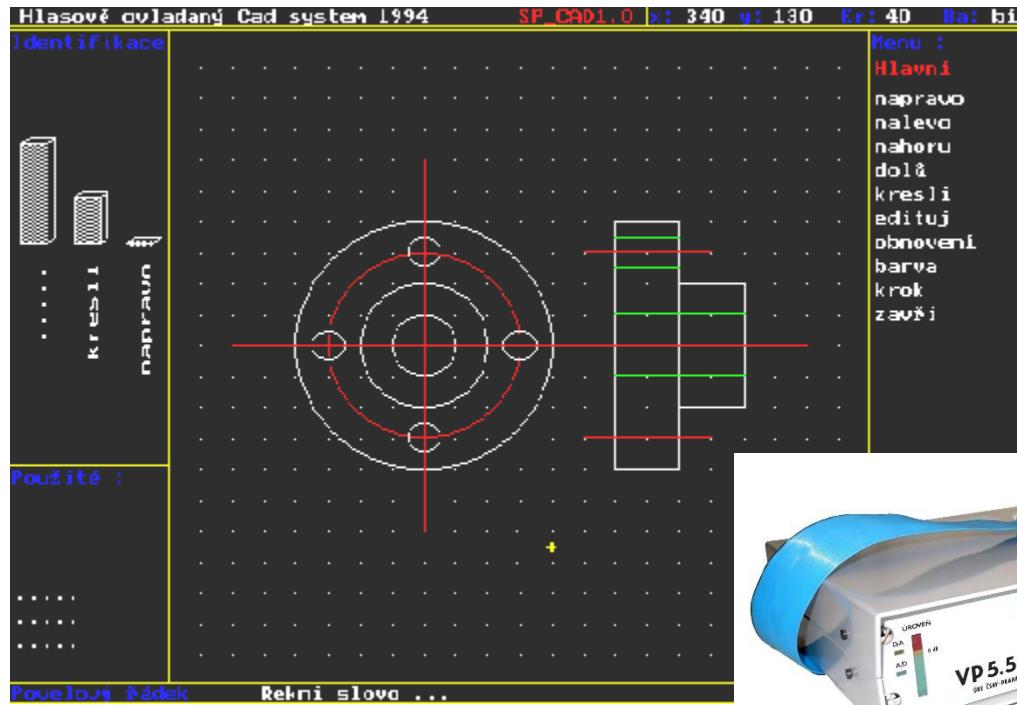
Hlavní oblasti výzkumu:

- **rozpoznávání řeči** se zaměřením na češtinu a evropské jazyky
- **automatické přepisy** (on- a off-line) zvukových streamů a záznamů
- **rozpoznávání osob** podle hlasu
- **audiovizuální** komunikace (rozpoznávání i syntéza s podporou vizuální informace)
- hlasové technologie na **pomoc postiženým**

Možnosti zapojení studentů:

- Témata pro semestrální, diplomové, disertační práce
- Placená spolupráce na projektech

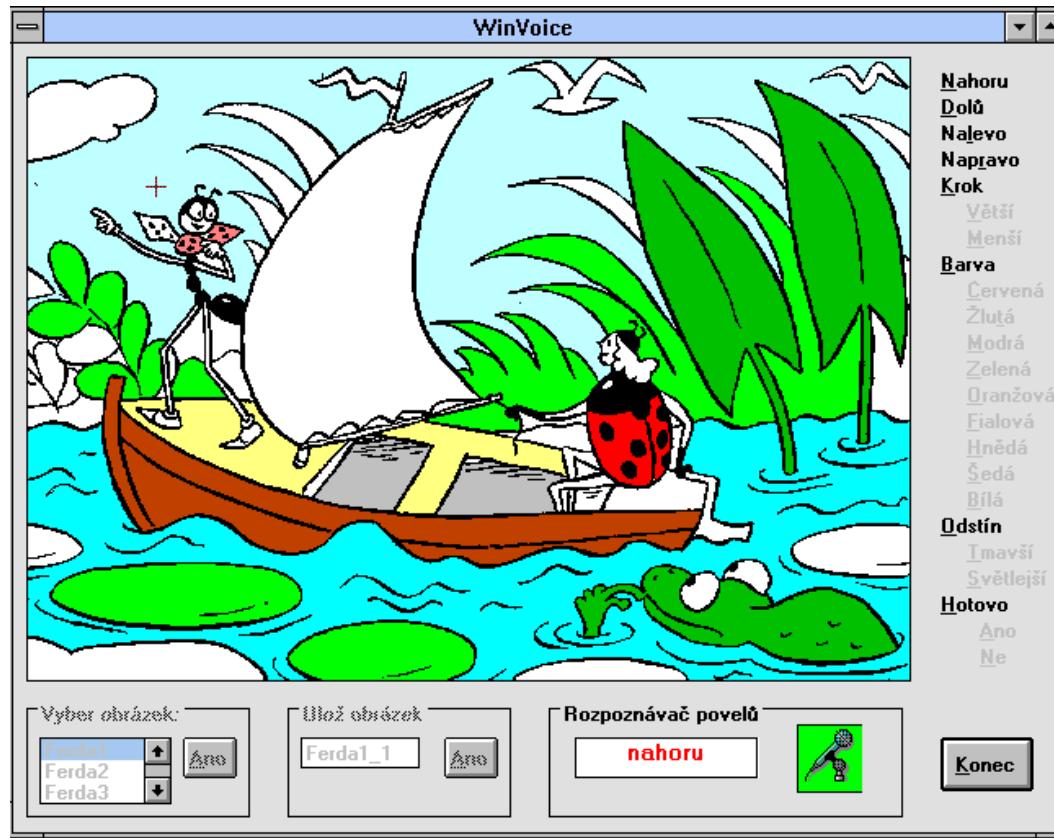
Z historie - 1994



VoiceCad - kreslení hlasem

PC 386 + speciální HW, 33 slov (libovolný řečník)

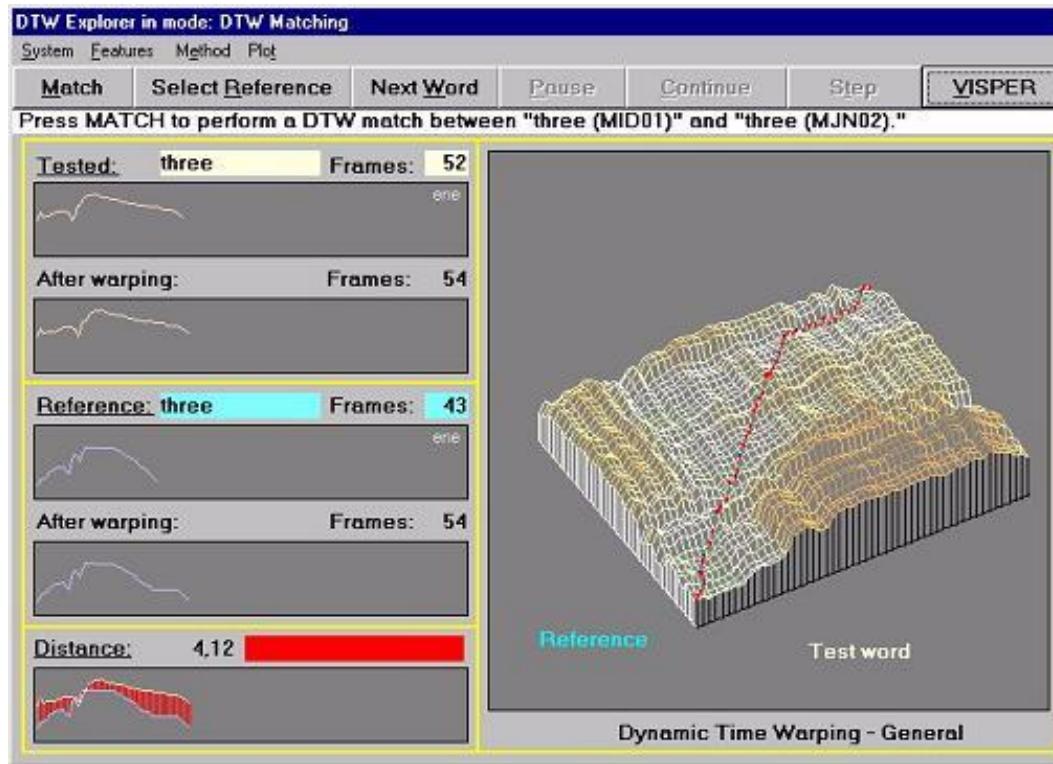
Z historie - 1996



Hry ovládané hlasem

Windows 3.1, desítky slov, určeno pro děti, exponát v NTM

Z historie - 1997



VISPER

systém pro vizualizaci algoritmů, výuku a experimentování

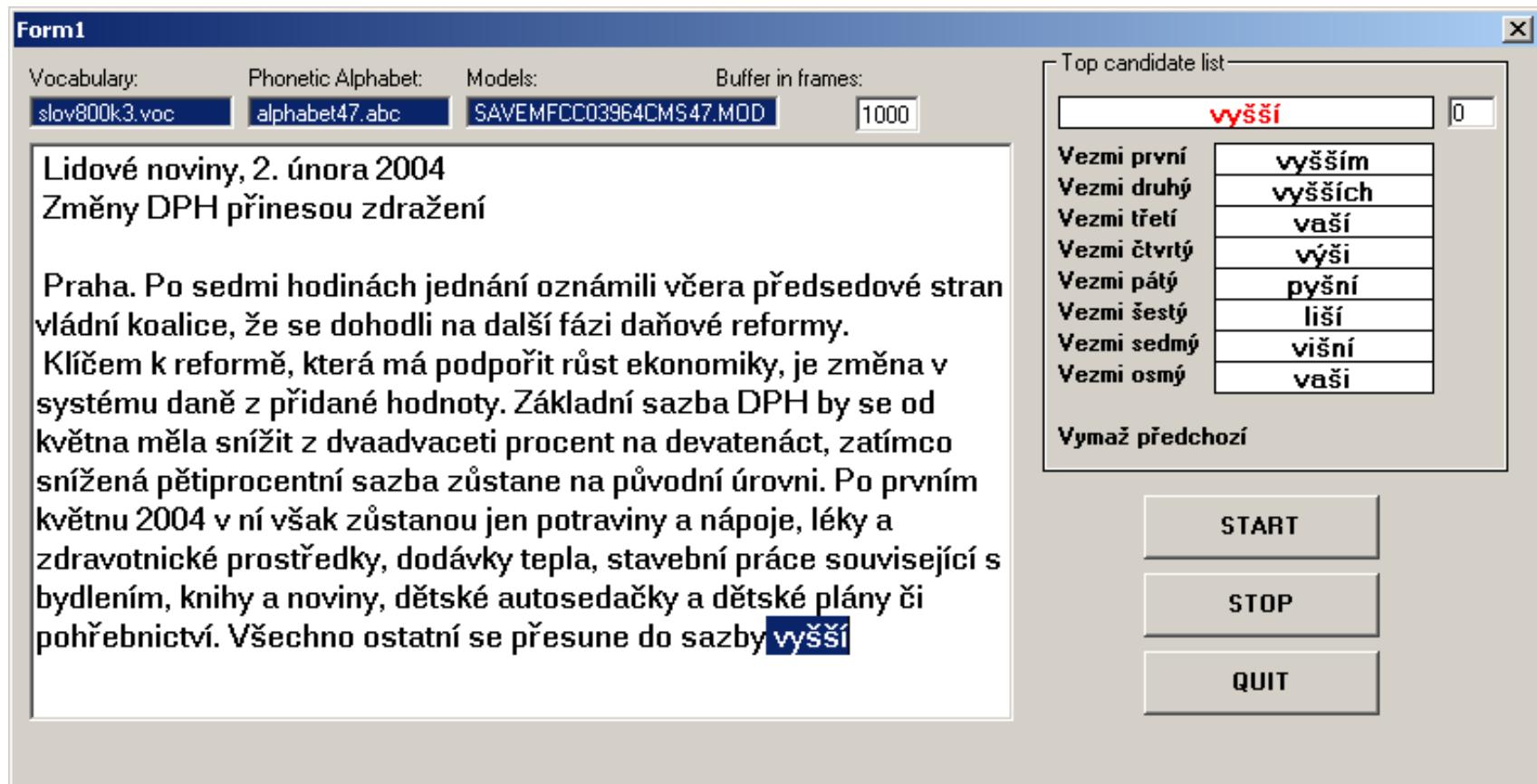
Z historie - 1999



InfoCity

první český telefonní systém založený na hlasovém dialogu s počítačem - v provozu v Liberci do roku 2006 (30 000 volání), jeho jádro je součástí několika komerčních aplikací

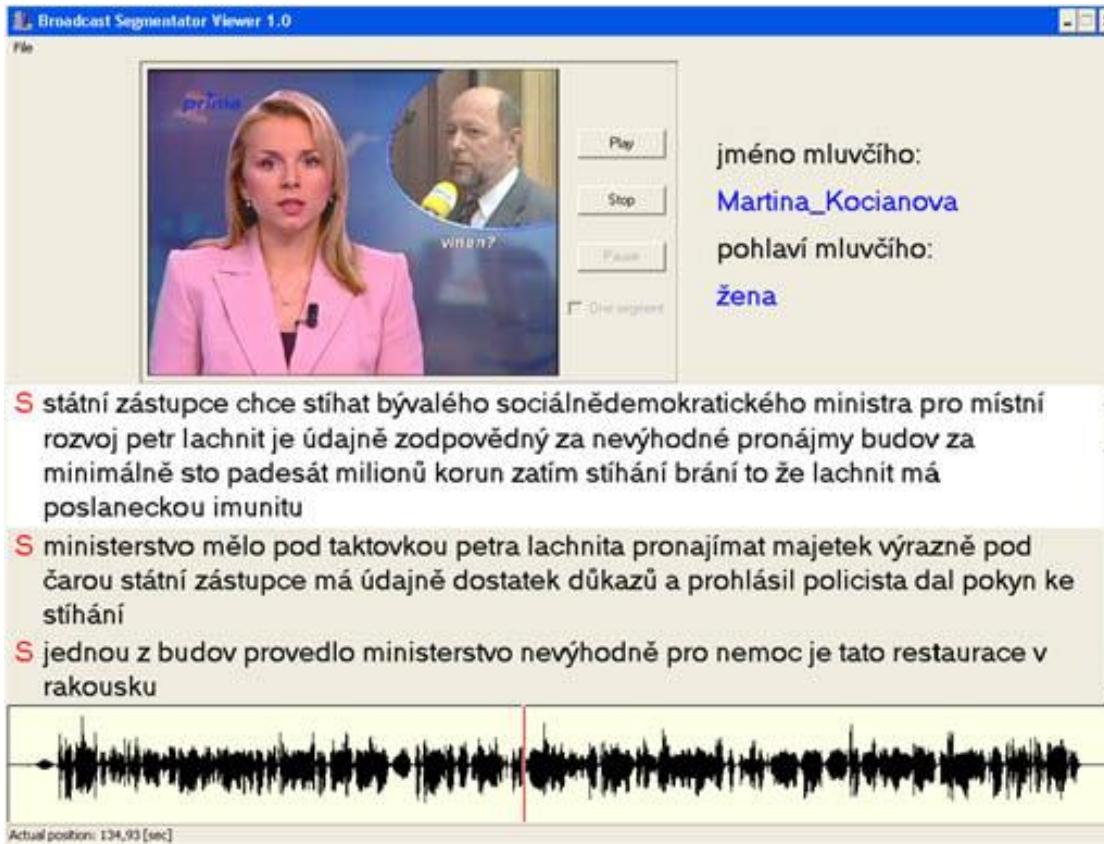
Z historie – 2003



Hlasový diktát pro češtinu

Zvládnutí izolovaného rozpoznávání s obrovským slovníkem (1 milion slov)

Z historie - 2004



Rozpoznávání plynulé češtiny s velkým slovníkem (100 K+)
Ukázkové aplikace: diktát, přepis zpráv

Z historie - 2005



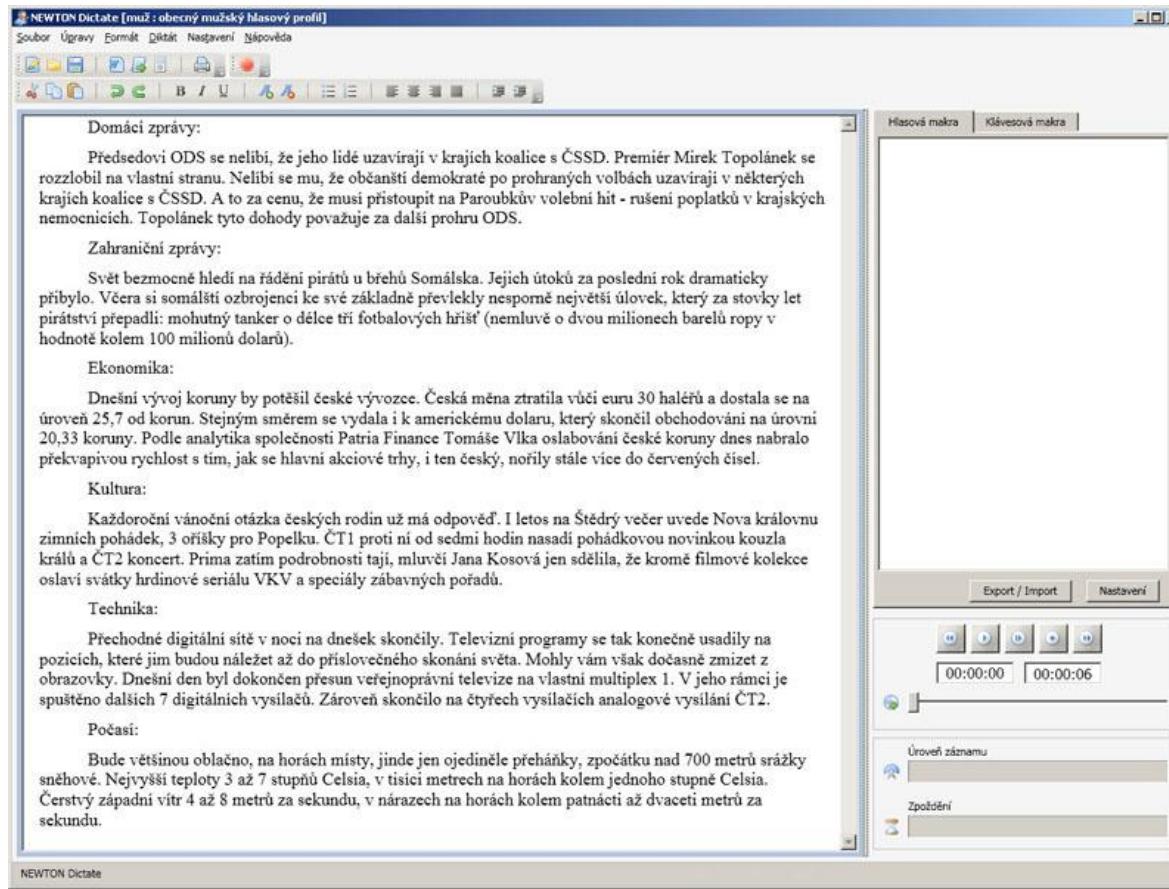
Systém MyVoice
– hlasové ovládání PC pro handicapované
(Cena Vládního výboru pro pomoc postiženým)

Z historie - 2007



Systém monitorování rozhlasových a televizních stanic
ATT – úplný přepis vysílání (24 hodin denně), komplexní
distribuovaný systém rozpoznávání řeči, řečníka, hudby, atd.

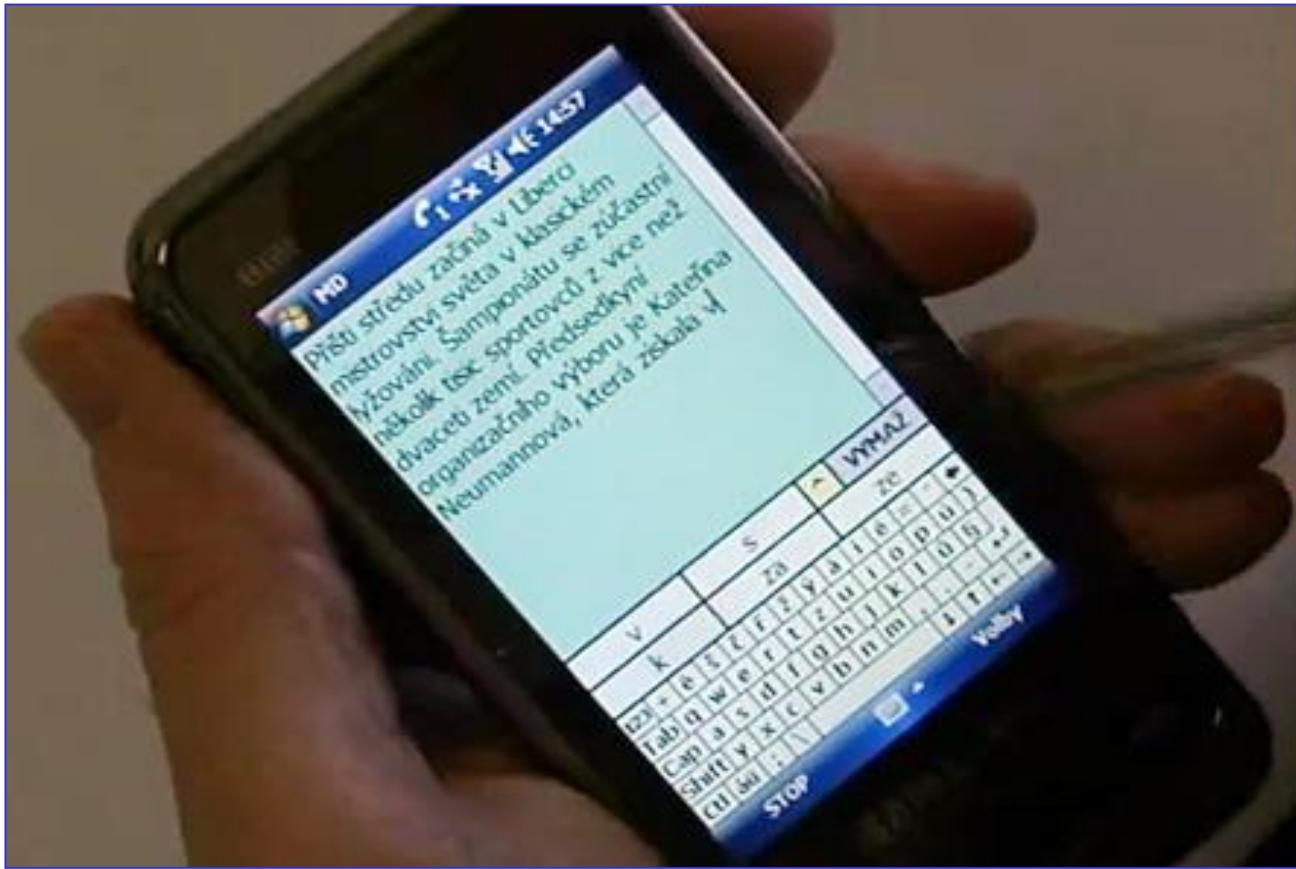
Z historie - 2009



Diktovací program Newton Dictate

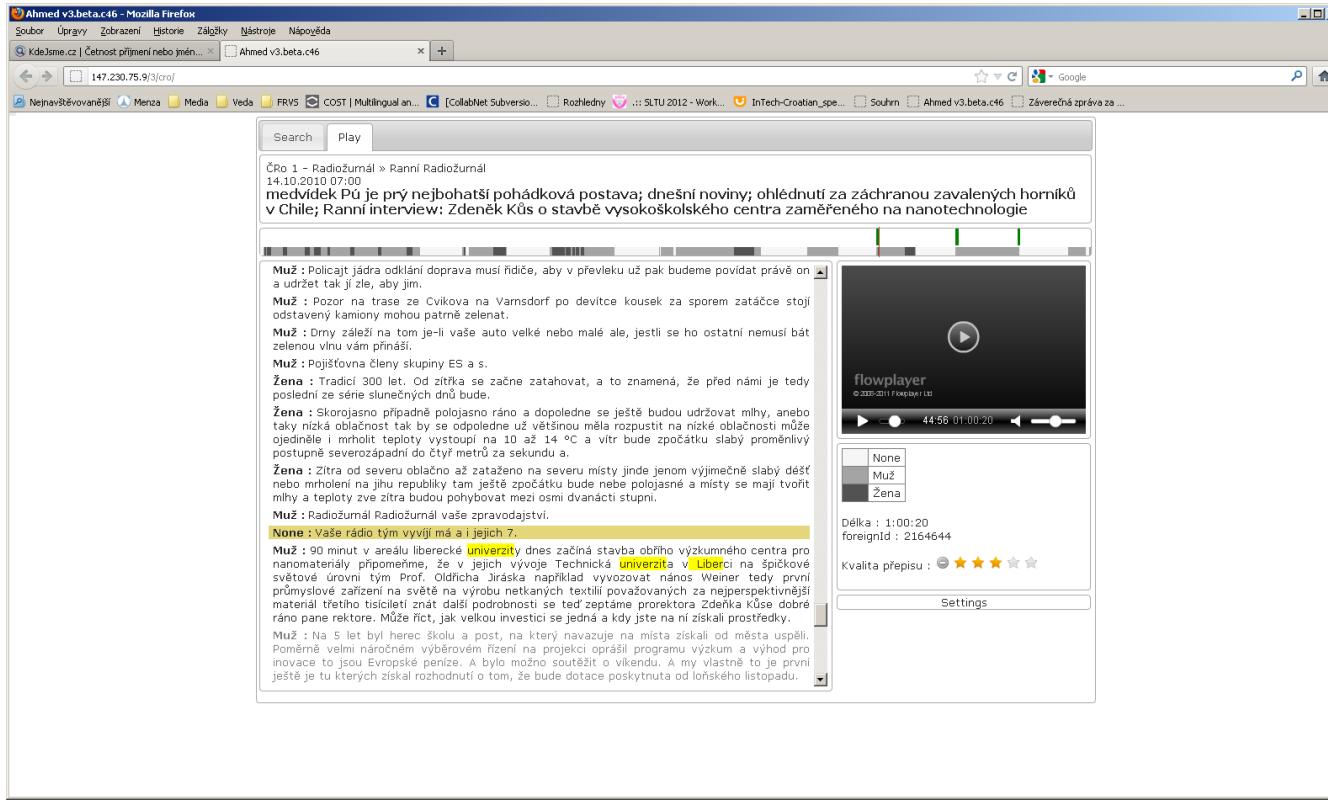
- umožňuje plynulé diktování v různých oblastech – novinářský text, justice, lékařství – v několika slovanských jazycích

Z historie - 2009



Diktovací program do smartphonu
- rozpoznávání přímo v mobilu nebo na serveru

Z historie 2014



Projekt zpřístupnění archivu ČRo (MK ČR)

Systém pro přepis a zaindexování více než 250.000 archivních nahrávek

Ze současnosti

Český rozhlas - Dvojka



RUNNING



--nonspeech--

Polskie Radio - Jedynka



RUNNING



w historii polskiej kinematografii pewnie od razu każdy rozpozna, o czym mowa o

Slovenský rozhlas - Slovensko



RUNNING



e aj prekážka na dažďové počúvate dopravné správy nová nehoda sa stala vo Svätom

RTV SLO - Radio Slovenija A1



RUNNING



enija je očitke o kršitvah evropskega prava vseskozi zavračala državno tožilstvo

HRT - HR1 - Prvi program



RUNNING



ostanite uz nas i provedite tridesetak ugodnih minuta

Studio B - Radio Studio B



RUNNING



saobraćaju Dejane hvala ako nešto bude bilo tu smo čekamo vaš poziv krov naravno

Radio Russia - Radio Russia



RUNNING



на местах а тому как можно решить проблему неравномерного питания коммерческого

On-line přepis vysílání desítek rozhlasových stanic ve 20 jazycích

Mluvený a psaný jazyk

Jak řeč vzniká? Řečový signál

Text a řeč

**Každý „civilizovaný“ jazyk má dvě podoby:
psanou (text) a mluvenou (řeč)**

**Mluvená je vývojově starší, textová přišla až se vznikem
písma, pro každý jazyk zvlášt'.**

**Mluvená podoba je daleko variabilnější než textová -
důvod: textová podoba je dána standardy**

Základní stavební jednotky jazyka:

Věta (promluva) – nese konkrétní výpověď

Slovo – základní významová jednotka jazyka

Slabika – nejmenší vyslovitelná jednotka řeči

Písmeno(znak, grafém) – nejnižší grafická jednotka textu

Hláska (foném) – nejnižší odlišitelná fonetická jednotka řeči

Fonetika a fonologie

Fonetika – zabývá se zvukovou stránkou jazyka (jazyků), zkoumá jak se tvoří a správně vyslovují hlásky.

Fonologie – zkoumá zvukovou stránku jazyka z hlediska dopadu na význam řeči, rozlišuje pouze takové zvuky, které se podílejí na významu řeči

např. v češtině rozlišujeme dva fonémy „a“ a „á“, protože tvoří významově různá slova („rada“ a „ráda“), zatímco v řadě jazyků délka samohlásek nehraje roli

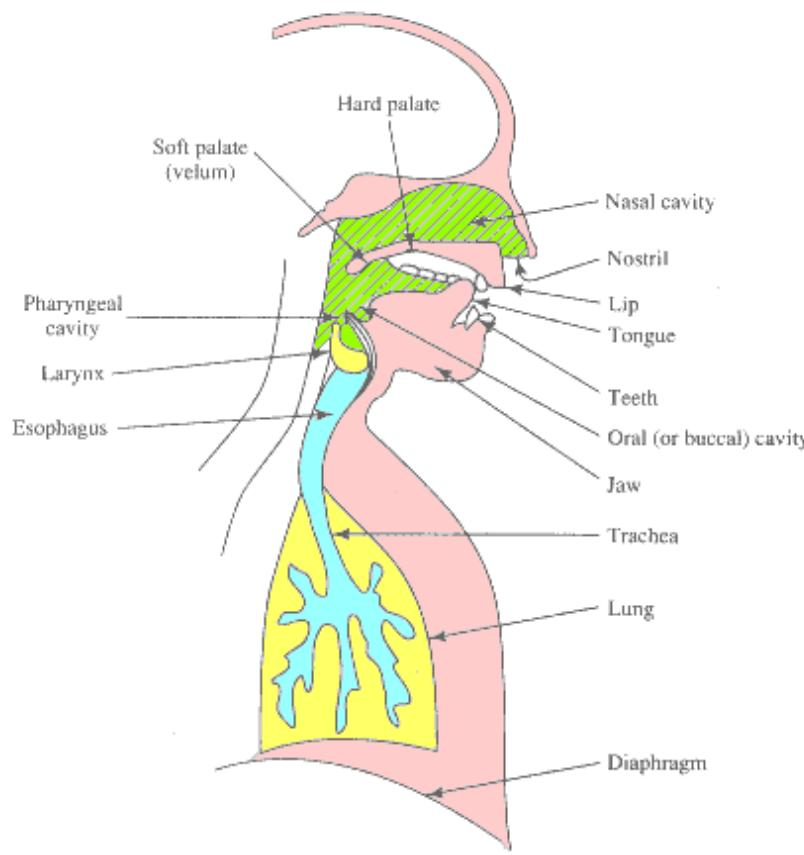
v ruštině rozlišují v řeči „i“ a „y“, češtině pouze jedený foném „i“

Grafémy – stavební jednotky pro zápis textu

Fonémy – stavební jednotky pro tvorbu řeči – k zápisu fonému ovšem používáme též znaky

Jak řeč vzniká (1)

Hlasový trakt - human vocal tract



EN

Lungs

Vocal cords

Pharyngeal cavity

Larynx

Oral cavity

Nasal cavity

Palate

Velum

Tongue

Teeth

Lips

CZ

Plíce

Hlasivky

Hrdelní dutina

Hrtan

Ústní dutina

Nosní dutina

Tvrdé patro

Měkké patro

Jazyk

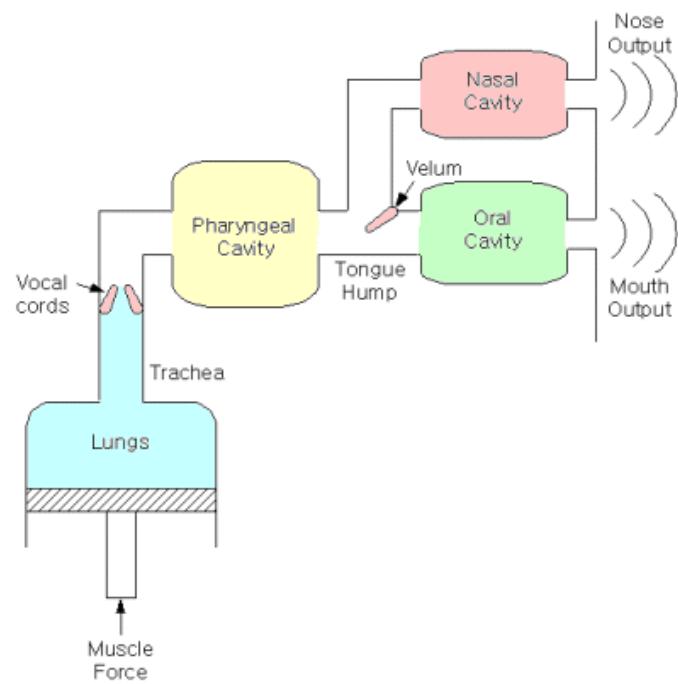
Zuby

Rty

Jak řeč vzniká (2)

Technický model

Technical model



EN	CZ
Lungs	Plíce
Vocal cords	Hlasivky
Pharyngeal cavity	Hrdelní dutina
Oral cavity	Ústní dutina
Nasal cavity	Nosní dutina
Palate	Tvrdé patro
Velum	Měkké patro
Tongue	Jazyk
Teeth	Zuby
Lips	Rty

Fonémy – základní jednotky řeči

Hlavní fonémové skupiny:

Samohlásky (Vowels) – a, e, i, o, u,

znělé (kvaziperiodické) zvuky vznikající vibrací hlasivek, proud vzduchu vycházející ústy (případně i nosem) není ničím omezován

Souhlásky (Consonants)

charakter zvuku je výrazně ovlivněn překážkami, které vzduchovému proudu kladou rty, zuby, pozice jazyku a patra, jde o zvuky s výraznou šumovou složkou a časově proměnným charakterem

Explozivy – p, t, k, ... hlásky jsou tvořeny nejprve uzávěrem proudu vzduchu a pak krátkým výbuchem

Frikativy – s, z, f, h ... cesta pro proud vzduchu je výrazně zúžena (např. jazykem, rty, zuby), čímž vzniká turbulentní proudění mající charakter šumu

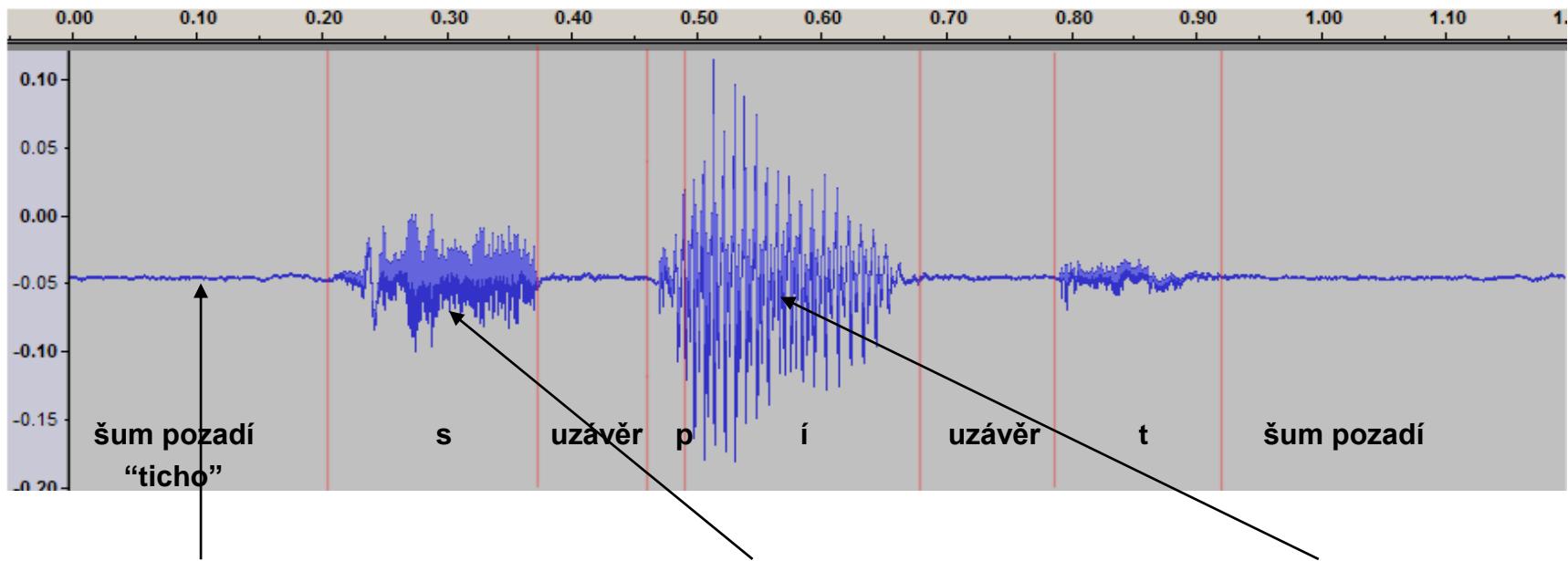
Nasály – n, m, vzduch proudí též nosní dutinou

Ostatní – l, j, r, ř, r a ř jsou tzv. vibranty, l a j tzv. approximanty
v češtině patří ke slabikotvorným souhláskám

Souhlásky mohou být znělé a neznělé (s účastí či bez účasti hlasivek)
mnohé tvoří páry: b – p, d – t, g – k, z – s,

Ukázka řečového signálu (1)

Slovo “speed” - znázorněné v časové oblasti



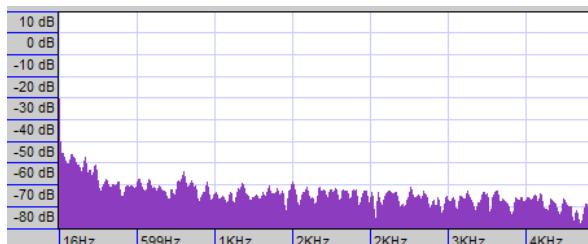
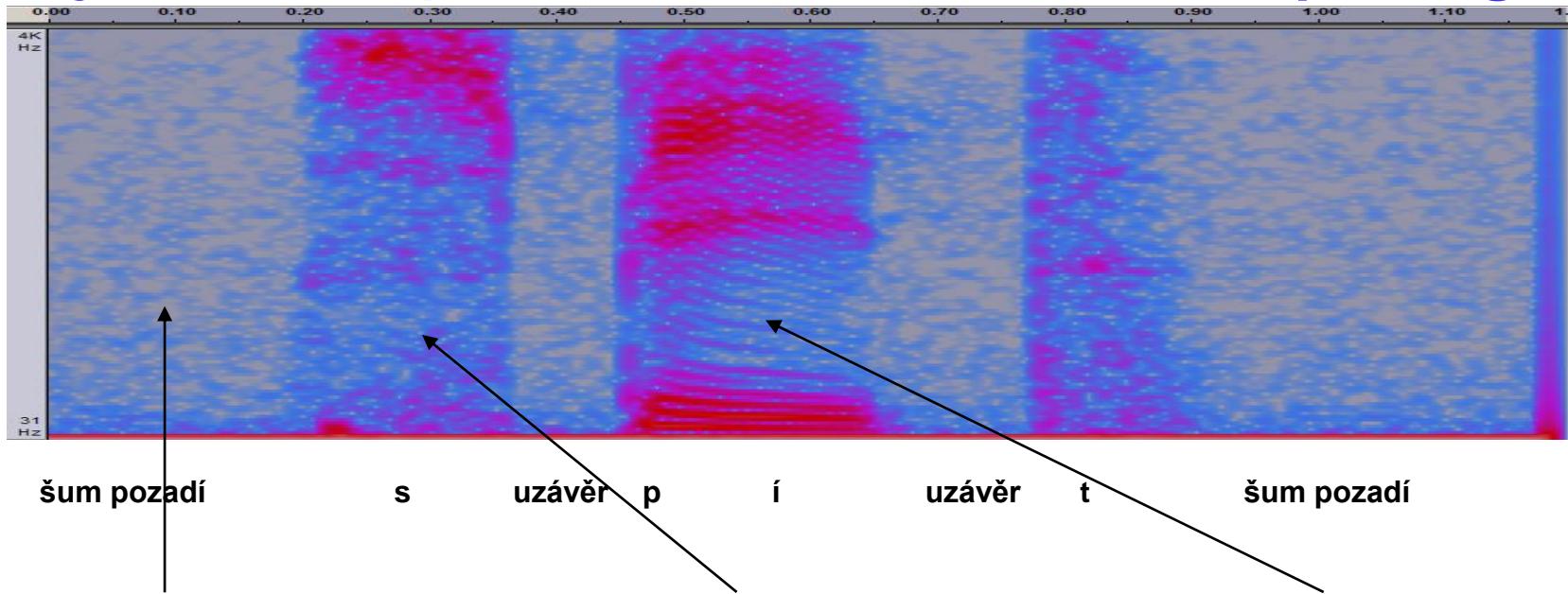
šum pozadí

s

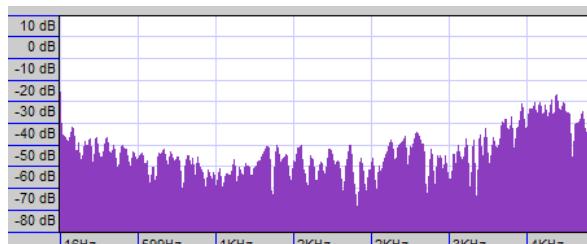
í

Ukázka řečového signálu(2)

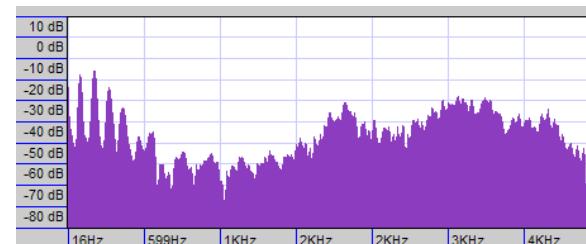
Stejné slovo “speed” ve frekvenční oblasti - spektrogram



šum pozadí



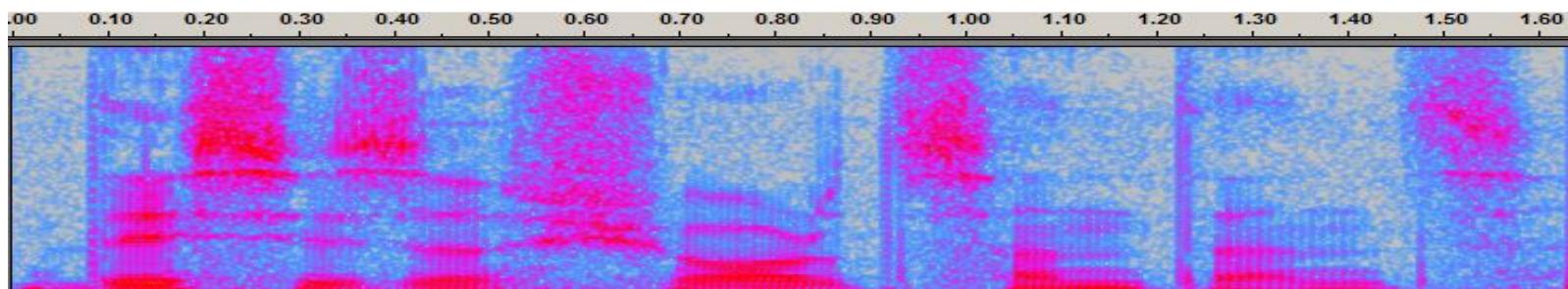
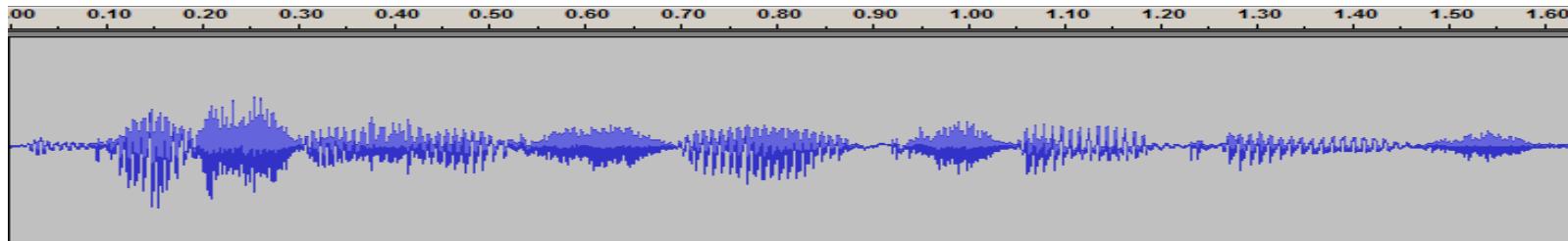
s



í

Řečový signál – časové poměry

Věta “This is a short sentence.”

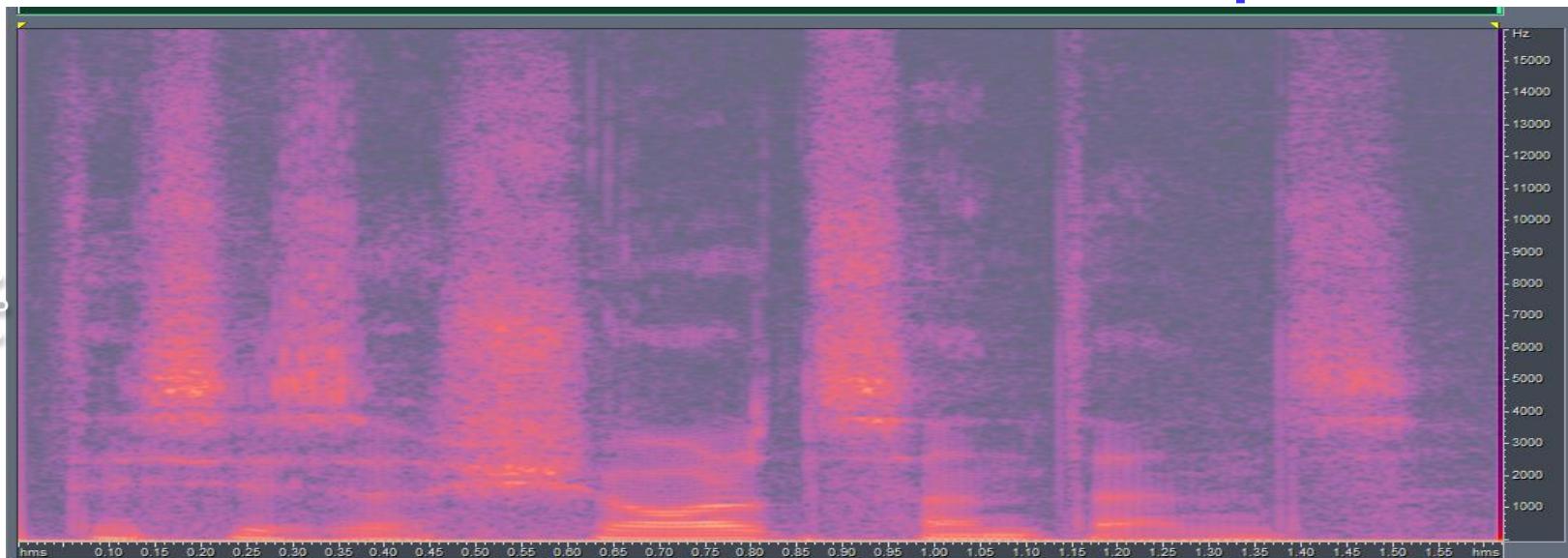


Všimněme si:

- 1) Do 1 sekundy se běžně vejde několik slov (5 – 30 fonémů podle tempa řeči)
- 2) V běžné plynulé řeči nejsou žádné “pauzy” (a tím pádem ani hranice) mezi jednotlivými slovy. (Je to jako kdyby nebyly mezery mezi slovy textu)
vběžné plynulé řeči nejsou žádné pauzy mezi jednotlivými slovy.

Řečový signál – frekvenční poměry(1)

Věta “This is a short sentence.” vzorkovaná při 32 kHz.



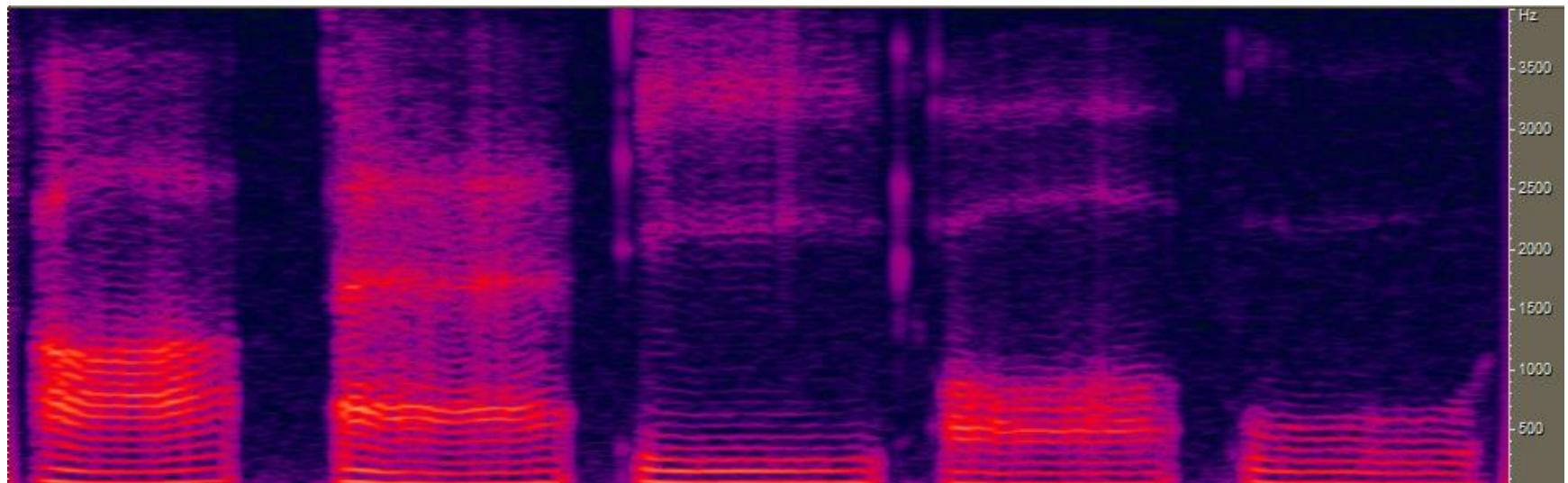
Všimněme si:

- 1) S výjimkou frikativ se signál dobře „vejde“ do pásma 0 - 8kHz, proto 16 kHz vzork. frekvence je vhodná a často používaná (Nyquistův teorém!)
- 2) 8 kHz vzorkovací frekvence se často používá v digitální telefonii, přenosové pásмо je pak zúženo na cca 3 kHz (standardně 300 – 3400 Hz).
- 3) Porovnejte si kvalitu a srozumitelnost zvuku: 32 kHz 16 kHz 8kHz 4kHz



Řečový signál – frekvenční poměry (2)

Sekvence samohlásek “A E I O U” (vzorkovaných 8 kHz)

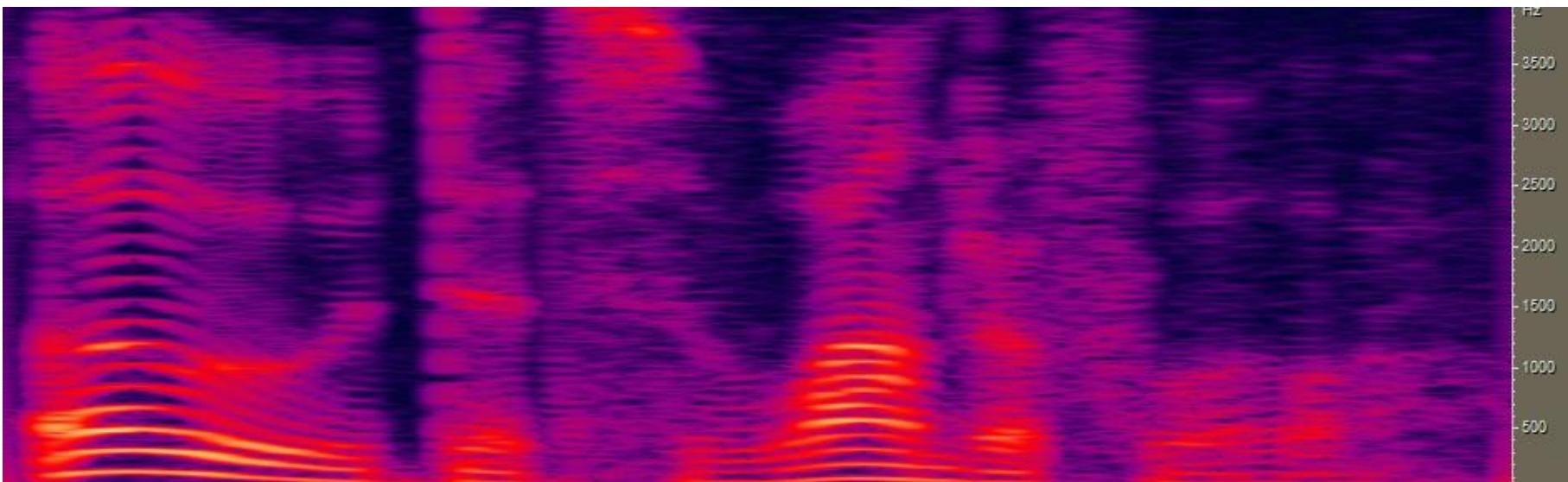


Všimněme si:

- 1) Samohlásky jsou kvaziperiodické signály. Proto, Fourierova analýza nachází u těchto signálů základní frekvenci F0 (zde okolo 100 Hz) a vyšší harmonické.
- 2) Různé samohlásky se liší ve spektru. Některé frekvenční oblasti jsou zesíleny, jiné potlačeny.
- 3) Zesílená frekvenční pásma se nazývají formanty. Odpovídají resonančním frekvencím hlasového traktu (jsou určeny především nastavením ústní dutiny).

Řečový signál – frekvenční poměry (3)

Věta “Oh, that’s wonderful!” (vzorkovaná 8 kHz)

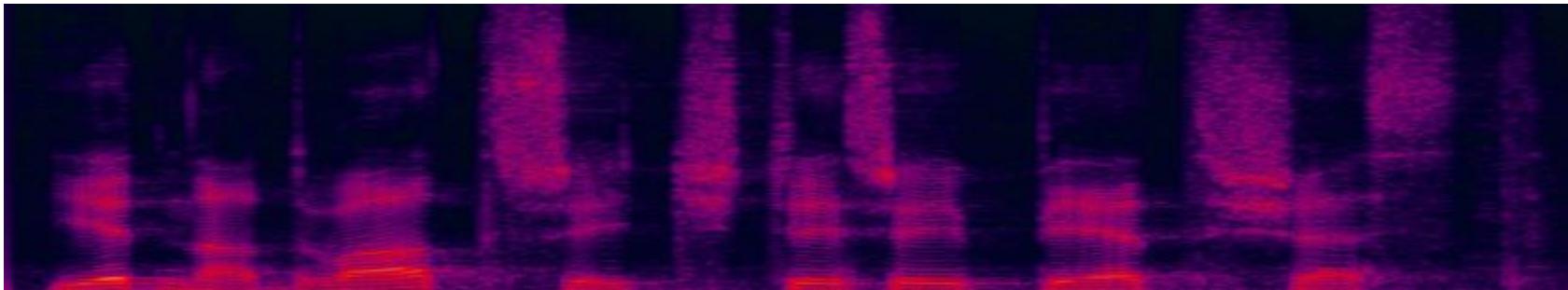


Všimněme si:

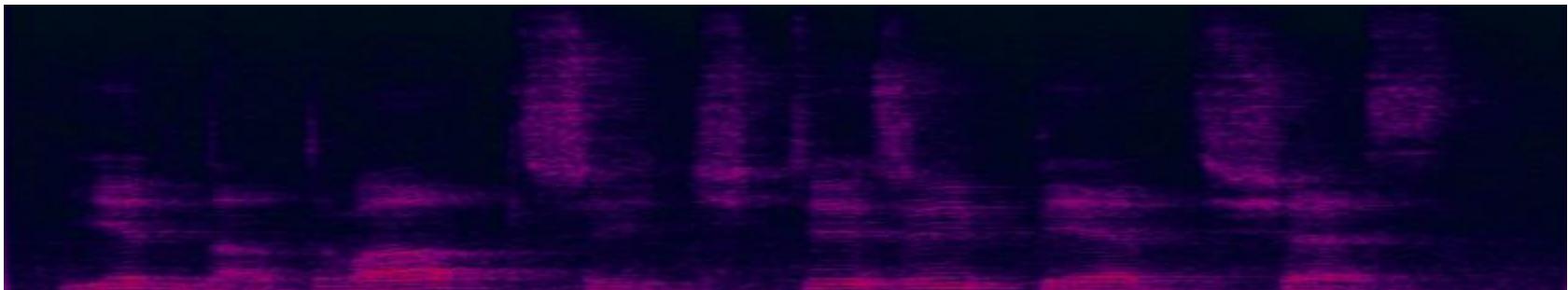
- 1) Základní frekvence F0 se mění v čase. Tomuto jevu se říká intonace (“melodie řeči”). Intonace spolu s dynamikou a tempem řeči určuje prozodii řeči.
- 2) Muži a ženy se liší v hodnotách F0 – průměrná hodnota pro mužské mluvčí je 120 Hz, pro ženy okolo 250 Hertz.
- 3) Zákl. frekvence F0 se mění ve velkém rozsahu při zpívané řeči (zpěvu).
- 4) Zákl. frekvence F0 není přítomná v šepтанé řeči (hlasivky zde nepracují).

Řečový signál – snímání mikrofonem

Blízký mikrofon (close-talk mic – do 5 cm od úst)



Vzdálený mikrofon (distant mic – více než 20 cm od úst)



Hlavní rozdíly:

- 1) Intenzita signálu klesá se čtvercem vzdálenosti
- 2) S vyšší vzdáleností jsou více potlačovány vyšší frekvence
- 3) Vzdálený mikrofon snímá i odrazy zvuku v místnosti

Zadání úlohy na cvičení (1)

Vytvořit v MATLABu program, který bude generovat hlasová hlášení, např.

- 1) hlášení o aktuálním času – „Je právě XX hodin, YY minut“.
- 2) hlášení o odjezdech vlaků – „Osobní vlak do Prahy odjíždí z 5. nástupiště.“
- 3) Hlášení čísel tažených ve Sportce – „Tažená čísla jsou 5, 24, 37 ...“
- 4) Hlášení od odletech, o počasí, apod.

Řešení: vytvořit sadu nahrávek a podle požadavků je pak vždy složit do jednoho hlášení.

AUDACITY - Program pro nahrávání a editaci (střih, úprava, apod)

<http://audacity.sourceforge.net/>

Zadání úlohy na cvičení (2)

Rady pro nahrávání:

1. Nahrávky je třeba pořizovat v klidných podmírkách, z blízkého mikrofonu, bez ruchů, je třeba nastavit vhodné zesílení (ne příliš potichu, ale také bez přebuzení).
2. Všechny nahrávky pořizujte za stejných podmínek (nejlépe v jedné seanci, aby bylo zajištěné stejné nastavení mixeru, stejná vzdálenost úst od mikrofonu, stejná barva hlasu)
3. Kvůli přirozenosti projevu pořizujte raději delší nahrávky, z nichž pak vyřízněte v programu Audacity menší jednotky. (Např. „Je | pět | hodin | deset | minut“)
4. Pamatujte také na intonaci („Je pět hodin.“ vs „Je pět hodin deset minut.“)
5. V Matlabu si vytvořte jednoduchý program, který nahrávky z disku nahraje a pak je podle potřeby spojí.

Odevzdání úlohy

Vytvořenou úlohu včetně všech nezbytných dat mi pošlete v jednom souboru ZIP nejdéle do pondělí 18.00.

Vyzkoušejte si, že v ZIPu je vše, aby program fungoval na cizím počítači (s nainstalovaným Matlabem)

Můžete si zvolit, kterou ze zmíněných aplikací vytvoříte. Pokud vás napadne nějaká jiná a zajímavá aplikace, bude to vítáno a případně i oceněno.

Za splnění budete dostávat bod (někdy i extra-bod)

Získané body budou rozhodovat o zápočtu a závěrečné známce.