

Veřejná nahrávka¹

- 1 Následující text jsem převzal z odborné literatury o vývoji počítačové syntézy z psaného projevu na řeč. Jedná se o systém, který má v budoucnosti zcela nahradit živého mluvčího. Jako interpretku jsem zvolil bývalou rozhlasovou hlasatelku a redaktorku Šárku Skalákovou.

běžnou mlhovinu snů, slunce svít na svíčkách, znamení milenců. A někdy méně, někdy více, má i své zatmění srdce. Krok dělil nás, jen jeden krok dálkou svou, a ta dálka byla v nás. Co je lehkomyslné láskou svou, a to spíš bude tím, že u každého dostaví se částečné zatmění srdce, a to vzápětí. Ani pět

slavných astronomů nemá na ten úkaz názor jednotný. Známy chirurg cestou domů podotknul jen, dejme tomu, že u každého dostaví se částečné zatmění srdce. Jenže u mne dostaví se úplné zatmění srdce. Pokud je vám Kabát bližší než Lenka Filipová, dočkáte se za nějaké tři minuty, během kterých

VEŘEJNÁ NAHRÁVKA

Schopnost mluvit je základním komunikačním prostředkem mezi lidmi. Není divu, že vytváření řeči již odedávna zajímalo naše předchůdce. Syntéza řeči představuje důležitou oblast problematiky zpracování řečového signálu a je již po dlouhá léta předmětem intenzivního výzkumu. Jde o proces, při němž se uměle vytváří řeč. V případě počítačové syntézy je řeč uměle vytvářena počítačem. Důvody enormního zájmu o syntézu řeči jsou nasnadě: řeč je nepřirozenější formou komunikace mezi lidmi. Je to tedy činnost člověku vlastní a přirozená. Umělé vytváření řeči počítačem si klade za cíl „zpřirozenit“ komunikaci člověka s počítačem a stát se tak rovnocenným partnerem tradiční vizuální komunikaci. Zařízení, které proces vytváření řeči provádí, nazýváme syntetizér řeči. Syntetizér je jádrem každého systému konverze textu na řeč. Konečným cílem syntézy řeči je vytvářet řeč v takové formě a kvalitě, aby nebyla rozpoznatelná od řeči člověka. Syntetická řeč by tedy neměla působit jednotvárně, měla by znít přirozeně a její poslech by neměl unavovat ani vyžadovat zvýšenou pozornost.

**á, ě, ě, bé, cé, čé, dě, é, e s čárkou,
e s háčkem, ef, gé, há, cha, í, i s čárkou, jé,
k, el, em, en, ěň, ó, ó, pé, kvé, er, eř, s, eš, té,
tě, ú, u s čárkou, u s kroužkem, f, dvojité vé,
iks, ypsilon, ypsilon s čárkou, s, žet²**

2 Přednáška je přerušena osmnácti vstupy, ve kterých hlasový syntetizátor předčítá čím dál komplexnější text. Jedná se převážně o sémanticky nepředvídatelné texty, na kterých se testuje schopnost systému správně vyhodnotit výslovnost daných slov a vět, respektive schopnost posluchačů jim porozumět.

Dávno předtím, než bylo vynalezeno elektronické zpracování signálu, se někteří lidé snažili sestavit stroje, které by vytvářely lidskou řeč. Podle raných legend o „mosazných hlavách“ mezi ně patřili papež Silvestr II., Albertus Magnus a Roger Bacon.

Mosazná, měděná či bronzová hlava je legendární robot z raného novověku připisovaný pozdně středověkým učencům s reputací génů. Mužská hlava, vyrobená z mosazi nebo bronzu, byla mechanická nebo magická. Stejně jako Mimirova hlava ve skandinávské pohanské tradici prý uměla správně odpovídat na jakoukoli otázku, která jí byla položena, ačkoli odpovědi byly někdy omezené na „ano“ nebo „ne“.

**á, f, se, na, že, je, ó, s, s, k, f, á, se, na, že,
ó, je, s, to, s, je, se, f, to, na, á, o, í, že, by,
á, f, se, na, je, že, ó, s, s, se**

Zřejmě první pokusy vytvářet řeč uměle provedl v roce 1779 německý profesor Christian Kratzenstein, který vysvětlil fyziologické rozdíly mezi pěti dlouhými samohláskami a sestrojil akustické rezonátory napodobující hlasový trakt člověka, pomocí nichž samohlásky uměle vytvářel. Rezonátory budil vibrujícími rákosy (podobně jako u hudebních nástrojů).

**sil sil
zet bé je
sbě
běr**

**jero
erod
es pé
rod p
odpad
pad
padu
a dusil
dusil
sil**

První „syntetizér řeči“ představil o několik let později v roce 1791 maďarský šlechtic narozený v Bratislavě a žijící ve Vídni, Wolfgang von Kempelen. Jeho „mluvícím strojem“ bylo mechanické zařízení, jímž se snažil napodobit artikulační orgány člověka. Lehce zdokonalenou verzi von Kempelenova mluvičího zařízení sestrojil Charles Wheatstone. Skládalo se z hlavních a pomocných měchů, rákosu, pryže, kožené trubice a malých trubiček. Hlavní měchy simulovaly funkci plic, rákos činnost hlasivek, ústa se modelovala pomocí pryže. Kožená trubice představovala rezonanční prostor. Malé trubičky plnily funkci nozder. Jednotlivé zvuky se vytvářely modifikací budicího zvuku v rezonanční trubicí. Znělé zvuky se vytvářely vibracemi rákosu, frikativy pak vypouštěním proudu vzduchu pomocnými trubicemi. Pomocné měchy simulovaly expanzi hlasového traktu po zavření úst při generování závěrových souhlásek. Celé zařízení se ovládalo pomocí pák a bylo údajně schopné generovat primitivní slova i věty. Výsledky von Kempelenových pokusů ukázaly, že hlasový trakt

(dutina mezi hlasivkami a rty) významně ovlivňuje vytváření řeči. Až do té doby byly za centrum řečové produkce považovány pouze hlasivky.

Po vzoru von Kempelena se až do začátku 20. století objevovala další mechanická zařízení, která se snažila zvýšit kvalitu vytvářených zvuků. Do výzkumu se zapojily i tak známé osobnosti jako Bell, Helmholtz, Faber nebo Riesz.

**van
pan
mas
ras**

**dam
kam
naj
kap**

**laň
saň
ňam
zaf**

**van pan mas ras
dam kam naj kap
laň saň ňam zaf**

Ve dvacátých letech 20. století se vědci začali zabývat elektronickými modely vytváření hlasu. První opravdové elektronické zařízení pro syntézu řeči

představil v roce 1922 Jay Q. Stewart. Používal bzučák jako buzení dvou rezonančních obvodů, které modelovaly první dva formanty hlasového traktu. Zařízení bylo schopné vytvářet jednotlivé samohláskové zvuky. V roce 1932 japonští vědci Obata a Teshima přidali třetí formant a zvýšili srozumitelnost generovaných samohlásek. Tři formanty jsou obecně považovány za dostačující pro vytváření srozumitelné syntetické řeči.

**plot plod plot plot plot plod plod plot plod
plot plod plot ples plot ples plod ples plot
dá plod dá plot dá plod
plot ples plod ples plot ples plod ples plot
ples plod dá plot dá dá da dá da**

První elektronický syntetizér souvislé řeči, Voder (Voice Operating Demonstrator), uvedl v roce 1939 na světovém veletrhu v New Yorku Homer Dudley z Bell Laboratories. Skládal se z deseti paralelně zapojených pásmových propustí, které byly buzeny buď periodickým signálem (v případě znělých zvuků), jehož perioda byla nastavována nožním pedálem, nebo v případě neznělých zvuků šumem. Výstupy jednotlivých pásmových propustí se nezávisle na sobě zesilovaly pomocí potenciometrů a sčítaly. Zesilovače se nastavovaly pomocí deseti tlačítek. Znělost se řídila pomocí přepínače. Další přidavná tlačítka se pak používala pro vytváření krátkodobých přechodových zvuků, jako v případě závěrových souhlásek. Voder byl (stejně jako stroj von Kempelena) schopen

obsluhovat pouze zkušený a trénovaný operátor. Nicméně tento syntetizér byl schopen produkovat do jisté míry srozumitelnou souvislou řeč, i když kvalitativně stále na hony vzdálenou dnešní úrovni. Voder je tak obecně považován za první zařízení, které se může honosit „titulem“ syntetizér řeči. Definitivně bylo prokázáno, že je možné uměle vytvářet srozumitelnou řeč.

**koleno
noha
dokola
koleno
noha
dokola
koleno noha dokola
koleno noha dokola
koleno
koleno noha
dokola
koleno noha dokola dokola la la la**

Franklin Cooper z Haskins Laboratories, počátkem padesátých let představil Pattern Playback, jakéhosi předchůdce současné konkatenační syntézy. Pattern Playback pomocí optického zařízení transformoval spektrogramy na řeč.

**laik – detail
kofein – detail
Alois – detail**

**laik – detail
kofein – detail
Alois – detail
laik – detail
kofein – detail
heroin – detail
pointa – detail
naivní detail
prozaik
detail**

Dalším milníkem ve vývoji syntézy řeči byl rok 1953, kdy Brit Walter Lawrence vyvinul první paralelní formantový syntetizér – PAT (Parametric Artificial Talker). Přibližně ve stejné době švédský vědec Gunnar Fant představil i kaskádní formantový syntetizér OVE I (Orator Verbis Electris), v roce 1962 zdokonalený jako OVE II. Tehdejší výzkumy ukázaly, že při pečlivém manuálním nastavení parametrů mohou formantové syntetizéry reprodukovat přirozenou promluvu v kvalitě v podstatě neodlišitelné od originálu. Toto zjištění bylo velkým povzbuzením pro budoucí výzkum.

**kost most radost
most radost
kost radost
radost radost**

V roce 1961 fyzik John Larry Kelly Jr. a jeho kolega Louis Gerstman použili k syntéze řeči počítač

IBM 704. Kellyho syntz... syntezátor³ přezpíval skladbu Daisy Bell, již hudebně doprovodil Max Mathews. Shodou okolností byl tehdy na návštěvě v Bellových laboratořích spisovatel Arthur C. Clarke. Ukázka výstupu tohoto syntezátoru Clarka natolik ohromila, že scénu použil ve svém románu *2001: Věsmírná odysea*, kde počítač HAL 9000 zpívá tutéž píseň.

**ráz šva
ráz šva
ráz šva
tuk dům
dítě kost
tygr muž
vino laňka
cena oči džbán
tango tramvaj abych byl
tři
krk vlk osum
vlak rok moře jev**

Zhruba v polovině šedesátých let nastal zlom ve vývoji syntézy řeči. S nástupem číslicových počítačů se výrazně změnil i přístup k syntéze řeči. Před vlastní hardwarovou realizací (anebo později místo hardwarové realizace) bylo možné syntetizéry simulovat pomocí číslicových počítačů. Objevují se digitální verze úspěšných elektronických syntetizérů PAT a OVE, ale i spousta dalších digitálních syntetizérů.

³ Přestože se Šárka Skaláková na čtení velmi pečlivě připravovala, v této jediné větě se přepřekla. Neplánovaná lidská chyba tak podtrhla rozdíl mezi umělou inteligencí a živým člověkem.

**pil pych pyj piš piv PIN
pes les ves bez děs rez
důl hůl vůl sůl půl kůl
rak tak sák vak lak pak**

První pokusy v oblasti konkatenční syntézy byly zaznamenány zřejmě v roce 1968, kdy Rex Dixon a David Maxey vytvořili syntetizér, který řetězil řečové jednotky na úrovni difonů parametrizovaných pomocí formantových frekvencí. V roce 1977 pak Joe Olive a jeho kolegové z Bell Laboratories navrhli konkatenční syntézu řetězící difony popsané pomocí parametrů lineární predikce.

**lev les lem lep led len
byt lid kit žid hit vid
loď lom lov lok los lob
kos bos los nos šos es ó es**

Paralelně s vývojem řečových modelů a s nimi spjatých technik pro analýzu a syntézu řeči se však také vyvíjí problematika syntézy řeči z textu (angl. text-to-speech – TTS), označovaná v české literatuře také jako konverze textu na řeč. Jde o důležitý krok ve vývoji, neboť do té doby se totiž výzkum soustředil spíše na reprodukci řečových promluv než na vytváření promluv nových. TTS se navíc zabývá také zpracováním přirozeného jazyka a fonetickým popisem promluvy a jejím cílem je vytvářet řeč z libovolného textu. První opravdový systém TTS vyvinula v roce 1968 v Japonsku Noriko Umeda. Systém byl založen

VEŘEJNÁ NAHRÁVKA

na artikulačním modelu a generoval celkem srozumitelnou řeč, jejíž kvalita však ani zdaleka nedosahovala úrovně dnešních systémů. V této době byly položeny základy moderních systémů TTS. Zasloužili se o to Allen, Holmes, Klatt, Mattingly, Shcarme, Umeda a další.

**bál šál tál vál sál kál
dub dur dus duc duch duň
mít mís mír míň míč mým
suk puk kuk luk muk fuk**

V roce 1976 se ukázala praktická využitelnost syntézy řeči, když Raymond Kurzweil sestrojil jedinečný „přístroj pro čtení tištěných stránek“, určený pro nevidomé. Šlo vlastně o jednoúčelový počítač, který pomocí optického scanneru snímal tištěné stránky, převáděl je do digitální podoby a potom je transformoval na syntetickou řeč. Otevřel tak nové obzory zejména pro slepé lidi, do té doby odkázané pouze na Braillovo písmo. Ačkoliv byl přístroj příliš drahý pro průměrného zákazníka (koncem osmdesátých let se jeho cena pohybovala stále okolo 30 000 dolarů), byl hojně využíván zejména v knihovnách a pomocných centrech pro zrakově postižené.

**sklo
stříkat
strádat
svět
stříkat svět**

VEŘEJNÁ NAHRÁVKA

**strádat svět
sklo sjet
strádat**

Sedmdesátým létům dominuje Dennis Klatt z MIT, který významnou měrou ovlivnil vývoj oboru. V roce 1979 spolu s Jonathanem Allenem a Sheri Hunnicuttovou vyvinul formantový systém TTS nazvaný MITalk. O dva roky později pak Klatt uvedl jeho vylepšenou verzi nazvanou Klattalk, která je považována za nejlepší systém TTS založený na formantové syntéze. Technologie použitá v těchto systémech se stala základem mnoha ještě dnes prodávaných systémů.

**tekstil
sex
pexeso
xylofon
fix černý
fix mám
fax upadl
sex byl**

Zřejmě prvním integrovaným obvodem pro syntézu řeči byl Votrax – tento čip obsahoval kaskádní formantový syntetizér a jednoduché obvody pro vyhlazování řeči dolní propustí. V roce 1978 Richard Gagnon uvedl nepříliš drahý systém „Type-n-Talk“, založený na čipu Votrax. O dva roky později výzkumníci z Texas Instruments představili levný syntetizér „Speak-n-Spell“ využívající čip pro syntézu řeči

VEŘEJNÁ NAHRÁVKA

pomocí koeficientů lineární predikce TMS-5100. Používal se především jako čtecí pomůcka pro děti a dočkal se poměrně značné pozornosti. Začátek osmdesátých let se nesl ve znamení nástupu komerčních systémů TTS. Na prvním místě je určitě zapotřebí jmenovat DECtalk společnosti Digital Equipment Corporation, uvedený v roce 1983, jehož základem byl právě Klattalk. Šlo pravděpodobně o nejpoužívanější systém TTS 20. století.

V roce 1982 Rolf Carlson, Björn Granström a Sheri Hunnicuttová ve Švédsku představili Infovox – první vícejazyčný formantový systém TTS, následníka Fantových systémů OVE.

**nejistější
nejjistější
a nikoli nejistější poddaní
a nikoli podaný
předtančit a nikoli přetančit
odtud nikoli
odtud rozzáří
nikoli bezzásadový a nikoli čtverrohý
pod domem
nad tebe
a nikoliv**

Moderní technologie syntézy řeči používají celkem složité a propracované metody a algoritmy. V devadesátých letech 20. století se objevují první systémy využívající korpusově orientovanou syntézu řeči. Inventáře řečových jednotek těchto systémů se vytvářejí

VEŘEJNÁ NAHRÁVKA

automaticky na základě rozsáhlých řečových korpusů a často obsahují různé dlouhé řečové jednotky. S výhodou se přitom začínají využívat techniky známé především oblasti automatického rozpoznávání řeči, jako například skryté Markovovy modely (HMM) či neuronové sítě (NN).

**mnohem lepší
lepší
trochu lepší
stejně
trochu horší
horší
mnohem horší**

příliš žluťoučký kůň úpěl ďábelské módy

**příliš
žluťoučký
kůň
úpěl
ďábelské
ódy**

Syntéza řeči z textu nebo též konverze textu na řeč je nejobecnější a také nejkomplikovanější úlohou počítačové syntézy řeči. Jde o úlohu, jejímž řešením by měla být schopnost systému automaticky převést libovolný psaný text na mluvenou řeč, a to nejlépe v takové formě a kvalitě, jakou by stejný text přečetl člověk s dobrým přednesem.

expedici K2, dodává manažerka kulturních akcí v Besedě Hana Rendlová. Na své si přijdou i dříve narození, kromě tradičních tanečních večerů každý měsíc jednou v pátek od šesti večer, tak ve spolupráci s krajskou radou seniorů Plzeňského kraje pořádáme dva koncerty, jeden na konci září

VEŘEJNÁ NAHRÁVKA

**špatná
horší
slušná
dobrá
výborná**

**baterie je pod zadním sedadlem
jídlo visí pod teplým listem
žvýkačka stéká za hýčkaným oknem
akrobat šišlá nad borovým kufrem**

**starý pán je totální cvok
hladící šašek je polární vládce
čtecí bojovník byl běžný parník
vyrovnaný novorozenec bude
gramotný hrdina**

Systémy akustické syntézy řeči samy o sobě nabízejí široké pole uplatnění, ať už v oblastech, kde jiný než hlasový způsob komunikace nepřichází v úvahu, ale i tam, kde možnost hlasové komunikace výrazně obohatí kvalitu dané lidské činnosti nebo prostě tam, kde hlasová syntéza může ulehčit život. Syntetizovaná řeč může nahradit skutečného lidského řečníka na širokém spektru různých pozic – od rutinního oznamování opakujících se informací (zastávky MHD, nádraží apod.) přes hlasový monitoring údajů (řídící střediska), informační a dialogové systémy (automatická spojovatelka, telefonní klientské či informační linky) až po vysoce propracované a přirozené čtení libovolných textů (knihy, novinové články, úřední

VEŘEJNÁ NAHRÁVKA

korespondenci, e-mail, SMS zprávy, GPS navigace apod.). Systémy TTS jsou dnes již zcela běžnou součástí nejpoužívanějších operačních systémů (Microsoft Windows, MAC OS, Linux nebo Android). V současné době nelze též syntéze řeči upřít stoupající uplatnění v zábavním průmyslu.

**zcela neadekvátní
značná
průměrná
bez značné námahy
bez jakékoliv námahy**

**vyzkoušejte různé odstíny a materiály
jezte lesní stoly i vejce
zdvihni rozpustilé čmeláky nebo běhny
porovnejte opeřené sukýnky a ocásky**

**jak změnit systém veřejných financí?
jak léčit stavbu velkých hostí?
kam uvolnit orchestr mdlých oveček?
proč potěšit zdechlinu nahryzaných kostlivců?**

**Vítězslav cítil potřebu se vsadit
holič viděl korozi si povídat
celník slyší domky se sušit
bába vnímá modřiny se zhoršovat**

Konečným cílem by tedy nemělo být monotónním způsobem sdělit potřebnou informaci, ale generovat řeč přirozenou. Nepřirozeně znějící řeči je sice

rozumět, ale může být velice únavná a „otravná“ na poslech. Protože však úloha konverze textu na řeč zatím nebyla zcela uspokojivě vyřešena (syntetickému signálu často chybí právě přirozenost), lidé často dávají přednost lidskému hlasu před syntetickým. Nicméně syntéza TTS je stále předmětem intenzivního výzkumu a všichni zainteresovaní doufají, že je jen otázkou času, kdy se povede libovolný text převést na kvalitní, od lidské řeči nerozlišitelnou řeč. Pak teprve nastane pravý rozmach a se systémy TTS se budeme moci setkávat opravdu na každém kroku. Jedním dechem je ale třeba dodat, že generování naprosto přirozené syntetické řeči s sebou přinese i jistá bezpečnostní úskalí. Dokonalá syntetická řeč kopírující hlas i styl mluvy konkrétního řečníka bude přímo lákat ke zneužití jeho identity.

**Vybrané pasáže z knihy Mluvíme
s počítačem česky a z internetové
encyklopedie Wikipedia četla
bývalá hlasatelka Českého rozhlasu,
Šárka Skaláková.**

Testy srozumitelnosti TTS systémů a úryvky
z učebnice Jazyk a jazykověda četl TTS systém
Eliška vyvinutý společností Acapela Group. Zvuková
stopa se skládala z referenčních signálů určených
pro testování kvality reproduktorů.

Český rozhlas. Tam za trati, staví si chrám, plechovej most, už po něm kráčí první host, tak ať
ho nohy nebolí. Prošla si peklem a kouzla zná, přejetí mince počítám, a kdo ji spatří, je ztracen.
S touhle jedinou jsem zemřel v jedinej den, a jestli vám to nestačí, kdyby tam stála stovka žen, vyzvu

je k tanci. A pamatovat si budu na tu chvíli, když hudba zněla jak Paganini, a já věděl, že jsem
ztracen. Zeptal se za kolik. S touhle bych zemřel v jedinej den, a jestli vám to nestačí, kdyby tam
stála stovka žen, vyzvu je k tanci, a to netančím. S touhle bych zemřel v jedinej den, a jestli vám