Netværk for Dansk Lydteknologi

INNOVATIONSPROJEKT

LYDINDEKSERING I STORE LYDARKIVER

Kognitive systemer, DTU Informatik, Lasse Mølgaard

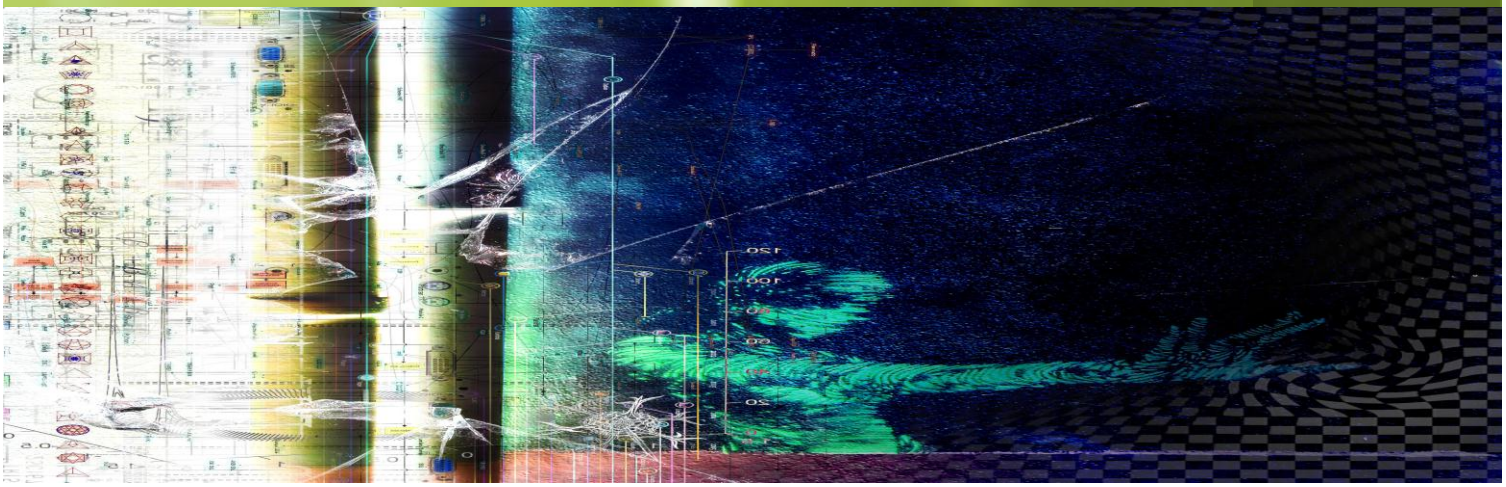
Aalborg Universitet, Zheng-Hua Tan

Danmarks Radio, Ivan Dehn

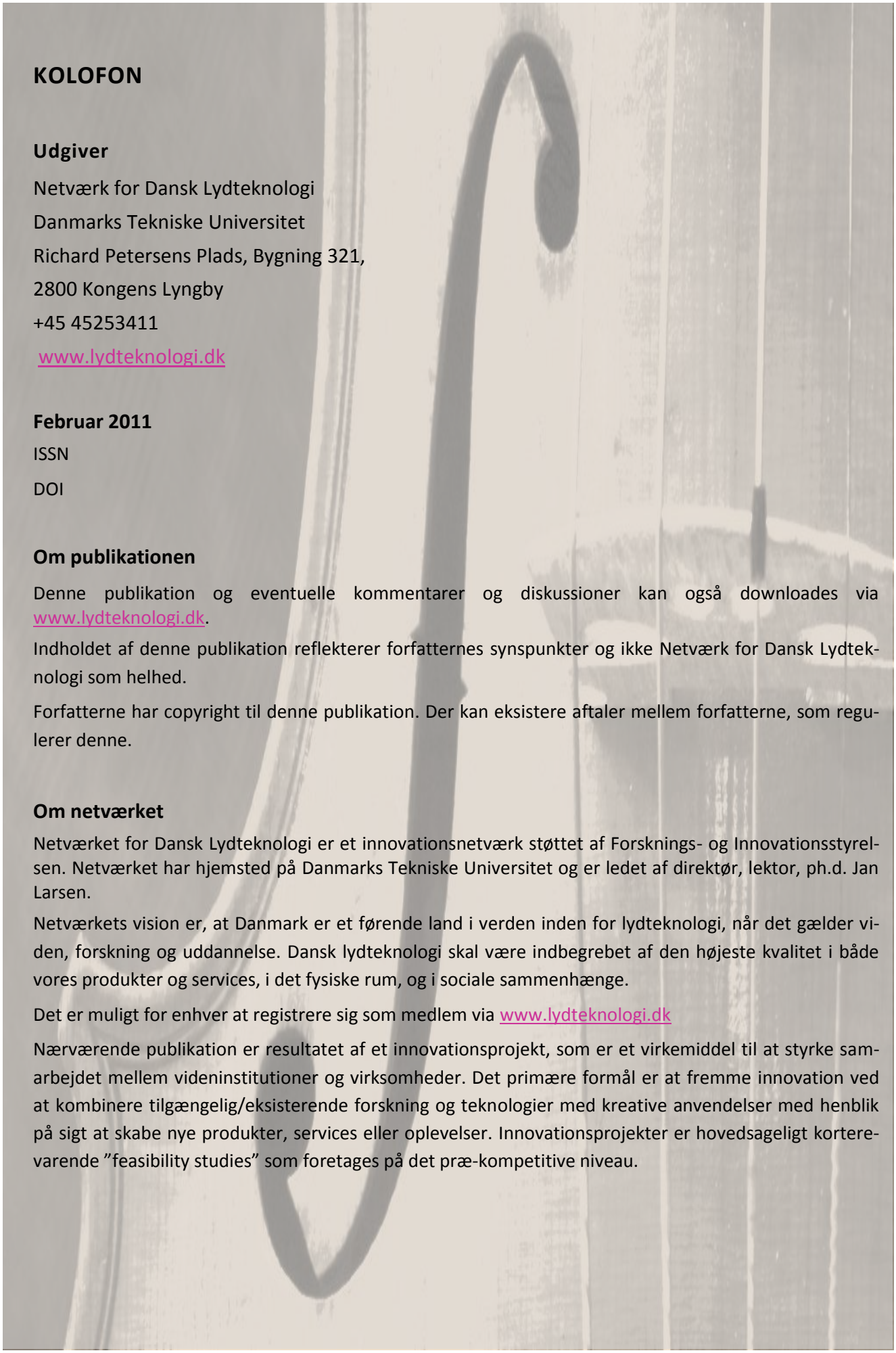
Geckon Aps, Peter Overgaard

Institut for Kunst og Kultur, Københavns Universitet, Bente Larsen, Anna Lawaetz

Danmarks Biblioteksskole, Birger Larsen



Projektet har implementeret et prototype-system, der for de involverede parter fik afdækket nogle af de aspekter af et lydindekseringssystem, der skal overvejes for at det kan benyttes af et større publikum. Systemet kan ud fra et input i form af en lydfil, generere en opdeling af lydfilen i sammenhængende segmenter, der kan beskrives ud fra nogle generelle labels. Segmenteringen finder segmenter med musik og tale – talesegmenterne opsummeres vha. en dansk automatisk talegenkender, for at kunne gøre talen søgbar. Det samlede system er implementeret som en web-baseret demo, som blev testet af en gruppe interesserede brugere ved en workshop. Projektet viste, at systemet kan give en opdeling der gør lydfilerne nemmere søgbare, men enkeldelene kræver stadig en del videreudvikling, før de er produktionsklare.



KOLOFON

Udgiver

Netværk for Dansk Lydteknologi

Danmarks Tekniske Universitet

Richard Petersens Plads, Bygning 321,

2800 Kongens Lyngby

+45 45253411

www.lydteknologi.dk

Februar 2011

ISSN

DOI

Om publikationen

Denne publikation og eventuelle kommentarer og diskussioner kan også downloades via www.lydteknologi.dk.

Indholdet af denne publikation reflekterer forfatternes synspunkter og ikke Netværk for Dansk Lydteknologi som helhed.

Forfatterne har copyright til denne publikation. Der kan eksistere aftaler mellem forfatterne, som regulerer denne.

Om netværket

Netværket for Dansk Lydteknologi er et innovationsnetværk støttet af Forsknings- og Innovationsstyrelsen. Netværket har hjemsted på Danmarks Tekniske Universitet og er ledet af direktør, lektor, ph.d. Jan Larsen.

Netværkets vision er, at Danmark er et førende land i verden inden for lydteknologi, når det gælder viden, forskning og uddannelse. Dansk lydteknologi skal være indbegrebet af den højeste kvalitet i både vores produkter og services, i det fysiske rum, og i sociale sammenhænge.

Det er muligt for enhver at registrere sig som medlem via www.lydteknologi.dk

Nærværende publikation er resultatet af et innovationsprojekt, som er et virkemiddel til at styrke samarbejdet mellem videninstitutioner og virksomheder. Det primære formål er at fremme innovation ved at kombinere tilgængelig/eksisterende forskning og teknologier med kreative anvendelser med henblik på sigt at skabe nye produkter, services eller oplevelser. Innovationsprojekter er hovedsageligt kortere-varende "feasibility studies" som foretages på det præ-kompetitive niveau.

FORORD

Denne publikation er resultat af innovationsprojektet 'Lydindeksering i store lydarkiver' finansieret af Netværk for Dansk Lydteknologi gennem en bevilling fra Forsknings- og Innovationsstyrelsen. Projektet er gennemført i perioden [indsæt periode] ledet af DTU, projektleder Lasse Mølgaard. Projektets øvrige projektdeltagere er: [indsæt projektdeltager], [indsæt projektdeltager], [indsæt projektdeltager].

SÆRLIGE OMSTÆNDIGHEDER

Beskriv evt. særlige forhold under tilblivelsen, udførelsen eller afslutningen på projektet.

INTRODUKTION

Dette projekt er startet for at undersøge mulighederne for at gøre lydarkiver tilgængelige og søgbare for almindelige, såvel som professionelle brugere.

BAGGRUND

I de hundrede år, der er blevet optaget lyd, er der opbygget massive mængder af data. Samtidig med eksplosionen af tekstuelt data på nettet vokser mængden af audio og video data tilsvarende. Anvendeligheden af søgemaskiner til audio er dog ikke i nærheden af det niveau, som det vi kender for tekstbaserede søgemaskiner.

Søgning efter og anbefaling af musik har et markedsmæssigt potentiale, der for nuværende bliver udnyttet i en række tjenester, der gennem brugerinteraktion bliver beskrevet og gjort søgbart. Desuden har man udviklet metoder til automatisk anbefaling af musik vha. sociale netværk at høste brugernes præferencer. Automatisk beskrivelse af musikkens indhold og popularitet er dog endnu ikke teknologi der har vist sig modent til et bredere publikum endnu.

Blandet lydmateriale der indeholder både musik, tale og andre lyde, som fx i radioudsendelser og podcasts findes i store mængder hos de forskellige større og mindre producenter af indhold. I den senere tid er der i større sprogområder gennemført projekter, der skal tilgængeliggøre den lydmæssige kulturarv primært igennem digitalisering af arkiverne, og udvikling af værktøjer til at gøre disse data søgbare. I USA har man fx i en del år arbejdet med projektet; The National Gallery of the Spoken Word (1), der bl.a. indeholder optagelser af en stor samling af præsidentielle taler og radioudsendelser fra hele det 20. århundrede. Ligeledes har de store nationale public service mediehuse store arkiver med lydoptagelser, der er meget sparsomt dokumenteret, og desuden er meget svært tilgængelige, indtil de bliver digitaliseret. I digitaliseringsprocessen er det nødvendigt at få et overblik over, hvad materialet indeholder, bl.a. for at kunne afgøre ophavsretsforhold, og faktisk at kunne genfinde bestemte udsendelser.

FORMÅL

Formålet er at udvikle et system for at kunne identificere sammenhængende segmenter i lydmiljøet, som brugere bevidst eller ubevidst registrerer, således at man kan finde kognitivt sammenhængende dele af lydoptagelser. Disse segmenter vil være dele, som kan beskrives dækkende af få, simple labels, der er intui-

tive for mennesker. Naturlige labels er fx lydtype (tale eller musik), identiteter af talere eller kunstneren der har skabt et stykke musik.

Denne segmentering vil blive baseret på en analyse af data gennem flere niveauer af signalbehandlingsfeatures. De karakteristika, der kan indgå spænder fra simple signal-features til karakteristika på højere abstraktionsniveauer, såsom semantiske emner, der tales om i lydoptagelserne.

Systemet udvikles på baggrund af radioudsendelser fra Danmarks radio. I denne sammenhæng er spørgsmålene: Hvad ønsker vi at få at vide om materialet? Hvordan kan analyse på store datamængder danne grundlag for humanistisk forskning? De store mængder data af varierende kvalitet og type, betyder at metoderne skal være både robuste og hurtige. Dette vil blive undersøgt i forbindelse med 3 cases:

- Popmusik-udsendelser: Hvordan kan det udviklede værktøj understøtte lokalisering af overgange mellem tale og musik? Kan værktøjet give os et bedre overblik over opbygningen af musikprogrammer? I hvor høj grad varierer de over tid? I forhold til den enkelte studievært? I forhold til bestemte, ekspliciterede programkoncepter? Kan musikken på en meget foreløbig måde genrebestemmes ved hjælp af værktøjet?

- Nyheder: Hvor langt kan kombinationen af segmentanalyse og talegenkendere bringe os i bestemmelsen af nyhedernes indhold og opbygning. Kan man på denne baggrund lokalisere variationer i opbygning af udsendelserne over tid? Kan man lokalisere lighedstræk (hverdag) og forskelle (ved fx skelsættende begivenheder) i opbygning og fremstilling? Og endelig: Hvor langt kan videreudviklingen af SPHINX bringe forhold til de transskriptionsprogrammer, som anvendes i lingvistisk forskning?

- Kulturprogrammer med blandede genrer: Harddisken '98, '99, '00, ’01: Kan man udpege en særlig programrytme - en særlig rytmisk signatur i netop denne periode, hvor Harddisken fremtræder som et markant og genremæssigt nytænkende radioprogram? Kan man bestemme udsendelsernes væsentlige temaer i perioden på baggrund af det udviklede værktøj? Kan man endelig bestemme de enkelte indslags tilhørsforhold til bestemte genrer (interviewet, montagen, reportagen osv.?)

Talegenkendelsen vil yderligere gøre det relevant for forskere fra Danmarks Biblioteksskole, som vil undersøge hvor godt de automatisk producerede transskriptioner egner sig til indholdsbeskrivelse af radiofonisk materiale i forhold til indeksering og informationssøgning. Det vil i særdeleshed blive undersøgt,

- Hvilke muligheder giver outputtet fra talegenkendelse anvendes til indeksering og genfinding af radiofonisk materiale på dansk?

- Hvordan kvaliteten af talegenkendelsen påvirker kvaliteten af søgeresultaterne (genfindingsperformance)?

Hvilke styrker og svagheder outputtet fra talegenkendelse har i forhold til andre typer metadata?

IMPACT/EFFEKT

At have et fungerende indekseringssystem er en vigtig byggesten i at kunne forbedre tilgængeligheden af multimedialt indhold på internettet, i tv-settop-bokse og andre medieplatforme. Der findes allerede en række video- og lyd-distributionsplatforme, men søgning er baseret på de metadata, som indholdsproducenten leverer. Disse data kan meget vel være brugbare for korte klip, fx et typisk youtube-klip, der viser en enkelt filmet scene, eller behandler et enkelt tema. Længere udsendelser kræver i stedet at producenten har tilføjet tidskoder, hvis slutbrugerne hurtigt skal finde et indslag, fx i en nyhedsudsendelse.

Et lydindekseringssystem kan således tænkes at blive et værktøj for journalister, researchere og andre, der har brug for at genfinde lydklip. Denne mulighed for at "browse" i udsendelser kan også benyttes i slutbrugerprodukter, f.eks. for at forbedre fast forward-funktionen i digitale afspillere.

METODE OG RESULTATER

Dette afsnit præsenterer de metoder, der er benyttet for at implementere den demo-applikation, som er outputtet af dette projekt. Desuden gennemgås de erfaringer, der er blevet gjort med det implementerede system.

TEORI

Systemet er baseret på en række metoder inden for signalbehandling og machine learning udviklet indenfor audioanalyse indenfor det sidste årti. Metoderne til systemet kan dels op i tre overordnede dele, den signalbehandlings-orienterede initielle opdeling af signalet, talegenkendelse og tekstbaseret emne-opdeling.

AUDIO-ANALYSE

Den første del af et søgesystem til tale benyttes til at identificere forskellige lyd klasser. De fire vigtigste klasser er tale, musik, baggrundsmiljø og stilhed, men afhængigt af produktionen af det audio signal, der betragtes, kan man betragte mere specifikke audioklasser, såsom tale i et støjfyldt miljø, tale over musik, og forskellige klasser af støj. Segmentering og klassificering af audio er blevet undersøgt i en række sammenhænge vha. en række forskellige signalbehandlingsmetoder over det seneste tiår.

Generelt kan klassifikations og segmenteringssystemer deles op i to dele, udtrækning af karakteristika (features) fra det primære signal, og den efterfølgende matematiske model, der benyttes til at træffe beslutning om der baseret på features er tale om den eller anden klasse af signaler.

Alt efter hvilke klasser af lyd, der skal adskilles er forskellige features blevet foreslået, baseret på betragtninger om de karakteristika, der adskiller tale, musik og andre mulige klasser af lyd. Disse features kan generelt deles op på grundlag af den tidshorisont, over hvilken de er udvundet, og hvor beregningskrævende de er. De mest basale features omfatter tidslige og spektrale features. Tidsdomæne-features inkluderer måling af energien i signalet eller nulpunktskrydsninger, mens spektrale features opsummerer frekvensindholdet på en passende måde. Typisk analyseres lyd på basis af lydbidder på 20-30 ms, idet signalet er stabilt på denne skala. For at kunne karakterisere egenskaber ved lyden er det fornuftigt at opsummere flere af disse korte klip, således at en klassifikation foregår på basis af 200ms til 1 sekund lyd.

Ud over de basale features har psykoakustisk inspirerede kepstrale koefficienter været anvendt med stor succes i talegenkendelsessystemer, og efterfølgende har kepstral analyse også vist sig at være effektiv i generelle lydklassifikationsopgaver (2). Andre features og måder opsummere features over tid er også blevet foreslået på grundlag af psykoakustiske observationer, se fx (3). En mere uddybende gennemgang af features til lydklassifikation kan findes i (4).

Givet et passende sæt af features kan data klassificeres vha. en klassificeringsalgoritme. Alt efter den kontekst metoderne er blevet implementeret i spænder de beskrevne tilgange til klassifikationen fra heuristiske regelbaserede metoder og mere strukturerede tilgange, der benytter data-drevne modeller.

De heuristiske metoder er generelt baseret på at udtrække nogle simple regler der benytter tærskelværdier for de værdier features kan antage. Da disse metoder afhænger af faste tærskelværdier, er de generelt ikke robuste over for skiftende forhold, men kan være en mulighed for realtidsafvikling af klassifikationen.

Modelbaserede tilgange kan inkludere generative modeller, såsom Gaussian Mixture Models (GMM), eller diskriminative modeller, såsom neurale netværk eller Support Vector Machines (SVM). Disse modeller er

velkendte metoder til at løse klassifikationsproblemer, og er udmærkede valg i denne sammenhæng. De nævnte modeller antager i udgangspunktet at features er uafhængige og ens fordelte (i.i.d.), hvilket kan være en rimelig antagelse. Generelt er det dog ikke tilfældet for audio, hvor der er en tidslig afhængighed i data. Nogle projekter har derfor også modelleret det tidsmæssige forløb af features, fx vha. Skjulte Markov Modeller (HMM), som dog er ganske beregningstunge.

En anden metode til generel opdeling af lyd i homogene segmenter, er at detektere skift i signalet. Dette kan udgangspunktet gøres mere uafhængigt af det indledende valg af klasser lyden kan inddeles i. Dette er fx fornuftigt for at kunne adskille forskellige talere i større samlinger af radioudsendelser, hvor de medvirkende personer i udsendelserne ikke er kendt på forhånd. I dette projekt benyttes denne tilgang derfor til at adskille talere.

Hvis de medvirkende talere i en udsendelse er kendt på forhånd, kan der trænes en superviseret model for hver taler, og audio signalet kan klassificeres i overensstemmelse hermed, som beskrevet ovenfor. Hvis identiteten af talerne ikke er kendt på forhånd må ikke-superviserede metoder anvendes. Metoder til ikke-superviseret detektion af skift mellem talere kan groft opdeles i to klasser, nemlig energi-baserede og metrik-baserede:

Energi-baserede metoder bygger på at detektere pauser, dvs. hvornår lydsignalets energi kommer under et vist niveau. Skift vil således typisk være baseret på at finde perioder med stilhed. I radioudsendelser kan klipningen være ganske aggressiv, så der kun er meget kort eller ingen stilhed mellem forskellige personer, hvilket gør denne fremgangsmåde mindre attraktiv.

Metrik-baserede metoder måler dybest set forskellen mellem to på hinanden følgende vinduesfunktioner, der flyttes langs lydsignalet. En række af afstand foranstaltninger er blevet undersøgt baseret på forskellige parametriske metoder.

En sidste strategi, der bl.a. er beskrevet i (5) og (6), er at udvikle hybridmetoder, der anvender ikke-superviserede metoder til at foretage en første opdeling, hvorefter disse segmenter benyttes til at opbygge superviserede modeller.

TALEGENKENDELSE

Der er ikke implementeret nogen videre behandling af de dele af udsendelsen, der indeholder musik, da der allerede eksisterer kommercielle systemer, der kan genkende fingerprints af musiknumre, fx Shazam (7), og karakterisere musik, fx Echonest (8).

Transskription af segmenterede tale sker ved hjælp af et automatisk system til talegenkendelse. I de seneste årtier er der forsket meget i at udvikle systemer med stort ordforråd, brugeruafhængig til brug for nyhedsudsendelser, og mange frit tilgængelige systemer er blevet udviklet. En stor del af problemerne for disse systemer er nyhedsudsendelser indeholder tale optaget i støjfyldte og andre ikke-ideal miljøer.

Talegenkendelsessystemer med stort ordforråd er typisk baseret på Hidden Markov Model-baseret afkodning. Praktiske implementeringer af disse systemer benytter forskellige heuristikker til at gøre dekodning mulig indenfor rimelig tid. Efterhånden er disse systemer blevet finpudset og optimeret, så de i dag har opnået et rimeligt performanceniveau, der dog stadig er langt fra menneskelig taleforståelse. Et af de primære uløste problemer for talegenkendelse er skiftende baggrundsstøj, hvilket kræver, at systemet kan detektere og tilpasse sig de skiftende omstændigheder.

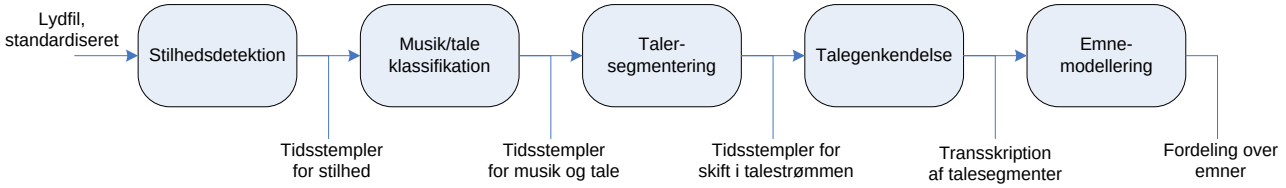
En række akademiske grupper har udviklet ASR-systemer, der frit til rådighed for eksempel under GNU Public License. Populære frit tilgængelige systemer (til akademiske formål) omfatter Hidden Markov model Toolkit (HTK), Sonic (9), og SPHINX (10). SPHINX systemet er et helt økosystem med applikationer til at træ-

ne nye akustiske modeller, modeller trænet på kendte databaser, og implementeringer af systemet til flere forskellige platforme, inklusiv en mobiltelefon-baseret variant.

Formålet med dette projekt har ikke været at udvikle eller forbedre talegenkendelse, derfor særlige optimering eller ændring af talegenkendelse. I stedet benyttes et fuldt udviklet ASR-system, der kan håndtere et stort ordforråd, er taleruafhængigt, kan håndtere almindelig tale og hurtigt kunne tages i brug.

IMPLEMENTERING

Systemet er implementeret som et web-baseret system, hvor brugeren kan uploade en lydfil til en server og få de genererede annoteringer af lydfilen præsenteret på en webside. Demoen er i udgangspunktet lavet til at kunne levere simple annoteringer til de uploadede filer, som så kan lægges hen over lydfilerne for



Figur 1 Pipeline benyttet for at generere output til websiden

at evaluere den struktur, som algoritmerne har frembragt.

Lydfilerne uploades via web-interfacet til serveren, hvorefter den bliver behandlet i den pipeline, der vises i Figur 1, med de følgende skridt:

STILHEDSDETEKTION – dette skridt er indført for at finde pauser i udsendelserne, som kan indikere, at der er et klart skift mellem to forskellige uafhængige dele af udsendelsen. Stilhedsdetektionen er implementeret ved at udregne signalenergien i et vindue på 20 ms, og registrere stilhed, når energien kommer under et fastlagt niveau. Dette er en meget simpel metode, der fungerer meget hurtigt, men ikke er alt for robust overfor støj. Denne del er implementeret i Matlab.

MUSIKKLASSIFIKATION – benyttes til at finde dele af udsendelserne, der indeholder musik. Denne opdeling er ganske naturlig, da musik typisk er et element i en lydoptagelse, der semantisk adskiller sig fra segmenter med tale, eller baggrundslyde. Klassifikationen af lyden i musik eller tale sker, baseret på en række features, der inkluderer korttidsenergi, nulpunktskrydsningsraten, spektral flux, og mel kepstrale koefficienter. Disse features blev undersøgt i (4), og var tilstrækkelige til at adskille musik og tale, og benyttes derfor her. Disse features udtrækkes for hver 20 ms vindue. Hvert sekund klassificeres baseret på middelværdien og varians af hver af de features, der er nævnt herover. Klassifikationen foregår vha. en lineær klassifikationsalgoritme, der er trænet på en samling bestående af 50 segmenter med musik og 50 segmenter med tale. Hvert segment var 45 sekunder langt, således at der i alt er brugt 4500 sekunder lyd. Træningen af klassifikationsalgoritmen resulterede i at 14 features blev valgt som input til klassifikationen. Metoderne til musikklassifikation er implementeret i Matlab, vha. pakken Voicebox (11) til udtrækning af features.

TALERSEGMENTERING – benyttes på de segmenter, der er imellem dele mellem stille passager og musiksegmenter. Semantisk er det simplest detekterbare skift i en samtale, skift mellem de talere der deltager. Et sådant skift kan i en nyhedsudsendelse være når en vært giver ordet videre til en reporter, eller bare skift mellem to personer i samtale.

Implementering af taleradskillelse er bygget på en ikke-superviseret metode, der i udgangspunktet er baseret på at sammenligne to vinduer, der bevæges hen over signalet. Når forskellen mellem disse to vinduer overstiger en tærskelværdi, antager vi, at der er sket et skift mellem talere. Kvantiseringen af

forskellen mellem de to vinduer sker, baseret på mel kepstrale koefficienter som features. Hvert vindue beskrives ved at foretage en k-means clustering på de MFCC'er der er indenfor hvert vindue, og herefter udregne forskellen mellem vinduer vha. vektor kvantiseringsafstanden, som beskrevet i (12). Metoden til at finde skift skal balancere imellem høj præcision, altså at de skift der foreslås er rigtige skift, og genkaldelsesraten, dvs. at man finder alle de rigtige skift der forekommer. For at opnå dette har vi implementeret en to-trins metode, hvor første trin, benytter en fast vinduesstørrelse og er justeret, således at genkaldelsesraten er høj. I andet trin evalueres de skift, der blev foreslået af første trin, ved at benytte et større vindue, hvorved vi kan fjerne nogle af de falske skift, som det første trin har foreslået. Denne metode er baseret på arbejdet beskrevet i (12). Metoderne til musikklassifikation er implementeret i Matlab, vha. pakken Voicebox (11) til udtrækning af MFCC'ere.

TALEGENKENDELSE – Udføres på de fundne talesegmenter, vha. en talegenkender der er baseret på SPHINX-systemet (10) som den underliggende engine. Da systemet skal kunne transskribere dansk tale er der i dette system trænet en dansk akustisk model, såvel som en sprogmodel baseret på et korpus med dansk tale bestående af omkring 10 timers transskriberet audio. Systemet kan potentielt genkende et ordforråd på 44239 ord.

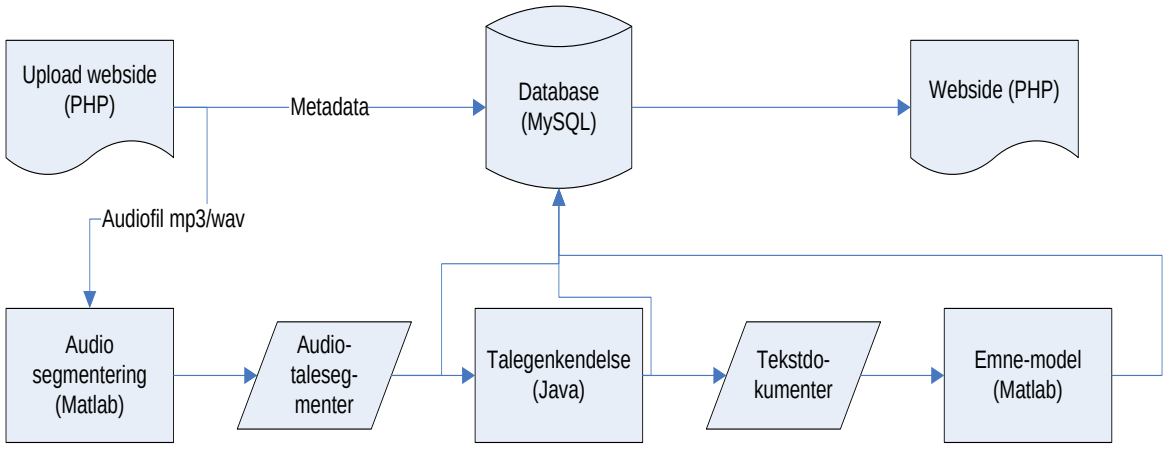
EMNE-GENKENDELSE – er implementeret ved at opbygge bag-of-words-repræsentationen for hver af de segmenter, der er blevet transskriberet, altså at tælle antallet af forekomster af ord, og opbygge en vektor med disse counts. Samlingen af segmenter giver dermed en matrix **X**, som kan benyttes til at lave en matematisk model, der finder semantiske emner.

Emnerne udtrækkes vha. Non-negative matrix factorisation (NMF), der faktoriserer dokument-matricen i to matricer, én der giver vægtningen af ordene indenfor emnerne **W**, og én der giver dokumenternes tilhørsforhold til emnerne **H**. Således at dokument-matricen **X** skrives som:

$$\boldsymbol{X} \cong \boldsymbol{WH}, \boldsymbol{X}_{ij} \geq 0, \boldsymbol{W}_{ik} \geq 0, \boldsymbol{H}_{kj} \geq 0$$

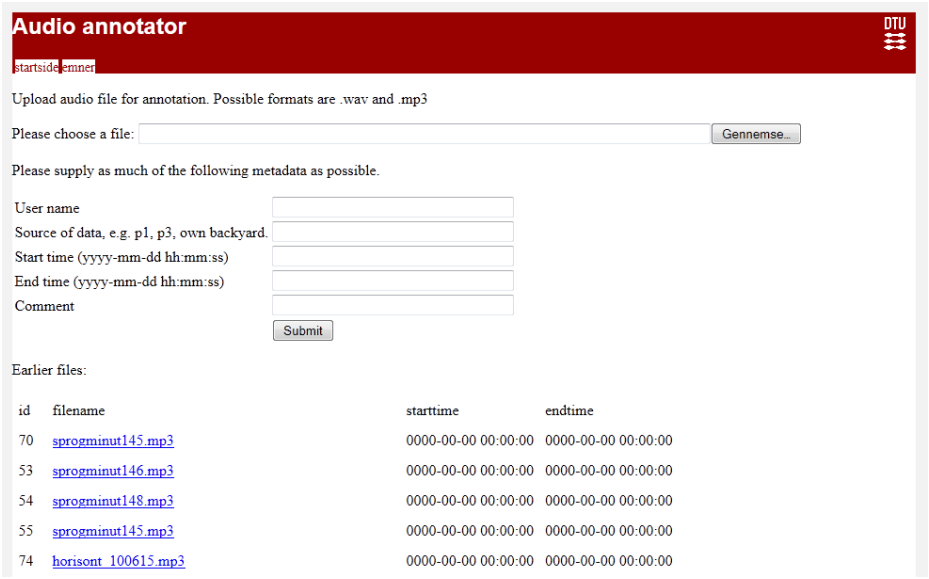
Udtrækningen af **W** og **H** matricerne sker ved hjælp af en stokastisk gradient metode, der bygger på at minimere kvadratet på fejlen mellem **X** og **WH**. Den benyttede NMF-metode er beskrevet i (13). Emne-udtrækningen er også implementeret i Matlab, ved brug af NMF-toolboxen.

Disse automatiske analyse-trin præsenteres for brugerne vha. en web-baseret demo, der er tilgængeligt på http://castsearch.imm.dtu.dk/larm. Output fra de forskellige analyse-trin bliver gemt i en MySQL-database, som kan tilgås via den hjemmeside der er programmeret i PHP, som vist i Figur 2.



Figur 2 Flowchart over implementering af web-baseret demo

Web-siden der bruges til at uploade filer til analyse er vist i Figur 3. Denne giver mulighed for at tilknytte nogle få metadata, der gør det muligt at genfinde den uploadede fil i systemet.



Figur 3 Web-interface til upload af filer, der skal analyseres.

En uploadet fil bliver analyseret på serveren, hvorefter outputs fra disse analyser kan ses på en webside, som den der er vist i . De segmenter som analysen har genereret kan hentes i tekst-format, således at de kan læses ind i et audio-program som labels, fx i Audacity[1], og ses repræsenteret i et enkelt layout på web-siden.

EKSPERIMENTER

Udviklingen af systemet er primært blevet testet ved en workshop afholdt for potentielle slutbrugere af systemet. Projektet var i udgangspunktet initieret for at undersøge mulighederne for at benytte automatisk lydindeksering for at tilgængeliggøre ikke-annoteret lyddata til slutbrugere.

De enkelte analyse-komponenter, der indgår i systemet, er blevet testet i løbet af implementeringen på små samlinger af lyd, der er indsamlet til de opgaver. Samlingerne er nærmere beskrevet i (4)

- Musikklassifikationen blev, som tidligere nævnt, testet på en samling af ren tale og ren musik á 4500 sekunder, hvor performance blev evalueret ved at klassificere hvert af disse sekunder som musik eller tale.
- Talesegmenteringen blev evalueret på en samling af tale fra nyhedsudsendelser, som manuelt er blevet annoteret med skift mellem talere. Samlingen består af 103 minutters tale, hvor der foregår 388 skift mellem talere.

De enkelte dele blev således testet så de var optimeret til de beskrevne samlinger. En samlet evaluering af delene sat sammen blev mere kvalitativt evalueret.

Projektets interessenter mødtes til workshop d. 14. maj 2010, på Institut for Kunst og Kultur, Københavns universitets Amager-campus. På den afholdte workshop blev web-demoen præsenteret, og de deltagende fra Institut for Kunst og Kultur og Biblioteksskolen fik mulighed for at teste systemet med lydfiler, som de havde interesse i at få analyseret.

Ved at downloade analyse-resultaterne, og læse dem ind i Audacity, kunne deltagerne evaluere hvor godt de automatiske analyser stemte overens med lydfilen. Disse evalueringer blev udført på 3 forskellige lydfiler, hvoraf en var en radioavis, den anden en morgenandagt, og en tredje et taleprogram fra DRs P1. Denne

[1] http://audacity.sourceforge.net/

evaluering blev primært af kvalitativ karakter, idet testbrugerne uploadede deres filer, hvorefter der var en diskussion af brugbarheden af de genererede annoteringer.

RESULTATER

Som afslutning på projektet var annoterings-demoen implementeret og den basale funktionalitet, med upload, registrering af uploadede filer, og de forskellige trin af annoteringen var implementeret og kørte på web-serveren, som beskrevet i Implementerings-afsnittet.

Evalueringerne på workshop'en af de automatiske annoteringer gav en række overordnede betragtninger om brugbarheden af systemet til analyse af lydarkiverne.

Overordnet set blev det tydeligt, at der er behov for et bedre værktøj til at evaluere de automatiske annoteringer. Audacity er en god brugergrænseflade til at overskue lydsignalet og spektrogrammet, men i dette tilfælde er det interessante at kunne benytte annoteringerne aktivt til at finde dele af lyden, og at kunne rette de automatiske annoteringer.

De enkelte filer blev ikke analyseret kvantitativt, men der var generelt det problem at filerne blev opdelt i for korte segmenter pga. pause-detekteringen. Hermed blev segmenterne hakket i meget korte dele, som nogle gange bestod af enkelte ord, hvilket ikke umiddelbart var brugbart for testerne.

Generelt var talegenkendelsen heller ikke af en kvalitet, der var brugbar til at genkende emnerne for det der der blev talt om i de afprøvede lydfiler. Hermed blev den automatiske emne-detektion heller ikke brugbar.

KONKLUSION

Projektet har implementeret et prototype-system, der for de involverede parter fik afdækket nogle af de aspekter af et lydindekseringssystem, der skal overvejes for at det kan benyttes af et større publikum.

Et aspekt er nødvendigheden af at have et standardiseret annoteringsværktøj, der let kan bruges til at evaluere og korrigere de automatisk genererede annoteringer. For at udvikle et større lydindekseringssystem vil det være nødvendigt med en organiseret manuel annoteringsindsats, hvilket også vil indebære, at et mere effektivt annoteringsprogrammel skal findes eller udvikles.

Da projektet havde som mål at implementere hele kæden af annoterings-metoder, blev de enkelte led i kæden ikke testet helt så indgående, som det kunne være ønsket. Dermed kunne vi se at nogle typer radioudsendelser, der har et specielt udtryk, som for eksempel en morgenandagt, hvor en taler har et næsten syngende udtryk vil blive forvekslet med musik.

DTU har fået mulighed for at rulle de teknologier, de kendte til, til et bredere publikum, og dermed få kvalitativ feedback på performance af disse metoder.

Aalborg Universitet har leveret talegenkendelse, og hermed testet en anvendelse af dansk talegenkendelse. Det var klart fra udgangspunktet, så ville systemet, baseret på de tilgængelige træningsdata ikke kunne opfylde de krav, der ville være til talegenkenderen i form af de støjfyldte taleoptagelser, den blev testet med. Det er således klart AAUs system mangler mere træningsdata, hvilket vil blive forsøgt indsamlet i et senere projekt.

DR og Geckon har i udviklingen af platformen til distribution af deres arkivmateriale fået evalueret mulighederne for automatisk at annotere de mange tusind timers materiale der er til rådighed. Dette gav bl.a. anledning til videreudvikling af brugerinterfacet for at kunne opfylde nogle af de ideer, der blev diskuteret til evaluerings-workshop'en.

Geckon er en nystartet virksomhed med 8 ansatte og dette projekt var det første samarbejde med videninstitutioner.

Institut for Kunst og Kultur og Danmarks biblioteksskole har deltaget i projektet ud fra en rolle som (super)brugere af det udviklede system. Dette har gjort det mere klart for de deltagende studerende og forskere, hvilke muligheder og begrænsninger, der teknologisk er med humanistisk forskning i de digitaliserede radioarkiver. Disse erfaringer kan videreføres i de projekter, der vil indgå i LARM-projektet.

Kort beskrivelse af impact/videreførelse for alle deltagere i projektet.

Følgende elementer skal inkluderes af hensyn til såkaldt performance regnskab (udelad kun hvis det ikke er relevant eller muligt at besvare):

Nye produkter, ydelser og processer

Nye ideer, der senere kan føre til innovation

Nye kompetencer, der senere kan føre til innovation

Deltagere, der deltager i deres første samarbejde med en videninstitution

Deltagende virksomheder med under 50 ansatte

REFERENCER

1. **Hansen, J.H.L., et al., et al.** "SpeechFind: Advances in Spoken Document Retrieval for a National Gallery of the Spoken Word. *Speech and Audio Processing, IEEE Transactions on.* Sept. 2005, s. vol.13, no.5, pp. 712- 730 .

2. **J.-L. Gauvain, L. Lamel, and G. Adda.** The limsi broadcast news transcription system. *Speech Communication,*. 37, 2002, Årg. 1, 2, s. 89–108.

3. **Breebaart, M. F. McKinney and J.** Features for audio and music classification. *Proc. of ISMIR.* 2003.

4. **Jørgensen, Kasper W. og Mølgaard, Lasse L.** *Tools for automatic audio indexing.* Kgs. Lyngby : DTU Informatics, 2006.

5. **T. Kemp, M. Schmidt, M. Westphal, and A. Waibel.** Strategies for automatic segmentation of audio data. *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing.* 3, 2000, s. 1423–1426.

6. **H.-G. Kim, D. Ertelt, and T. Sikora.** Hybrid speaker-based segmentation system using model-level clustering. *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing.* 2005, Årg. 1, s. 745–748.

7. *Shazam.* [Online] http://www.shazam.com/.

8. *Echonest.* [Online] http://the.echonest.com/.

9. **Pellom, B. og Hacioglu, K.** Recent improvements in the CU Sonic ASR system for noisy speech: the SPINE task. *Proceedings, 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03).* 2003, Årg. vol.1 pp. I-4- I-7.

10. **Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, Joe Woelfel.** *Sphinx-4: A Flexible Open Source Framework for Speech Recognition.* s.l. : SUN Microsystems, 2004. TR2004-0811 .

11. VOICEBOX: Speech Processing Toolbox for MATLAB. [Online] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

12. **Jørgensen, Kasper W., Mølgaard, Lasse L. og Hansen, Lars Kai.** Unsupervised Speaker Change Detection for Broadcast News Segmentation. *Proceedings of European Signal Processing Conference.* 2006.

13. **Mølgaard, Lasse L., Jørgensen, Kasper W. og Hansen, Lars K.** Castsearch - Context Based Spoken Document Retrieval. *Proceedings of International Conference on Acoustics, Speech, and Signal Processing.* 2007.

APPENDIKS