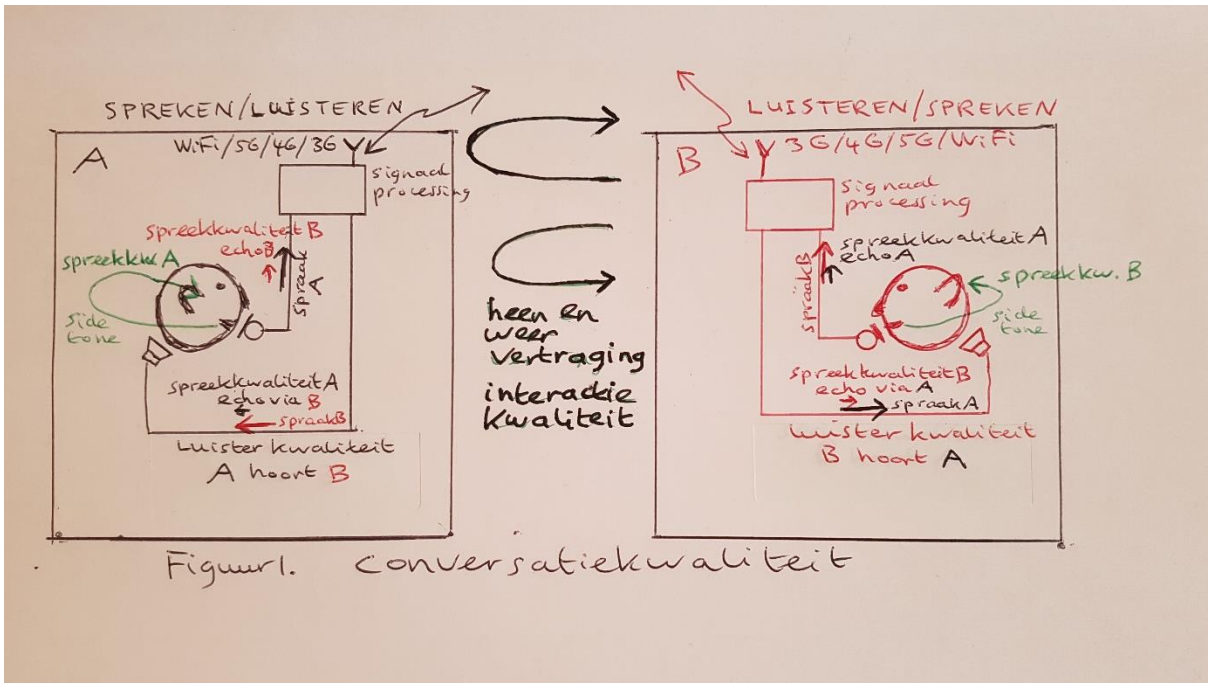
HiFi over WiFi ?

Waarom worden we toch zo moe van Zoomen, Teamsen, Facetimen, Skypen, enz.? We zien vaak argumenten gerelateerd aan het cameragebruik, dat het vervelend is om jezelf te zien, dat het beeld niet synchroon loopt met het geluid, of dat bijvoorbeeld oogcontact maken moeilijk is. Zet je die camera echter uit, dan blijkt: die spraakverbinding over je computer is gewoon slecht, slechter dan een telefoonverbinding. Nu is een telefoonverbinding weliswaar een 1-op-1-verbinding, maar hoe moeilijk kan het zijn om een stem op te nemen en met hoge kwaliteit weer te geven? In principe is het niet zo moeilijk, en toch is de spraakkwaliteit van Zoom- of Teams-verbinding vaak heel slecht. Hoe komt dat toch?

Er wordt veel onzin beweerd in de wereld van HiFi maar een stem goed opnemen en weergeven is niet echt moeilijk, als je zestiger jaren hits hoort dan valt vaak de perfecte stem op.. Hoe komt het dan dat al die moderne apps zo slecht klinken? Komt dat misschien omdat HiFi over WiFi moeilijk is? Dat speelt een rol maar het is zeker mogelijk om goede audiokwaliteit over WiFi leveren. De werkelijke oorzaken van slechte audiokwaliteit, of het nu over telefonie gaat of over het weergeven van muziek, liggen dieper en zijn lastig kort en helder uit te leggen.

Om te beginnen gaat het bij die apps bijna altijd om conversatiekwaliteit, iets waar spraakkwaliteit een belangrijk onderdeel van is, maar niet het enige. Er zijn drie elementen die een rol spelen bij conversatiekwaliteit: luisterkwaliteit, spreekkwaliteit en interactiekwaliteit (zie Figuur 1).



Een merkwaardige, maar belangrijke eigenschap van conversatiekwaliteit is de spreekkwaliteit. Deze hangt direct samen met hoe goed je jezelf terug hoort. Bij een live gesprek is deze spreekkwaliteit prefect. Als je echter over een app met iemand communiceert wordt de spreekkwaliteit, en daarmee de conversatiekwaliteit minder. Als je een koptelefoon gebruikt dan mis je een directe terugkoppeling van wat je zegt. Klassieke telefoons lossen dit op door je eigen stem ook terug te voeren in de luidspreker van de hoorn. Daardoor blijf je jezelf goed horen, zelfs als je de telefoon strak tegen je oor aan houdt en er veel omgevingslawaai is. Dit noemen we een side toon. Bij het gebruik van een koptelefoon, of een “oortje”, wordt deze side toon vaak weggelaten, simpelweg omdat het niet eenvoudig is om je eigen

stem op een goede manier terug te koppelen naar je eigen oor. Voor artiesten die live optreden met een oortje is dat ook vaak een probleem. Denk maar terug aan het Eurovisie songfestival, waarbij een slecht afgeregeld oortje vaak de oorzaak is van het mislukken van (proef)optredens.

De beste spreekkwaliteit krijgen we door het pad van je eigen stem naar je eigen oren niet te blokkeren. Dit kan bijvoorbeeld door luidsprekers te gebruiken, alleen die hebben jammer genoeg als groot nadeel dat er echo's kunnen ontstaan bij de andere partij. Deze echo's kunnen met geavanceerde signaalprocestechnieken worden onderdrukt, maar dit gaat weer ten koste van de spraakkwaliteit die je gesprekspartner ervaart.

Naast spreekkwaliteit is er nog een belangrijke factor die de conversatie kwaliteit kan domineren: de interactiekwaliteit. Deze wordt voor het grootste deel bepaald door de vertraging in de verbinding. Om snel "heen en weer" te kunnen praten moet de vertraging laag zijn. De maximaal acceptabele vertraging is echter moeilijk te definiëren: voor een monoloog is het geen enkel probleem als de vertraging langer is dan een seconde, en als je de gesprekstrategie aanpast, door voor lief te nemen dat je bij iedere vraag even moet wachten op een antwoord, is een seconde vertraging ook niet erg. Als je echter snel en efficiënt wilt onderhandelen is een halve seconde al te veel. De Japanse PTT (NTT) heeft begin jaren negentig al aangetoond dat een "heen en weer"-vertraging van 0.1 seconde al waarneembaar is en dat voor een goede conversatiekwaliteit de vertraging onder de 0.2 seconde moet liggen. Je kunt controleren of de vertraging in een verbinding laag genoeg is door interactief tot 10 te tellen (ik 1, jij 2, ik 3,…totdat je 10 hoort), dat kan in 4 à 5 seconden in een live setting. Bij 6 seconden is de vertraging al waarneembaar en bij 10 seconden is de vertraging onacceptabel lang.

Nu zijn alle moderne netwerken (3G, 4G, 5G, WiFi) ontworpen voor efficiënt datatransport en dat betekent grote pakketten, met hertransmissie bij pakketverlies, wat resulteert in een hoge "heen en weer" vertraging. Voor diensten als radio, TV, Netflix, Spotify, of in het algemeen voor alle streamingsdiensten, levert een vertraging in het netwerk nauwelijks problemen op. Bij conversatiediensten zoals telefonie en beeldbellen/vergaderen moet de "heen en weer"-vertraging onder de halve seconde blijven en als je beeld en geluid synchroon wilt laten lopen is dat een zware eis. In de praktijk wordt dat probleem opgelost door spraak en video te comprimeren in kleine pakketten en die zo snel mogelijk over het Internet te sturen, en als een pakket verloren gaat wordt het "gat" wat ontstaat vaak gevuld met een schatting van het signaal.

Zelfs als de "heen en weer"-vertraging onder de halve seconde blijft, de spreekkwaliteit in orde is, en de spraak efficiënt en correct wordt verstuurd via een netwerk zodat alle pakketten aankomen, zelfs dan is de conversatiekwaliteit vaak slecht. Dat komt doordat bij veel spraakverbindingen de microfoon te ver af staat van de mond. Je zou zeggen dat is geen probleem, als ik op 1 meter afstand met iemand praat klinkt het goed, waarom moet die microfoon dan zo dichtbij? Het is in het algemeen binnen de HiFi audio wereld een bekend fenomeen, als je iets opneemt op een plaats waar het goed klinkt dan klinkt die opname bij het afspelen bijna altijd heel slecht. We kunnen dit aantonen met een eenvoudig luister experiment, pak je smartphone en maak een opname van een stem op 1 cm en op 1 m. Als je jezelf terugluistert dan klinkt de 1 m opname heel slecht. Dit komt doordat onze oren zo goed ontwikkeld zijn dat ze altijd proberen om het geluid wat we horen te compenseren voor de ruimte waarin we ons bevinden en dat kan bij een opname helaas alleen voor de ruimte waarin het geluid wordt weergegeven en niet voor de ruimte waarin het geluid is opgenomen.

Je kunt er natuurlijk voor zorgen dat je de microfoon dicht bij je mond houdt en dat de pakketten van conversatiediensten voorrang krijgen in het netwerk (Quality of Service). Het resultaat hiervan kennen we al, namelijk de klassieke telefoniediensten (Plain Old Telephony Service, POTS). Het beste is om

een klassieke telefoonhoorn te gebruiken waarbij je stem op minder dan 1 cm van de microfoon wordt *ingekoppeld* en het spraakgeluid dicht op de oorschelp wordt *uitgekoppeld*. Als je met zo'n telefoon in een rumoerige omgeving zit druk je de hoorn tegen je oor en de gesprekskwaliteit blijft perfect. Het is zelfs zo dat bij een klassieke telefoon je eigen stem correct wordt weergegeven in de luidspreker van de hoorn zodat je ook jezelf goed blijft horen, precies de side tone die vaak ontbreekt in online gesprekken. Ondanks dat de moderne smartphone minder goed is voor het in- en uitkoppelen van spraak van je mond en naar je oor, kunnen we toch in veel gevallen een hoge conversatiekwaliteit halen dankzij moderne signaalprocestechnieken. Dit effect is vooral merkbaar als we twee smartphones gebruiken met High Definition voice (HD voice). Stel dus bij de eerstvolgende uitnodiging om te Zoomen voor om gewoon te bellen, bij voorkeur met een HD voice setting, wat vaak automatisch het geval is zolang je beide met dezelfde operator belt.

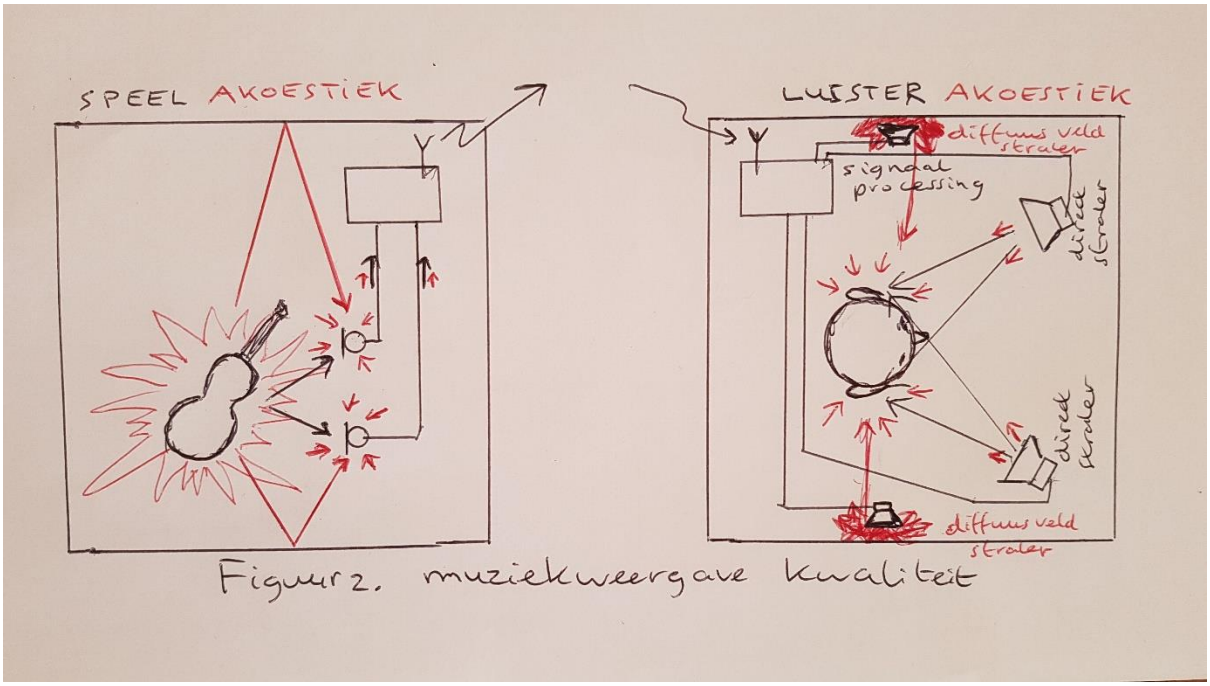
Als we analyseren waarom een HD voice verbinding goed klinkt komen we op twee hoofdpunten, luidheid en natuurlijkheid/timbre (klankkleur). Beide zijn in orde en soms nog beter dan de "werkelijkheid". De luidheid van een telefoongesprek is tot wel 20 decibel luider dan een natuurlijk gesprek, op het gevaarlijke af met als voordeel dat slechthorenden nog kunnen converseren over een telefoonverbinding als ze daar in een "live" situatie al niet meer toe in staat zijn. De natuurlijkheid van een stem over een telefoonverbinding is prima omdat we de stem vlakbij opnemen en dus geen last hebben van reflecties in de opnamekamer en het timbre is prima omdat een groter spectrum opgenomen kan worden met de introductie van HD voice, van 300 tot 3400 Hz bij een ouderwetse telefoonverbinding naar 50 tot 7000 Hz, en soms zelfs ruimer, voor een HD voice verbinding.

Kunnen we onze HiFi phone analyse uitbreiden naar muziek weergave? In bepaalde opzichten is HiFi muziekweergave eenvoudiger te realiseren dan een HiFi telefoon: we hoeven namelijk niet heen en weer. Door eerst wat pakketten binnen te halen, te bufferen, kunnen we netwerkvariaties opvangen en ziet een film op Netflix er perfect uit met het juiste High Quality abonnement. Waarom klinkt tegenwoordig bij veel mensen thuis de weergave dan zo belabberd? Dat is een simpel gevolg van een gebrek aan aandacht, onze aandacht gaat tegenwoordig naar data over 4G/5G/WiFi, bye bye HiFi.

Wat is het aspect dat we nodig hebben om te komen tot HiFi muziek weergave? Allereerst moeten we ons realiseren dat er twee types van HiFi bestaan: de illusie "hier en nu" (augmented reality) en de illusie "daar en toen" (virtual reality). Voor een stem kiezen we vrijwel altijd voor "hier en nu", we nemen hem zo direct mogelijk op, zonder reflecties (ook wel dood genoemd) en geven hem weer over een koptelefoon of luidspreker die goed koppelt aan onze oren. Voor muziek is dat een verkeerde, en vaak onuitvoerbare, benadering, probeer maar eens alle instrumenten van het Concertgebouw Orkest droog op te nemen en dan weer te geven over een koptelefoon of luidspreker, dat klinkt belabberd. Er zijn twee oorzaken waarom het slecht klinkt, de eerste is dat een dode opname van een muziekinstrument van zijn afstraling in de ruimte een enkel punt in de ruimte opneemt, maar dat de klank op die plaats vaak niet optimaal is. Heel veel muziekinstrumenten hebben een grillig afstraalpatroon waardoor ze op ieder punt in de ruimte anders klinken en de beste uitkomst is om al die klanken uit te middelen, zoals dat in een concertzaal met een goede akoestiek gebeurd (zie Figuur 2). De tweede reden is dat we bij muziek, in tegenstelling tot spraak, graag worden omspoeld door het geluid, op dezelfde manier als dat in een concertzaal gebeurd. We weten al dat wanneer je iets opneemt op een plaats waar het goed klinkt dat die opname bij het afspelen vaak heel slecht klinkt. Mensen rekenen de akoestiek van de ruimte waarin ze luisteren weg. Pas als het je lukt om exact dezelfde signalen aan onze beide oren aan te bieden als die van de opname, dan lukt het opnieuw om bij de weergave de ruimte weg te rekenen.

Je zou denken dat voor een illusie “daar en toen” een koptelefoon de beste benadering is, maar het is niet eenvoudig om bij een weergave exact de twee oorsignalen van de live-beleving te reproduceren. Bij een koptelefoon treedt vrijwel altijd een “in je hoofd”-effect op. Alleen met de juiste individuele filtering van de twee oorsignalen, gecombineerd met een “head tracker” die de signalen aanpast aan de positie van je hoofd, kun je een redelijk goede “daar en toen” illusie creëren. Laag frequente signalen tot 100 Hz blijven moeilijk om weer te geven, omdat deze vaak via botgeleiding gaan in plaats van via het directe oorpad.

En hoe zit dat met luidsprekerweergave? Helaas, ondanks de introductie van quadrafonie in de jaren zeventig, de ontwikkeling van surround sound in de afgelopen dertig jaar, object-based audio in de afgelopen twintig jaar en de vele standaarden hoe geluid op te nemen en te weergeven en diverse eigen ontwikkelingen van HiFi fabrikanten zoals Dolby Surround en ATMOS, is het niet gelukt om dat gevoel van omspoeling goed op te nemen en weer te geven. Dat komt omdat bij al deze systemen teveel nadruk wordt gelegd op lokalisatie, nodig bij een film, maar bij muziekweergave is het effect van omspoeling veel belangrijker. Sterker nog, de meeste mensen luisteren nog naar muziek over een gewone stereo opstelling met twee luidsprekers, en de simpelste en effectiefste uitbreiding is om met een eenvoudig algoritme wat diffuus veld toe te voegen met extra luidsprekers. Verder wordt de klank van een luidspreker voor een groot deel bepaald door de akoestiek van de weergavekamer en is het aan te bevelen om de ruimte droog te maken met demping op de grond en het plafond en het geluid te diffuseren met grote objecten waardoor het omspoelingseffect beter wordt.



Figuur 2. muziekweergave kwaliteit