Compared to many existing audibility-based models, the suitability of the ESTI for use in reverberant conditions is an important advantage. When compared to other modulation-based models, the use of short time windows makes the ESTI-model better suited for use in interrupted noise. However, several of these models tend to perform better when speech is masked by noises with speech-like characteristics. When compared to models based on machine learning, the simplicity and easy usage of the ESTI are the main advantages.
An important aspect that was largely left unaddressed in the current work is the applicability of the model for sensorineurally hearing-impaired persons. The use of the individual tone audiogram can be useful, but this does not account for the temporal and spectral distortions that usually occur with this type of hearing loss. An additional individual distortion factor based on the hearing loss for speech in noise might improve the model further.
The ESTI proved to be a valuable extension of the classic STI for the use in non-stationary noises. With only a single recording of the background noise and an impulse response measurement, a reliable ESTI-value can be obtained. This value can then be used to predict speech intelligibility in a variety of non-stationary background noises. The addition of context improved the model further, but at the cost of higher complexity. It did show how a context model can be successfully used in combination with traditional methods for estimating speech intelligibility.

Samenvatting

De Speech Transmission Index of STI is een veelgebruikte maat die ontworpen is voor de evaluatie van de kwaliteit van spraak tussen een spreker en een luisteraar. De STI kan worden toegepast in de communicatietechnologie en zaalakoestiek om een transmissiekanaal te evalueren zonder spraakverstaanbaarheidsmetingen te hoeven uitvoeren voor een specifieke conditie. Voorbeelden van het gebruik zijn het ontwerpen en de constructie van auditoria en theaters, en de evaluatie van een werkomgeving wanneer er problemen optreden bij het verstaan van spraak.
De STI is een relatief eenvoudige maar robuuste index tussen 0 en 1 die is gebaseerd op de vermindering van modulaties in de spraak als gevolg van achtergrondruis en/of nagalm. Deze modulatiereductie blijkt gecorreleerd met de mate van spraakverstaan. De basis van de STI is de Modulatie Transfer Functie. Deze MTF wordt berekend op basis van opnames die traditioneel werden gemaakt met behulp van de *directe* meetmethode, door een gemoduleerde ruis als meetsignaal te gebruiken. Een andere benadering is de *indirecte* meetmethode, waarbij een separate opname van de achtergrondruis wordt gemaakt in combinatie met een meting van de impulsrespons. Wanneer de MTF bekend is, wordt deze omgerekend naar een STI-waarde in verschillende stappen. Een STI-waarde van 0 staat voor een situatie waar spraakverstaanbaarheid onmogelijk is. Een index van 0.75 of hoger staat voor omstandigheden waarbij goede tot excellente spraakverstaanbaarheid mogelijk is. Een index onder de 0.30 duidt op slechte tot zeer slechte spraakverstaanbaarheid. Merk op dat dit enkel van toepassing is op normaalhorende luisteraars in de eigen moedertaal. Wanneer een transferfunctie tussen de STI en de spraakverstaanbaarheid bekend is voor een bepaald spraakcorpus, kan de spraakverstaanbaarheid geschat worden met behulp van de STI-waarde.
Ondanks de robuustheid en grondige evaluatie van de STI zijn er verschillende zwakke punten. Wanneer non-lineaire vervormingen zoals compressie of spectrale subtractie plaatsvinden, gaat de nauwkeurigheid van de STI achteruit. Echter, het huidige onderzoek richtte zich op zaalakoestiek, waardoor alleen vervormingen die lineair van aard zijn, werden meegenomen. Een andere beperking van de STI is de bruikbaarheid wanneer achtergrondruis niet stationair is. Dit aspect was de belangrijkste aanleiding voor het huidige onderzoek. De achtergrond van deze beperking is tweeledig. Ten eerste is de traditionele, directe meetmethode met gemoduleerde ruis als meetsignaal gevoelig voor fluctuaties in de achtergrondruis. Dit kan leiden tot onder- of overschatting van de STI. Ten tweede neemt de spraakverstaanbaarheid bij normaalhorende luisteraars toe wanneer er dips in de ruis worden

geïntroduceerd. Deze winst wordt veroorzaakt door het vermogen om spraakfragmenten waar te nemen op de momenten dat de spraak het minst wordt beïnvloed door de ruis. De STI kan niet omgaan met de winst in het verstaan door deze fluctuaties in de ruis, waardoor de uitkomst niet representatief is voor de werkelijke omstandigheden. Het doel van het huidige werk was om de toepasbaarheid van de STI in niet-stationaire achtergrondruizen te verbeteren door zowel de meetmethode als de winst door fluctuaties in de ruis te onderzoeken.

In **hoofdstuk 2** werden de condities onderzocht waaronder de STI nauwkeurig gemeten kan worden. Hierbij lag de focus op de indirecte meetmethode, waarbij de MTF werd afgeleid van een meting van de impulsrespons en een langdurige opname van de achtergrondruis. Deze methode is minder gevoelig voor fluctuaties in de ruis, maar onder welke omstandigheden de metingen betrouwbaar uitgevoerd kunnen worden, was niet bekend. Om dit verder te onderzoeken werden twee experimenten uitgevoerd. Metingen van de impulsrespons (met een zogenaamd *sweep-signaal*) en ruisopnames werden uitgevoerd in een ruimte met variabele absorptie, verschillende niveaus van stationaire en fluctuerende ruis, en verschillende niveaus van het sweep-signaal. Om de experimentele bevindingen te kunnen extrapoleren naar andere akoestische condities, werd een groot aantal andere omstandigheden gesimuleerd. De experimenten en simulaties toonden aan dat de minimale impuls-ruisverhouding van +25 dB (overeenkomend met een sweep-ruisverhouding van –4 tot +15 dB) in niet-stationaire ruis benodigd was om de STI nauwkeurig te kunnen meten.

De Extended STI of ESTI werd geïntroduceerd in **hoofdstuk 3**. Het doel van het aangepaste model was om rekening te houden met de verbeterde spraakverstaanbaarheid wanneer er dips in de ruis aanwezig zijn. De belangrijkste aanpassing was de berekening van de STI per tijdsinterval van 2 ms, in plaats van voor het volledige signaal. De uiteindelijke ESTI-waarde was het gemiddelde van alle lokale STI-waarden. Om rekening te houden met abrupte fluctuaties in de ruis werd ook voorwaartse maskering (*forward masking*) toegevoegd aan het model. De fijnafstelling van de modelparameters werd gedaan op basis van nieuw uitgevoerde spraakverstaanbaarheidsmetingen bij normaalhorenden waarbij zinnen werden aangeboden. Door middel van een adaptieve procedure werd de signaal-ruisverhouding gemeten waarbij 50% van de zinnen werden verstaan (cSNR). De spraak werd vervormd door verschillende ruizen te gebruiken (stationaire ruis en twee soorten niet-stationaire ruis), dan wel nagalm toe te voegen in vijf gradaties. Evaluatie van het model vond plaats met behulp van data van 10 studies uit de bestaande literatuur over spraakverstaanbaarheid. De ESTI voorspelde de spraakverstaanbaarheid beter dan de klassieke

STI in alle soorten niet-stationaire ruis die werden gebruikt. Onnauwkeurigheden in de voorspellingen werden geobserveerd wanneer achtergrondruis spraakachtige eigenschappen vertoonde. Wanneer enkel de omhullende karakteristieken van spraak vertoonde, was het verschil tussen de waargenomen en voorspelde cSNR ongeveer 3 – 4 dB. Wanneer de fijnstructuur van de ruis ook op spraak leek, werd er een extra verschil van 5 – 7 dB gevonden. De hypothese van de auteurs was dat deze onnauwkeurigheden het resultaat waren van modulatiemaskering, *informational masking* en/of contexteffecten.

In **hoofdstuk 4** werd de hypothese getest dat deze onnauwkeurigheden veroorzaakt werden door contexteffecten. Bij het waarnemen van spraakfragmenten in niet-stationaire ruis met hoge modulatiefrequenties heeft de luisteraar toegang tot delen van alle spraakelementen. Deze elementen kunnen fonemen als deel van een woord zijn, maar ook woorden als deel van een zin. Wanneer de modulaties in de ruis traag zijn, neemt de waarschijnlijkheid toe dat elementen in de spraak volledig gemaskeerd worden door de langere ruisfragmenten en daardoor niet verstaanbaar zijn. De luisteraar kan in dat geval terugvallen op contextuele informatie om het gemiste spraakelement te "raden". Om dit mechanisme bij het verstaan mee te nemen, werd context toegevoegd aan het ESTI-model. Eerst werd de ESTI per spraakelement berekend in plaats van voor het gehele signaal. Daarna werd een transferfunctie geschat om de ESTI per element te koppelen aan de elementscore in isolatie. Wanneer deze score bekend was, kon de verstaanbaarheid van de gehele uiting geschat worden met behulp van een context model. Om de prestatie te evalueren van dit op context gebaseerde ESTI-model (cESTI-model) werd bestaande spraakverstaanbaarheidsdata van betekenisvolle, monosyllabische woorden in onderbroken ruis geanalyseerd. De toevoeging van twee verschillende contextmodellen werd vergeleken, aangeduid als $cESTI_1$ (met het Bronkhorst contextmodel) en $cESTI_2$ (met het Boothroyd en Nittrouer contextmodel). De nauwkeurigheid van de voorspellingen van het nieuwe model verbeterde aanzienlijk voor interruptiefrequenties lager dan 5 Hz. De prestaties van de twee modelversies $cESTI_1$ en $cESTI_2$ waren vergelijkbaar.

Het cESTI-model werd geëvalueerd in **hoofdstuk 5** met behulp van nieuw gemeten CVC-woorden in stationaire en onderbroken ruis. Zowel nonsens als betekenisvolle woorden werden gebruikt. Alleen $cESTI_2$ (met het eenvoudigere van de twee contextmodellen uit hoofdstuk 4) werd gebruikt in dit hoofdstuk. Het model presteerde beter bij zowel nonsens als betekenisvolle woorden. Echter, ondanks de verbeterde prestaties werd de verstaanbaarheid van betekenisvolle woorden bij interruptiefrequenties onder de 5 Hz nog steeds onderschat. Hogere contextwaarden bleken mogelijk meer geschikt te zijn bij lage interruptiefrequenties en leidden tot een verbeterde nauwkeurigheid van

de voorspellingen. Verder lieten modelvoorspellingen een duidelijke afname zien bij interruptiefrequenties van 8 en 16 Hz, mogelijk gerelateerd aan een overschatting van het effect van voorwaartse maskering. De nauwkeurigheid van het model nam toe bij het gebruik van een alternatieve functie voor de voorwaartse maskering.

In de hoofdstukken 4 en 5 werd het cESTI-model uitsluitend geëvalueerd met behulp van monosyllabische *woorden*. Echter, de oorspronkelijke reden voor het toevoegen van context was de onnauwkeurigheid van de voorspellingen van *zinnen* die gemaskeerd werden door ruizen met spraakachtige kenmerken in hoofdstuk 3. Om deze reden werd in **hoofdstuk 6** de prestatie van het cESTI-model onderzocht op basis van de spraakverstaanbaarheidsdata van zinnen uit hoofdstuk 3. Zowel de $cESTI_1$ als de $cESTI_2$ werden opnieuw geëvalueerd. De nauwkeurigheid van de voorspellingen van $cESTI_1$ bleef vergelijkbaar of nam toe in vergelijking met de ESTI-voorspellingen. Daarentegen nam de nauwkeurigheid van het $cESTI_2$-model juist af voor niet-stationaire achtergrondruizen zonder spraakachtige kenmerken. Echter, $cESTI_2$ presteerde beter dan $cESTI_1$ bij spraakachtige ruizen, zoals bijvoorbeeld een andere spreker. In het algemeen bleven de voorspellingen met betrekking tot spraakachtige ruizen relatief onnauwkeurig. Het aanpassen van de contextwaarden en de transferfuncties kunnen mogelijk tot een verbetering leiden, maar het is waarschijnlijk dat modulatiemaskering en informational masking een belangrijke rol blijven spelen in de discrepantie tussen spraakverstaanbaarheid en de modelvoorspellingen.

In vergelijking met veel modellen gebaseerd op hoorbaarheid, is de toepasbaarheid van de ESTI in condities met nagalm een belangrijk voordeel. Vergeleken met andere modellen gebaseerd op spraakmodulaties, maakt het gebruik van korte tijdsintervallen de ESTI beter geschikt voor onderbroken ruis. Echter, een aantal van deze modellen lijkt beter te presteren wanneer spraak gemaskeerd wordt door ruizen met spraakachtige kenmerken. Wanneer de ESTI vergeleken wordt met modellen gebaseerd op machine learning, zijn de eenvoud en het gebruiksgemak van de ESTI de belangrijkste voordelen.

Een belangrijk aspect dat nauwelijks genoemd werd in het huidige werk is de toepasbaarheid van het model bij mensen met een perceptief gehoorverlies. Het gebruik van het individuele toonaudiogram zou bruikbaar kunnen zijn voor deze groep, maar deze benadering houdt geen rekening met de temporele en spectrale vervorming die normaliter optreedt bij dit type gehoorverlies. Het toevoegen van een vervormingsfactor op basis van het gehoorverlies voor spraak in ruis kan het model mogelijk verder verbeteren.

De ESTI is een waardevolle toevoeging op de klassieke STI voor het gebruik in niet-stationaire ruizen. Met een enkele opname van de achtergrondruis en een

impulsresponsmeting kan een betrouwbare ESTI worden berekend. Dit resultaat kan vervolgens worden ingezet om de spraakverstaanbaarheid te voorspellen in een verscheidenheid aan niet-stationaire achtergrondruizen. Het toevoegen van context verbeterde het model verder, maar wel ten koste van een hogere complexiteit. Het liet wel zien hoe een contextmodel succesvol gebruikt kan worden in combinatie met traditionele methodes voor het voorspellen van spraakverstaan.