Microsoft Speech containers

On-premise speech API's	
Systeemvereisten:	Docker Engine
Ontwikkeld door:	Microsoft
Pay-per-use	Contactpersoon:	Bert.Vanhalst@smals.be

Functionaliteiten

In deze Quick Review bespreken we de Microsoft Cognitive Services Speech containers. Het betreft Speech service API's voor het omzetten van spraak naar tekst (*speech-to-text*) en tekst naar spraak (*text-to-speech*). Standaard wordt die functionaliteit aangeboden als een dienst in de Azure cloud. Microsoft biedt met deze containers de mogelijkheid om die diensten ook op andere locaties te draaien, zoals in het eigen datacenter. Mogelijke beweegredenen om dat te doen zijn vereisten rond data governance of compliance (de gegevens mogen niet in handen komen van een derde partij), of vereisten voor een snelle verwerking door ervoor te zorgen dat de speech services fysiek dicht bij de data en business logica draaien.

De *speech-to-text* container biedt, net zoals de public cloud versie, de mogelijkheid om spraak in real time of in batch om te zetten naar tekst. Met de *text-to-speech* container kan tekst dan weer omgezet worden naar spraak. Beide containers bieden ondersteuning voor meerdere talen, waaronder Engels, Frans, Nederlands en Duits. De spraakfuncties kunnen toegevoegd worden aan een eigen toepassing op basis van REST API's of de Speech SDK die beschikbaar is voor verschillende programmeertalen en platformen.

Om de accuraatheid te verbeteren biedt Microsoft de mogelijkheid om het basismodel voor *speech-to-text* te customiseren door het bij te trainen. Dat bijtrainen kan op basis van audiobestanden met de bijhorende transcriptie, of tekstbestanden met voorbeeldzinnen of phonetische notatie van woorden en afkortingen. Daarnaast kan ook de *text-to-speech* component gecustomiseerd worden. Op basis van audiobestanden kan je een aangepaste stem bouwen die meer aansluit bij het eigen merk of product.

Conclusies & Aanbevelingen

De Speech containers doen wat ze beloven, zijn gemakkelijk te deployen als Docker container en gemakkelijk te gebruiken. Dit laat toe om gebruik te maken van spraakfunctionaliteit zonder de spraakgegevens te moeten delen met een derde partij. De kost is identiek aan de overeenkomstige public cloud diensten. Voor het aanrekenen van het verbruik is evenwel nog een connectie nodig met de Azure servers, waardoor we niet kunnen spreken van een volledig offline scenario. Customisatie kan de accuraatheid nog verhogen. Op dit moment zijn de mogelijkheden in het Nederlands hiervoor beperkt.

Testen & Resultaten

Om aan de slag te kunnen met de speech containers is er een Azure subscription nodig en moet er (voorlopig althans) een online aanvraag ingediend worden bij Microsoft. We testten de Nederlandstalige versie van de *speech-to-text* en *text-to-speech* containers uit. De installatie van de containers kan eenvoudig op basis van de images die beschikbaar zijn in Docker Hub. Om de container te starten moeten er parameters meegegeven worden voor het aanrekenen van het verbruik: het *billing endpoint* waar de gegevens over het verbruik naartoe gestuurd worden en een API key die gelinkt is aan de Azure account.

Eens de containers draaien kunnen we een client toepassing voorzien die de speech API's van de containers aanroept. Dit zijn dezelfde API interfaces als bij de public cloud diensten, met een paar kleine nuances:

- Er zijn aparte containers per taal, terwijl er bij de public cloud versie één API is die de taal als parameter aanvaardt. Het is natuurlijk mogelijk om verschillende containers naast elkaar te draaien met elk een specifieke taal.
- De API's die de containers aanbieden zijn standaard niet beveiligd; die beveiliging moet je zelf nog voorzien.
- Als klant sta je zelf in voor de onderliggende infrastructuur en het deployen van updates van de containers.

Als test voorzien we een webtoepassing die gebruik maakt van de Javascript SDK om de speech container API's aan te roepen. In een eerste stap wordt Nederlandstalige spraak-input omgezet naar tekst via de *speech-to-text* API. Die tekst wordt dan in een tweede stap terug uitgesproken via de *text-to-speech* API. Waar we bij de *speech-to-text* container gebruik kunnen maken van een websocket connectie, kan dat niet bij de *text-to-speech* API. Daar kunnen we enkel gebruik maken van de REST API.

Op basis van de beperkte testen die zijn uitgevoerd, kunnen we stellen dat de kwaliteit van de *text-to-speech* vrij goed lijkt. Maar er sluipen nog fouten in de *speech-to-text*. Het betreft vooral specifieke termen en afkortingen eigen aan het business domein waarin we werken. Zo wordt "RSZ" omgezet naar "ingezet" of "recent". Maar soms loopt de omzetting ook mis bij woorden waar je dit niet zou verwachten. Zo wordt "België" soms omgezet naar "belgi", of is er verwarring tussen "dit" en "het".

Microsoft biedt verschillende mogelijkheden om de speech services te customiseren, maar niet alle features worden ondersteund in elke taal. Bij het Nederlands zijn de mogelijkheden beperkt. Vooral de mogelijkheid om phonetische uitspraak toe te voegen als trainingsdata is een beperking. Hoewel we bij de testen een custom model wisten te trainen via de Speech Studio, zijn we er niet in geslaagd om de accuraatheid van de spraakherkenning significant te verbeteren.

Gebruiksvoorwaarden & Budget

De kosten voor het gebruik van de speech containers worden verrekend volgens een *pay-per-use* model. De tarieven zijn dezelfde als die voor de public cloud diensten. Het verbruik van beide (containers en public cloud diensten) wordt samengeteld en maandelijks afgerekend.