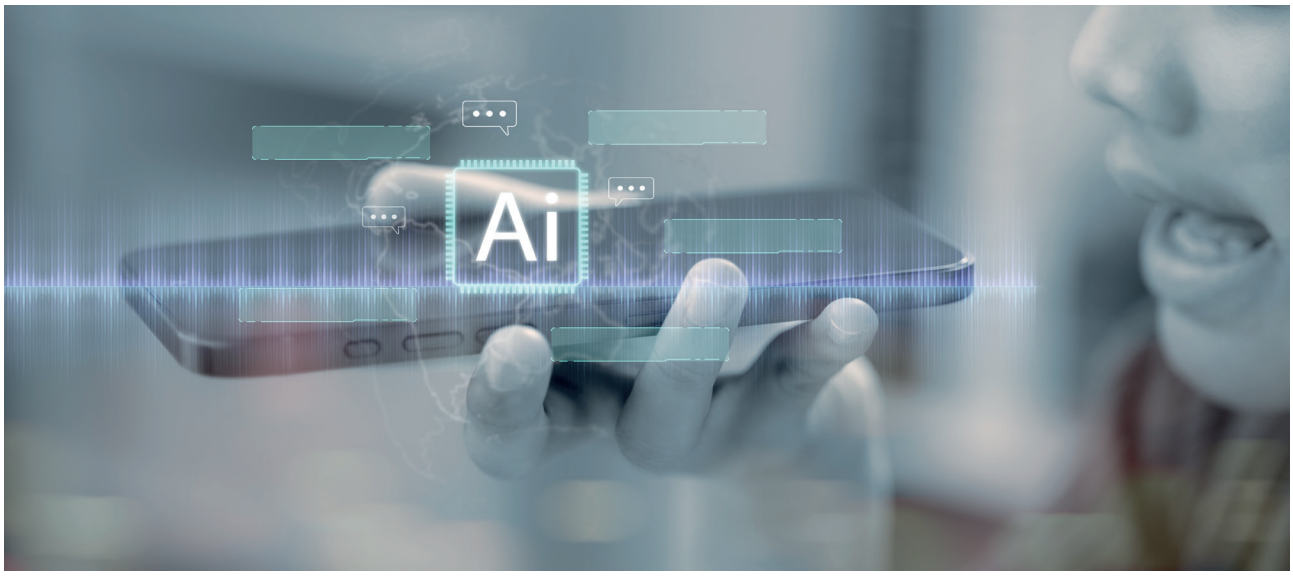
Jak sztuczna inteligencja uczyła się mówić



Technologia głosowa stała się w naszych czasach istotną platformą komunikacji między człowiekiem a sztuczną inteligencją. Współczesne systemy rozpoznawania i syntezy mowy nie tylko słuchają, rozumieją i odpowiadają, ale też są zdolne do rozpoznawania i naśladowania wielu różnych parametrów głosu – dzięki czemu posługują się sztucznie syntetyzowaną mową w sposób właściwy człowiekowi. Tworzy to niezwykłą, można wręcz powiedzieć intymną więź między człowiekiem a maszyną.

Jacek Grabowski

z wykształcenia specjalista gazownictwa i górnictwa naftowego, przygodę z informatyką rozpoczął w końcu lat 80. XX wieku od współpracy z wydawnictwem „Lupus", gdzie publikował teksty głównie w dwutygodniku „PCkurier" i miesięczniku „Enter". Współtwórca pierwszego w Polsce informatycznego czasopisma B2B „MRK" (1997). Były redaktor naczelny miesięcznika „Reset", współpracownik wielu innych tytułów (magazyn „WWW", „IT Reseller", „Komputer Świat"). Obecnie freelancer, współpracuje m.in. z warszawską komunikacją miejską.

Mowa jest jednym z najważniejszych narzędzi porozumiewania się ludzi. Głos to naturalny i intuicyjny sposób komunikacji rozwijający się od wczesnego dzieciństwa i towarzyszący nam przez całe życie. W rozmowie znaczenie mają nie tylko słowa, lecz także intonacja, tempo, pauzy czy akcent, które niosą dodatkowe informacje o naszych intencjach, stanie emocjonalnym i kontekście sytuacyjnym. Coś, co dla nas jest naturalne i normalne, dla maszyny jest jednak wyzwaniem: aby uzyskać realistyczne brzmienie i rozumienie wszystkich aspektów mowy ludzkiej, komputer musi wykonać szereg skomplikowanych operacji. Stąd synteza mowy była zawsze trudnym zadaniem, Jednak zanim w ogóle stała się możliwa, najpierw trzeba było przekształcić analogowy dźwięk w sygnał cyfrowy, możliwy do przetwarzania w komputerze.

Cyfrowe przetwarzanie dźwięku

Teoretyczne początki cyfrowego przetwarzania dźwięku sięgają pierwszej połowy XX w. Istotną rolę odegrała teoria próbkowania opracowana przez Harry'ego Nyquista (1928) i Claude'a Shannona (1949), zgodnie z którą sygnał analogowy może zostać wiernie odwzorowany w postaci cyfrowej, jeżeli częstotliwość próbkowania jest co najmniej dwukrotnie większa od najwyższej częstotliwości w sygnale źródłowym. Próbka dźwięku to jednorazowy pomiar wartości fali akustycznej, który jest zapisywany w formie cyfrowej podczas procesu próbkowania. Rolę częstotliwości próbkowania można porównać do tworzenia filmu z poszczególnych klatek obrazu: kiedy klatek jest zbyt mało, ruchomy obraz przeskakuje albo jest spowolniony. Dopiero odpowiednia liczba klatek zapewnia wrażenie ciągłego ruchu. Podobnie jest z cyfrową reprezentacją dźwięku – im większa liczba próbek, tym dokładniejsze jest cyfrowe odtworzenie analogowego sygnału. Poza częstotliwością próbkowania istotnym parametrem cyfrowego dźwięku jest jego rozdzielczość, czyli liczba bitów opisujących każdą próbkę.

Wraz z rozwojem cyfryzacji dźwięku ewoluowały także metody matematyczne umożliwiające analizę i przetwarzanie sygnałów. W latach 60. pojawiły się m.in. algorytmy szybkiej transformaty Fouriera (*Fast Fourier Transform*, FFT), opracowane przez Jamesa Cooleya i Johna Tukeya (1965), które ułatwiły praktyczne zastosowanie przetwarzania sygnału cyfrowego w obliczeniach komputerowych. Największą zaletą FFT była prędkość działania – umożliwiło to przetwarzanie cyfrowego sygnału dźwiękowego w czasie rzeczywistym i analizę dużych bloków danych.

W latach 70. pojawiły się pierwsze dedykowane układy przetwarzania sygnału (DSP) – specjalizowane procesory, zdolne do wykonywania operacji mnożenia i akumulacji z dużą prędkością. Jednym z pierwszych ich szerokich zastosowań konsumpcyjnych były odtwarzacze płyt CD wprowadzone w 1982 r. Standard CD obejmował zapis i odtwarzanie dźwięku z częstotliwością próbkowania 44,1 KHz i rozdzielczością 16 bitów. W kolejnych latach rozwój internetu i technologii cyfrowej zaowocował pojawieniem się algorytmów stratnej kompresji dźwięku, które umożliwiły znaczną redukcję rozmiaru plików audio przy zachowaniu zadowalającej jakości odsłuchu. Było to istotne m. in. ze względu na powolną transmisję danych sieciowych – zastosowanie kompresji ułatwiało przesyłanie plików dźwiękowych. Najpopularniejszym stratnym formatem służącym do rozpowszechniania plików dźwiękowych w sieci stał się MP3, opracowany w Instytucie Fraunhofera. Rozwój komputerów osobistych i oprogramowania DAW (Digital Audio Workstation) wprowadził przetwarzanie cyfrowego dźwięku „pod strzechy" – również amatorzy dostali do ręki wyrafinowane narzędzia umożliwiające edycję dźwięku oraz remiksowanie i komponowanie własnych utworów.

Współcześnie szczególnie dynamicznie rozwijają się metody wykorzystania sztucznej inteligencji w przetwarzaniu dźwięku cyfrowego. Algorytmy uczenia maszynowego umożliwiają automatyczną separację źródeł dźwięku, inteligentną korekcję barwy, a nawet generowanie realistycznych kompozycji muzycznych. Współczesne badania koncentrują się na połączeniu klasycznych metod DSP z głębokimi sieciami neuronowymi, tworząc nową gałąź – przetwarzanie dźwięku za pomocą sztucznej inteligencji. Technika ta ma istotne znaczenie również w zastosowaniu głosu do komunikacji z SI.

Choć komputerowe rozpoznawanie mowy istniało już od lat 50. XX w., to początkowo technologia ta w ogóle nie miała powiązania ze sztuczną inteligencją, Pierwsze programy uznawane za początki sztucznej inteligencji komunikującej się z człowiekiem mogły dialogować tylko za pomocą tekstu. Przełomem było opracowanie i wdrożenie do rozpoznawania mowy ukrytych modeli Markowa (HMM), co nastąpiło na szerszą skalę w dekadzie lat 90. XX w., a także opracowanie prostych reguł rozumienia języka przez sztuczną inteligencję. Umożliwiło to powstanie pierwszych systemów IVR (Interactive Voice Response) rozpoznających niewielką liczbę poleceń głosowych. Nie przypominało to jeszcze swobodnego dialogu, ale systemy informatyczne stały się zdolne do zanalizowania poleceń głosowych i wykonywania na tej podstawie zdefiniowanych zadań.

Na przełomie XX i XXI w. pojawiły się bardziej zaawansowane rozwiązania łączące technologię rozpoznawania i analizy mowy ze sztuczną inteligencją. Istotnym krokiem było wprowadzenie statystycznych metod badania języka naturalnego (*NLP – Natural Language Processing*, warto zwrócić uwagę, że skrót NLP odnosi się często także do tzw. programowania neurolingwistycznego, które jest osobną dziedziną). Mówiąc ogólnie, metoda NLP polega na rozbijaniu języka na krótsze, bardziej podstawowe elementy, podejmowaniu prób zrozumienia zależności pomiędzy poszczególnymi częściami składowymi oraz badaniu sposobu, w jaki łączą się one ze sobą, tworząc sens. Dzięki rosnącym korpusom językowym i zbiorom nagrań do trenowania SI, rozpoznawanie mowy zaczęło współpracować z modułami rozumienia języka NLP. W ten sposób pojawiła sie pierwsza generacja systemów, które były zdolne nie tylko do rozpoznawania słów, lecz również zrozumienia mowy ludzkiej. Jednak w tym pierwszym okresie komunikacja człowieka z maszyną za pomoca mowy była wciąż jeszcze bardzo ograniczona i „sztywna".

Głos i sztuczna inteligencja

Głos ludzki to zjawisko akustyczne powstające w wyniku przepływu powietrza z płuc przez krtań, gdzie drgają fałdy głosowe, a następnie dźwięk jest modyfikowany w rezonatorach (jama ustna, nosowa, gardłowa) oraz artykułowany przez język, wargi i żuchwę. Fizyczne cechy głosu obejmują m.in. wysokość dźwięku, głośność i barwę (tembr). Tembr głosu jest cechą unikalną dla każdego człowieka i zależy od budowy jego narządów głosowych. Wysokość głosu i sposób artykulacji zależy od cech fizycznych, właściwych np. dla płci (głos kobiet jest ogólnie wyższy niż mężczyzn), zmienia się także wraz z wiekiem człowieka – np. na starość głos staje się słabszy, mniej stabilny i wyraźny. Na jakość mowy mogą wpływać choroby krtani, gardła i jamy ustnej, a także inne czynniki. Trudno nawet w skrócie opisać wszystkie aspekty mowy i głosu, które wpływają na jego brzmienie, intonację, artykulację itd. Nic więc dziwnego, że właściwe odwzorowanie głosu ludzkiego w interakcjach z komputerami ma długą i skomplikowaną historię.

Siri – pierwszy interaktywny asystent głosowy

Rewolucją w komunikowaniu się człowieka z maszyną był asystent głosowy Apple Siri, wprowadzony do smartfonów Apple w 2011 r. Apple przejęło Siri w 2010 r. Firma Siri rozwijała już wcześniej technologie rozpoznawania mowy i NLP w połączeniu z uczeniem maszynowym i uczeniem głębokim, a także syntezą TTS (*Text-To-Speech*). Ze względu na konieczność korzystania z wielkich zbiorów danych i odpowiednio silnych komputerów do ich przetwarzania, Siri działała w chmurze, wymagając stałego połączenia z internetem.

W tamtych czasach systemy głosowe opierały się na dwóch rodzajach syntezy, Pierwszą, najstarszą i jednocześnie najbardziej rozpowszechnioną w technologii asystentów głosowych, była tzw. synteza konkatenacyjna, polegająca na cięciu nagrań na fonetyczne dźwięki i łączeniu ich (konkatenacja) w nowe słowa i zdania. W metodzie tej system korzysta z dużej bazy nagrań lektora. Każde słowo, sylaba, fonem lub difon (przejście między dwoma fonemami) jest wcześniej nagrane i opisane. Wprowadzany tekst jest dzielony na jednostki fonetyczne, a system wyszukuje w bazie najlepiej pasujące fragmenty dźwięku i łączy je płynnie, tworząc kompletną wypowiedź. Wadami konkatenacji są mała skalowalność (trudna adaptacja do nowych języków i głosów), konieczność tworzenia olbrzymich baz z nagraniami, niewielkie możliwości oddania emocji i problemy z nietypowymi zwrotami (np. skrótami czy nazwami własnymi). Zaletą jest naturalnie brzmiący dźwięk i wysoka zrozumiałość syntetycznej mowy.

Drugim sposobem syntetyzowania mowy był tzw. wokoder, czyli procesor tworzący mowę parametrycznie, bez wykorzystania wcześniejszych nagrań jak w konkatenacji, lecz poprzez łączenie źródła z odpowiednimi filtrami sztucznie modelującymi kanał głosowy. W klasycznych wokoderach wykorzystywano technikę liniowego kodowania predykcyjnego (*LPC – Linear Predictive Coding*). Metoda ta wykorzystuje fakt, że ludzki głos ma dużą redundancję i polega na przewidywaniu kolejnych próbek sygnału na podstawie poprzednich, co pozwala na jego efektywne zakodowanie i odtworzenie. Tak zsyntetyzowany sygnał mowy charakteryzował się często sztucznym, „metalicznym" brzmieniem, jednak dzięki temu, że wokoder generuje wszystko sztucznie był znacznie bardziej elastyczny od metody konkatenacyjnej i – co ważne – nie wymagał żadnych zewnętrznych baz nagrań. Synteza za pomocą wokodera byla więc wykorzystywana najczęściej w systemach, dla których priorytetem była elastyczność, skalowalność i rozmiar danych. Tam, gdzie priorytetem była zrozumiałość i naturalność wypowiedzi (np. w asystentach głosowych) stosowano metodę konkatenacyjną.

Wprowadzenie głębokiego uczenia i modeli językowych

W połowie drugiej dekady XXI w. pojawiły się pierwsze systemy rozpoznawania mowy bazujące na głębokich i konwolucyjnych sieciach neuronowych (DNN i CNN) oraz modelach pamięci długotrwałej krótkoterminowej (*LTSM – Long Short Term Memory*). Jednym z pierwszych był WaveNet Google – model generatywny wytrenowany na próbkach mowy ludzkiej. Tworzy on przebiegi wzorców mowy, przewidując, które dźwięki najprawdopodobniej będą następować po sobie, budowane po jednej próbce na raz, z prędkością do 24 tys. próbek dźwięku na sekundę. W 2017 r. Google opracował także pierwsze transformery wykorzystujące tzw. mechanizm uwagi (*selfattention*), który symuluje działanie ludzkiej uwagi poprzez przypisywanie różnych poziomów ważności różnym słowom w zdaniu. Zastosowanie tych rozwiązań spowodowało, że SI już nie tylko reagowała na polecenia głosowe, lecz zaczęła prowadzić konwersację z człowiekiem – rozumieć kontekst, zapamiętywać wcześniejsze wypowiedzi i adaptować odpowiedzi do użytkownika. Mogła łączyć ze sobą dane z różnych źródeł, generując wypowiedzi uwzględniające wiele naturalnych czynników. Dzięki temu rozmowa ze sztuczną inteligencją stała się wielomodalnym dialogiem przypominającym do złudzenia normalne porozumiewanie się ludzi.

Współczesna neuronowa synteza głosu dzieli się na dwie główne techniki – starszą dwustopniową, w której mamy system przetwarzania tekstu na mowę (TTS) i wokoder, oraz najnowszą end-to-end, czyli przetwarzanie jednostopniowe, w którym nie stosuje się już osobnego wokodera. W metodzie dwustopniowej system przetwarzania TTS tworzy tzw. mel-spektrogram, czyli wizualizację dźwięku używającą skali Mel na osi Y zamiast standardo-

wej skali liniowej częstotliwości w Hz. Skala melowa jest nieliniowa i lepiej odpowiada ludzkiemu słuchowi, ponieważ naśladuje sposób, w jaki człowiek odbiera wysokość dźwięku. Model przewiduje, jak powinna brzmieć wypowiedź, jakich użyć akcentów i pauz itd., a następnie wokoder na podstawie tych danych syntetyzuje mowę. Metoda ta daje na wyjściu naturalne brzmienie, umożliwia klonowanie głosu (czyli syntetyzowanie głosu konkretnej osoby z jego wszystkimi cechami), uwzględnia także dobre odwzorowanie emocji. Połączenie TTS z wokoderem daje jednak czasem niepożądane artefakty, poza tym wymaga większych zasobów niż wcześniejsze metody.

W metodzie jednostopniowej zastosowany jest tylko jeden duży model, który od razu przetwarza tekst do sygnału mowy bez współudziału modelu wokodera. Taka metoda używana jest np. przez model VALL-E Microsoftu czy ChatGPT. Metoda jednostopniowa jest szybsza, zapewnia lepszą spójność mowy, praktycznie unika powstawania artefaktów, umożliwia klonowanie głosu z kilku sekund nagrania, a także łatwe sterowanie stylem mówienia i emocjami zawartymi w głosie. Wymaga jednak bardzo wysokiej mocy obliczeniowej do obsługi wielkiego modelu.

Co dalej?

Sposób generowania mowy i jej rozpoznawania przeszedł więc długą drogę aż do czasów, kiedy w technologii głosowej używa się powszechnie sieci neuronowych i modeli językowych. Ich zastosowanie umożliwiło osiągnięcie naturalności „wypowiedzi" komputerowej i zwiększyło możliwości przetwarzania sygnału mowy. Obecnie jesteśmy w stanie skopiować głos i styl wypowiedzi konkretnej osoby w taki sposób, że złudzenie odbiorcy może być praktycznie pełne, co ułatwia tworzenie deep-fake'ów. Komputery „mówią" płynnie, bez zacięć i nienaturalnych dźwięków, ale ma to też niestety swoje wady. Warto bowiem pamiętać, że sztuczna inteligencja nie jest świadoma tego, że mówi – to jest tylko funkcja realizowana przez odpowiednie oprogramowanie i sprzęt, jednak odbiorca może rozumieć ten fakt zupełnie opacznie. Swobodny dialog z robotem, który jest w stanie symulować emocje i kształtować swoje wypowiedzi identycznie jak człowiek, sprawia wrażenie rozmowy z istotą faktycznie rozumiejącą i czującą podobnie, co prowadzi czasem do fatalnych w skutkach nieporozumień.

Etyczne problemy, na jakie napotyka synteza mowy, to tylko jeden z wielu kłopotów. Rozpoznawanie mowy osiągnęło pożądaną precyzję w laboratoriach i kontrolowanych warunkach, ale w realnym świecie bywa znacznie trudniejsze. Hałas ulicy, dźwięki w tle, echo w pomieszczeniach, a nawet różnice w akustyce pomieszczenia mogą znacząco obniżyć skuteczność systemu. Wyzwanie to staje się szczególnie widoczne w zastosowaniach medycznych, motoryzacji czy w przemyśle, gdzie pomyłki mogą prowadzić do poważnych konsekwencji.

Systemy głosowe w praktyce napotykają także trudności w rozpoznawaniu mniej rozpowszechnionych języków, regionalnych dialektów czy indywidualnych akcentów. W dodatku rozpoznanie słów to jedno, a zrozumienie sensu – drugie. AI musi nie tylko przetworzyć ciąg fonemów, ale też uchwycić kontekst, intencję i ewentualne niuanse emocjonalne. Choć obecnie udaje się to dość dobrze, niedokładności w analizie kontekstu i rozpoznaniu języka mogą prowadzić do błędnych odpowiedzi lub innych niepożądanych efektów działań systemu, co jest szczególnie krytyczne w zastosowaniach medycznych lub finansowych.

Sztuczna inteligencja musi być również wystarczająco elastyczna, aby adaptować się do zmian ludzkiego głosu w zależności od wieku, emocji, zdrowia czy warunków otoczenia. To wymaga ciągłego uczenia maszynowego i aktualizacji modeli, co łączy się z dużym nakładem mocy obliczeniowej i wiążącymi się z tym kosztami. Optymalizacja tych parametrów jest kluczowa, aby systemy głosowe mogły działać płynnie i responsywnie w codziennym użyciu. Tak więc mimo, że osiągnęliśmy już bardzo wiele w połączeniu sztucznej inteligencji z naturalnie brzmiącymi wypowiedziami głosowymi w różnych językach, to nadal technologie te wymagają dalszych badań i modyfikacji, żeby mogły się rozpowszechnić w wielu zastosowaniach.