(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 420 559**

(51) Int. CI.:

H04M 3/533 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **12.02.2007 E 07712713 (2)**

(97) Fecha y número de publicación de la concesión europea: **24.04.2013 EP 1992154**

(54) Título: **Un sistema a gran escala, independiente del usuario e independiente del dispositivo de conversión del mensaje vocal a texto**

(30) Prioridad:

10.02.2006 GB 0602682
09.01.2007 GB 0700376
09.01.2007 GB 0700377

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
23.08.2013

(73) Titular/es:

SPINVOX LIMITED (100.0%)
Wethered House Pound Lane
Marlow, Buckinghamshire SL7 2AF , GB

(72) Inventor/es:

DOULTON, DANIEL MICHAEL

(74) Agente/Representante:

LEHMANN NOVO, María Isabel

DESCRIPCIÓN

Un sistema a gran escala, independiente del usuario e independiente del dispositivo de conversión del mensaje vocal a texto.

ANTECEDENTES DE LA INVENCIÓN

1. Campo de la invención

La invención se refiere a un sistema, a gran escala, de mensajería vocal, independiente del usuario e independiente del dispositivo, que convierte mensajes vocales no estructurados en texto para su presentación visual en una pantalla. Conviene señalar inicialmente los retos operativos con los que se enfrenta un sistema de mensajería vocal, independiente del usuario, a gran escala que puede convertir mensajes vocales no estructurados en texto. En primer lugar, 'a gran escala' significa que el sistema debe ser susceptible de su escalamiento a muy grandes números, a modo de ejemplo, 500,000 o más abonados (normalmente se trata de abonados a un operador de telefonía móvil) y no obstante, permitir tiempos de procesamiento efectivo y rápido, siendo un mensaje, en general, solamente de utilidad si se recibe dentro de 2 a 5 minutos desde que se deja. Ésta es una exigencia bastante más estricta que la mayor parte de las puestas en práctica del sistema de reconocimiento ASR. En segundo lugar, la expresión 'independiente del usuario' significa que no hay absolutamente ninguna necesidad para un usuario para capacitar al sistema para reconocer sus modelos de expresión o voz (a diferencia de los sistemas de dictado de voz convencionales). En tercer lugar, la expresión 'independiente del dispositivo' significa que el sistema no está obligado a recibir entradas desde un dispositivo de entrada particular; algunos sistemas de la técnica anterior requieren la entrada desde, a modo de ejemplo, un teléfono de tono táctil. En cuarto lugar, el término 'no estructurado' significa que los mensajes no tienen ninguna estructura predefinida, a diferencia de la respuesta a las solicitudes vocales. En quinto lugar, la expresión 'mensajes vocales' se refiere a un campo de aplicaciones muy específico y bastante estrecho que plantea diferentes retos a quienes tienen que enfrentarse a numerosos sistemas de reconocimiento de voz automatizado (ASR) convencionales. A modo de ejemplo, los mensajes de correo de voz, para un teléfono móvil, suelen incluir vacilaciones, 'ers' y 'ums'. Un método de ASR convencional tendría que convertir fielmente todas las expresiones orales, incluso sonidos sin significado. El conjunto de la transcripción verbal o prolija caracteriza el método de la mayoría de los participantes en el campo del reconocimiento de voz automático ASR. No obstante, en realidad, no es adecuado, en absoluto, para el dominio de la mensajería vocal. En el dominio de la mensajería vocal, el reto creativo no es una transcripción exacta o prolija en absoluto, sino que, en cambio, captura el significado en la manera más útil para el destinatario previsto.

Solamente mediante un enfoque satisfactorio de estos cinco requisitos es posible tener una puesta en práctica correcta.

2. Descripción de la técnica anterior

La conversión desde voz a texto (STT) utiliza el reconocimiento de voz automático (ASR) y se ha aplicado, hasta ahora, principalmente a las tareas de dictado y de órdenes. El uso de la tecnología de ASR para convertir un correo de voz en texto es una nueva aplicación con varias características que son específicas de las tareas. Puede hacerse referencia al documento WO 2004/095821, que da a conocer un sistema de correo de voz, a través de Spinvox Limited, que permite que el correo de voz, para un teléfono móvil, se convierta en texto de SMS y se envíe al teléfono móvil. La gestión del correo de voz en la forma de texto es una opción atractiva. Suele ser más rápido leer que escuchar mensajes y, una vez en forma de texto, los mensajes de correo de voz se pueden memorizar y buscar tan fácilmente como un correo electrónico o un texto de mensajes cortos SMS. En una forma de realización, los abonados al servicio SpinVox desvían su correo de voz a un número de teléfono de SpinVox dedicado. Los abonados llamantes dejan mensajes de correo de voz como es habitual para el abonado. A continuación, SpinVox convierte los mensajes de voz a texto, con el objetivo de capturar el significado completo así como los elementos estilísticos e idiomáticos del mensaje, pero sin convertirlo necesariamente palabra por palabra. La conversión se realiza con un nivel importante de entrada por operadores humanos. El texto se envía luego al abonado como un texto de mensajes cortos SMS o correo electrónico. En consecuencia, los abonados pueden gestionar el correo de voz tan fácil y rápidamente como los mensajes de texto y de correo electrónico y pueden utilizar aplicaciones de clientes para integrar su correo de voz – ahora en forma de texto archivable y susceptible de búsqueda – con sus otros mensajes.

El problema con los sistemas de transcripción que están basados significativamente en operadores humanos, sin embargo, es que pueden ser costosos y difíciles de establecer a más escala – p.e., a una base de usuarios de 500,000 o más abonados. En consecuencia, no es factible para los principales operadores de telefonía móvil o celular ofrecerlos a su base de abonados porque, para los tiempos de respuesta rápida requeridos, es simplemente demasiado caro tener operadores humanos a la escucha y transcribiendo la integridad de cada mensaje; de este modo, el coste por mensaje transcrito sería prohibitivamente alto. Por lo tanto, el problema técnico fundamental es diseñar un sistema basado en tecnología de la información IT que permita al operador humano de transcripción actuar con la máxima eficacia.

En el documento WO 2004/095821 se consideraba algún grado de procesamiento de extremo frontal de ASR combinado con operadores humanos: esencialmente era un sistema híbrido; la presente invención desarrolla esta idea inventiva y define tareas concretas que el sistema de tecnología de la información IT pueda hacer, con lo que se aumenta, en gran medida, la eficacia del sistema completo.

Los sistemas híbridos son conocidos en otros contextos, pero el método convencional para la conversión de voz es eliminar por completo el elemento humano; éste es el reto operativo para los expertos en las técnicas de ASR, en particular, las técnicas de STT. Por lo tanto, consideraremos ahora algunos de los antecedentes técnicos para STT.

La tecnología básica de la conversión de voz a texto (STT) es la clasificación. La clasificación tiene como objetivo determinar a qué 'clase' pertenecen algunos datos dados. La estimación de probabilidad máxima (MLE), como numerosas herramientas estadísticas, utiliza un modelo subyacente del proceso de generación de datos – bien sea la acción de tirar una moneda o el sistema de generación de la voz humana. Los parámetros del modelo subyacente se estiman con el fin de hacer máxima la probabilidad de que el modelo generara los datos. Las decisiones sobre la clasificación se realizan, a continuación, comparando las características obtenidas a partir de los datos de la prueba con los parámetros del modelo obtenidos de los datos de capacitación para cada clase. Los datos de pruebas se clasifican luego como pertenecientes a la clase con la mejor coincidencia. La función de la probabilidad describe cómo la probabilidad de observar los datos varía con los parámetros del modelo. La máxima probabilidad puede encontrarse desde los puntos de inversión en la función de probabilidad si la función y sus derivadas están disponibles o pueden estimarse. Métodos para la estimación de la probabilidad máxima incluyen el descenso del gradiente simple así como los métodos de Gauss-Newton más rápidos. Sin embargo, si la función de probabilidad y sus derivadas no están disponibles se pueden utilizar algoritmos basados en los principios de Expectativa – Maximización (EM) que, comenzando desde una estimación inicial, convergen en un máximo local de la función de probabilidad de los datos observados.

En el caso de STT, se utiliza una clasificación supervisada en donde las clases se definen por los datos de capacitación más comúnmente como unidades trifónicas, lo que significa un fonema particular hablado en el contexto del fonema precedente y siguiente. (La clasificación no supervisada, en donde las clases se deducen por el clasificador, se puede considerar como un agrupamiento de los datos). La clasificación en STT se requiere no solamente para determinar a qué clase trifónica pertenece cada sonido en la señal de voz sino que, es muy importante conocer qué secuencia de trifonos es la más probable. Esto último se suele conseguir creando modelos de voz con un modelo de Markov oculto (HMM) que representa la manera en la que las características de la voz varían con el tiempo. Los parámetros del modelo HMM se pueden encontrar utilizando el algoritmo de Baum-Welch, que es una forma de EM.

La tarea de clasificación, dirigida por el sistema SpinVox puede declararse en una forma simplificada como: "De todas las cadenas de texto posibles, que podrían utilizarse para representar el mensaje, ¿qué cadena es la más probable dada la señal de voz de correo de voz registrada y las propiedades del idioma utilizado en el correo de voz?". Resulta inmediatamente evidente que éste es un problema de clasificación de grandes dimensiones y complejidad.

Los motores de reconocimiento de voz automáticos (ASR) han estado bajo desarrollo durante más de veinte años en laboratorios de investigación en todo el mundo. En los últimos años, las aplicaciones de solicitar voz continua, ASR de amplio vocabulario han incluido sistemas de dictado y automatización de centros de llamadas de los que son importantes ejemplos "Naturally Speaking" (Nuance) y "How May I Help You" (AT&T). Se hizo evidente que el desarrollo satisfactorio de sistemas basados en la voz depende, en gran medida, del diseño del sistema como ocurre con el rendimiento de ASR y posiblemente debido a este factor, los sistemas basados en ASR no han sido todavía utilizados por la mayoría de los usuarios de telecomunicaciones y de tecnología de la información IT.

Los motores de ASR tienen tres elementos principales. 1. La extracción de características se realiza en la señal de voz de entrada aproximadamente cada 20 ms para extraer una representación de la voz que sea compacta y tan libre como sea posible de interferencias incluyendo la distorsión de fase y las variaciones del aparato telefónico. Se suelen elegir coeficientes cepstrales del algoritomo de Mel-frecuencia y es conocido que se pueden realizar transformaciones lineales sobre los coeficientes antes del reconocimiento con el fin de mejorar su capacidad para discriminación entre los diversos sonidos de la voz. 2. Los motores de ASR emplean un conjunto de modelos, que se suelen basar en unidades trifónicas, que representan todos los diversos sonidos de la voz y sus transiciones precedentes y siguientes. Los parámetros de estos modelos se aprenden por el sistema antes de su desarrollo utilizando ejemplos de voz de capacitación adecuada de la voz. El procedimiento de capacitación estima la probabilidad de ocurrencia de cada sonido, la probabilidad de todas las transiciones posibles y un conjunto de reglas gramaticales que restringen la secuencia de la palabra y la estructura de sentencias de la salida de ASR. 3. Los motores de ASR utilizan un clasificador de modelos para determinar el texto más probable dada la señal de voz a la entrada. Los clasificadores del modelo de Markov Oculto se suelen preferir puesto que pueden clasificar una secuencia de sonidos con independencia de la velocidad al hablar y presentan una estructura muy adecuada para crear modelos de voz.

Un motor de ASR proporciona, a la salida, el texto más probable en el sentido de que se optimiza la coincidencia entre las características de la voz de entrada y los modelos correspondientes. Además, sin embargo, ASR debe tener también en cuenta la probabilidad de ocurrencia del texto de salida del reconocedor en el idioma objetivo. A modo de ejemplo simple, el texto en inglés "see you at the cinema at eight" es un texto mucho más probable que "see you at the cinema add eight", aunque el análisis de la forma de onda de la voz más probablemente detectaría "add" que "at" en el uso del inglés ordinario. El estudio de la estadística de ocurrencia de elementos de idioma se refiere como una modelización del idioma. Es frecuente en ASR utilizar la modelización acústica, que hace referencia al análisis de la forma de onda de la voz, así como la modelización del idioma para mejorar notablemente el rendimiento del reconocimiento.

El más sencillo modelo del idioma es un modelo de unigrama que contiene la frecuencia de ocurrencia de cada palabra

en el vocabulario. Dicho modelo se construiría analizando textos amplios para estimar la probabilidad de ocurrencia de cada palabra. Una modelización más sofisticada emplea modelos de n-gramas que contienen la frecuencia de ocurrencias de cadenas de n elementos en longitud. Es frecuente utilizar n = 2 (bigrama) o n = 3 (trigrama). Dichos modelos de idioma son de bastante mayor coste en el sentido de cálculo informático pero son capaces de calcular la utilización del idioma mucho más concretamente que los modelos de unigramas. A modo de ejemplo, los modelos de palabras de bigramas son capaces de indicar una alta probabilidad de que 'grados' irán seguidos por 'centígrados' o 'Fahrenheit' y una baja probabilidad de que vaya seguido por 'centípedo' o 'foráneo'. La investigación en la modelización del idioma está en curso a escala mundial. Las cuestiones incluyen la mejora de la calidad implícita de los modelos, la introducción de limitaciones estructurales sintácticas en los modelos y el desarrollo de formas informáticamente eficientes para adaptar los modelos de idiomas a diferentes idiomas y acentos.

Los sistemas de ASR de voz continua independientes del más amplio vocabulario de quien habla reivindican tasas de reconocimiento superiores al 95%, lo que significa menos de un error de palabra entre veinte. Sin embargo, esta tasa de errores es demasiado alta para ganar la confianza del usuario necesaria para una utilización a gran escala de la tecnología. Además, el rendimiento del ASR disminuye, en gran medida, cuando la voz contiene ruido o si las características de la voz no se adaptan adecuadamente con las características de los datos utilizados para capacitar los modelos del reconocedor. Un vocabulario especializado o coloquial tampoco está bien reconocido sin una capacitación adicional.

Para construir y desarrollar sistemas de voz basados en ASR satisfactorios se necesita claramente una optimización específica de la tecnología para la aplicación y una fiabilidad añadida y solidez obtenida al nivel del sistema.

Hasta la fecha, nadie ha investigado a fondo los requisitos del diseño práctico para un sistema de mensajería de voz híbrido, independiente del usuario, a gran escala, que pueda convertir mensajes de voz no estructurados en texto. Las aplicaciones claves son la conversión del correo de voz enviado a un teléfono móvil en texto y correo electrónico; otras aplicaciones, en donde un usuario desea comunicar oralmente un mensaje en lugar de introducirlo a través de un teclado (de cualquier formato) son también posibles, tales como mensajería instantánea, en donde un usuario comunica oralmente una respuesta que se captura como parte de un denominado *IM thread*; comunicar oralmente un texto, en donde un usuario comunica un mensaje que pretende que sea enviado como un mensaje de texto, como una comunicación de origen o una respuesta a un mensaje de voz o un texto o alguna otra comunicación; o bien, el sistema denominado '*speak-a-blog*', en donde un usuario comunica las palabras que desea que aparezcan en un *blog* y dichas palabras se convierten luego a texto y se añaden al *blog*. En realidad, en donde existe un requisito, o potencial ventaja a obtenerse de permitir a un usuario comunicar oralmente un mensaje en lugar de tener que introducir directamente ese mensaje como texto y tener ese mensaje convertido a texto y que aparezca en pantalla, en tal caso, se pueden utilizar los sistemas de mensajería de voz híbridos, independientes del usuario, a gran escala, de la clase descrita en la presente especificación técnica.

SUMARIO DE LA INVENCIÓN

En un primer aspecto de la idea inventiva, se da a conocer un sistema de mensajería de voz, independiente del usuario e independiente del dispositivo, a gran escala en conformidad con lo estipulado en la reivindicación 1. En un segundo aspecto de la idea inventiva, se da a conocer un método para proporcionar mensajería de voz utilizando un sistema de mensajería vocal, independiente del usuario e independiente del dispositivo, a gran escala, en conformidad con lo estipulado en la reivindicación 17.

En una forma de realización se da a conocer un sistema de mensajería vocal, independiente del usuario e independiente del dispositivo, a gran escala que convierte mensajes vocales no estructurados en texto para su presentación visual en una pantalla; el sistema que comprende (i) subsistemas puestos en práctica por ordenador y también (ii) una conexión de red a operadores humanos que proporcionan transcripción y control de calidad; estando el sistema adaptado para optimizar la eficacia de los operadores humanos comprendiendo, además: 3 subsistemas básicos, a saber (i) un extremo frontal de pre-procesamiento que determina una estrategia de conversión adecuada; (ii) uno o más recursos de conversión y (iii) un subsistema de control de calidad.

Otras formas de realización se proporcionan en el Anexo III. La invención es una contribución al campo del diseño de un sistema de mensajería vocal, independiente del usuario e independiente del dispositivo, a gran escala, que convierte mensajes vocales no estructurados en texto para su presentación visual en una pantalla. Según se explicó anteriormente, este campo presenta numerosos retos operativos diferentes para el diseñador de sistemas en comparación con otras áreas en las que se ha desarrollado el sistema ASR con anterioridad.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

La invención se describirá con referencia a los dibujos adjuntos, en donde las Figuras 1 y 2 son vistas esquemáticas de un sistema de mensajería vocal, independiente del usuario e independiente del dispositivo, a gran escala, que convierte mensajes vocales no estructurados en texto para su presentación visual en una pantalla, según se define por esta invención. Las Figuras 3 y 4 son formas de realización, a modo de ejemplo, de cómo el sistema presenta las posibles elecciones de palabras y de frases a un operador humano para su aceptación o rechazo.

DESCRIPCIÓN DETALLADA DE LA INVENCIÓN

Los diseñadores del sistema SpinVox tuvieron que hacer frente a numerosos retos operativos:

Modelos de reconocimiento de voz automático y de idiomas

En primer lugar y ante todo, era evidente para los diseñadores que la tecnología de ASR establecida, por sí misma, no era suficiente para proporcionar una STT fiable para el correo de voz (y otras aplicaciones de mensajería vocal, independientes del usuario, a gran escala). ASR confía en los supuestos establecidos a partir de modelos teóricos de la voz y del idioma incluyendo, a modo de ejemplo, modelos de idiomas que contienen probabilidades de palabras anteriores y reglas de gramática. Numerosos, si no todos, estos supuestos y reglas no son válidos en general para la voz del correo de voz. Los factores encontrados en la aplicación de STT de correos de voz, que están más allá de las capacidades de la tecnología de ASR estándar incluyen:

- la calidad de la voz está sujeta al ruido medioambiental, variación de codificación-decodificación y del aparato telefónico, interferencias de redes, incluyendo ruido y desvanecimientos vocales;

- los usuarios no saben que están hablando un sistema de ASR y se sienten cómodos dejando un mensaje que utiliza un lenguaje natural y a veces deficientemente estructurado;

- el propio lenguaje y acentos utilizados en el correo de voz no están restringidos ni son predecibles;

- variaciones del vocabulario ocurren con rapidez incluso dentro del mismo lenguaje de modo que, a título de ejemplo, las estadísticas del lenguaje pueden variar debido a importantes hechos sobre asuntos de actualidad.

Infraestructura de tecnología de la información IT

El diseño de la infraestructura de IT para mantener la disponibilidad y calidad del servicio de SpinVox establece demandas de exactitud en el poder de cálculo, red y ancho de banda de almacenamiento y disponibilidad de servidor. La carga en el sistema SpinVox está sujeta a máximos impredecibles así como a variaciones cíclicas más predecibles.

Mensajes no convertibles

Ha de preverse que una fracción de los mensajes sea no convertible. Estos podrían ser mensajes vacíos, tales como ‘slam-downs’, mensajes en un lenguaje no soportado o llamadas marcadas de forma no intencionada.

Evaluación de la calidad

La evaluación de la calidad en cada etapa del sistema SpinVox es, en sí misma, un reto operativo. El procesamiento de la señal proporciona numerosas técnicas de análisis que se pueden aplicar a la señal vocal, variando desde la medición de SNR directa a técnicas más sofisticadas que incluyen la detección explícita de interferencias comunes. Sin embargo, las mediciones directas, tales como éstas, no son significativas en sí mismas, sino que necesitan evaluarse en términos de su impacto sobre el proceso de conversión posterior. Análogamente, la confianza de ASR puede medirse en términos de la probabilidad de salida de hipótesis de reconocimiento alternativas pero, como anteriormente, es importante medir la calidad en términos de impacto sobre la conversión de texto global y la complejidad del control de calidad necesario para alcanzarlo.

Experiencia del usuario y factores humanos

El valor del sistema para los clientes viene influido, en gran medida, por el nivel de éxito con el que los factores humanos se admiten por el diseño. Los usuarios perderán rápidamente su confianza en el sistema si reciben mensajes distorsionados o encuentran que el sistema no es transparente ni de fácil uso.

Las anteriores dificultades se han superado en el diseño del sistema SpinVox como sigue:

Diseño del sistema. En la Figura 1 se representa un diagrama de bloques simplificado del diseño del sistema SpinVox que muestra las principales unidades funcionales. El núcleo básico está constituido por el motor de ASR 1. El sistema SpinVox establece una clara distinción entre ASR y una conversión de STT completa. ASR 1 es un subsistema que genera texto ‘bruto’, dada la señal de voz a la entrada. Éste es un elemento clave utilizado para la conversión de STT, pero es solamente uno de varios importantes subsistemas que se necesitan para conseguir una conversión de STT fiable. El subsistema de pre-procesamiento de extremo frontal 2 puede realizar una amplia clasificación de la señal de voz que se puede utilizar para determinar la estrategia de conversión en función de la elección de una combinación de motor ASR, conjunto de modelos y procesamiento de mejora de la voz. El subsistema de control de la calidad 3 mide la calidad de la voz de entrada y la confianza en el ASR, a partir de cuya información se puede determinar una estrategia de control de calidad .El subsistema de control de calidad 4 opera a la salida de ASR. Su finalidad es generar texto semánticamente correcto, significativo e idiomático, para representar el mensaje dentro de las restricciones del formato

de texto. El conocimiento del contexto, incluyendo el identificador ID del usuario que llama, los modelos de lenguajes específicos del destinatario y de la parte que hace la llamada, basados en el tiempo, se pueden utilizar para mejorar la calidad de la conversión, en gran medida, en comparación con la salida de ASR bruta. El texto convertido se proporciona, por último, desde el subsistema de procesamiento de lenguaje pos-conversión 5 a un texto de SMS y pasarela de correo electrónico.

Características principales. Las principales características del método adoptado por SpinVox son:

- Conversión de mensaje significativa

La conversión de texto captura el mensaje, su significado, estilo e idiomas, pero no es necesariamente una conversión de palabra por palabra del correo de voz.

- Tiempo de ida y vuelta

El tiempo de ida y vuelta para la conversión de un mensaje está garantizado.

- Fiabilidad

El sistema nunca puede enviar un mensaje de texto distorsionado. Se avisa a los abonados con respecto a los mensajes no convertibles, que se pueden escuchar en la forma convencional.

- Lenguaje estándar

Los mensajes se envían en lenguaje estándar y no están 'textificados'.

- Amplia disponibilidad

El sistema funciona íntegramente en la infraestructura y no realiza ninguna exigencia operativa sobre el aparato telefónico o la red que no sea la de desvío de llamadas.

- Operación adaptativa

El sistema puede optimizar su rendimiento utilizando estrategias de control de la calidad incorporadas, impulsadas por el conocimiento obtenido en el transcurso del tiempo respecto a la modelización de lenguajes de correos de voz, en general, así como la modelización de lenguajes específicos del usuario que llama. Además, el sistema puede elegir de entre varias posibles estrategias de conversión de voz a texto, basadas en las características del mensaje de correo de voz. Los datos de mensajes de correo de voz y las correspondientes conversiones de texto se analizan continuamente con el fin de actualizar y adaptar el sistema STT de SpinVox.

- Supervisión de la calidad

La calidad de la conversión de voz a texto se puede supervisar en cada etapa y de este modo, se puede realizar eficientemente un control de la calidad, por agentes humanos o por agentes automáticos.

- Procesamiento del lenguaje

El procesamiento del lenguaje pos-conversión 5 se puede realizar para mejorar la calidad del texto del mensaje convertido, eliminar redundancias obvias y validar los elementos del mensaje tales como las estructuras de saludos que suelen utilizarse.

- ASR de tecnología moderna

Los motores de ASR comerciales se pueden utilizar con el fin de obtener una ventaja competitiva de la tecnología de ASR más moderna. Diferentes motores de ASR pueden solicitarse para gestionar mensajes diferentes o incluso partes diferentes del mismo mensaje (con la unidad de decisión 2 decidida en cuanto a qué motor utilizar). Los operadores humanos, por sí mismos, podrían considerarse también como una instancia operativa de un motor de ASR, que es adecuado para algunas tareas, pero no para otras.

- Estable y seguro

El servicio se realiza en servidores de Unix seguros y muy estables y pueden adaptarse a la demanda para varios lenguajes puesto que diferentes máximos de experiencia en zonas temporales se dispensan a través de cada periodo de 24 horas.

Control de la calidad. El sistema SpinVox ha desarrollado un conocimiento detallado de las expectativas y deseos de los

usuarios con respecto a sistemas de mensajería basados en telefonía. Han identificado una tolerancia cero de los usuarios para la conversión de voz a texto sin sentido; en particular cuando hay prueba evidente de que se han introducido errores por una máquina que no sea por error humano. El control de la calidad del texto convertido es, por lo tanto, de importancia clave. Se pueden utilizar tres estrategias de calidad alternativas; la unidad de decisión 2 selecciona la óptima. (i) mensajes para los que la confianza en la conversión de ASR es suficientemente alta se pueden comprobar automáticamente por el subsistema de evaluación de la calidad 3 en cuanto a su conformidad con las normas de calidad. (ii) Los mensajes para los que la confianza en la conversión de ASR no es suficiente alta, se pueden encaminar a un agente humano 4 para su comprobación y, si fuera necesario, su corrección. (iii) Los mensajes para los que la confianza en la conversión de ASR es muy baja se marcan como no convertibles y se informa al usuario de la recepción de un mensaje no convertible. Los mensajes no convertibles se pueden escuchar por el usuario, si así lo desea, utilizando la pulsación de una sola tecla. El resultado de estas estrategias es que el sistema SpinVox está diseñado de modo que un fallo en la conversión sea favorecido generando una conversión que contenga errores. La confianza del usuario en el sistema está, por lo tanto, protegida. Las estadísticas de SpinVox indican que un importante porcentaje de correos de voz se convierten de forma satisfactoria.

Una de las importantes herramientas utilizadas por SpinVox para mejorar la calidad de los mensajes convertidos es el conocimiento del lenguaje (frases comunes, saludos comunes y cierres de la transmisión, etc.) utilizados en los mensajes de correos de voz. A partir de los datos acumulados en el transcurso del tiempo, se pueden desarrollar modelos de lenguajes estadísticos específicos para la voz del correo y luego, utilizarse para guiar el proceso de conversión de STT. Esto mejora, en gran medida, la exactitud de la conversión para construcciones de lenguajes no estándar.

La característica más evidente de SpinVox es que suministra un servicio que numerosos usuarios no se percataban de que lo necesitaban, pero pronto encuentran que no pueden efectuar ninguna gestión sin dicho servicio. Se trata del primer sistema en tiempo real que proporciona una conversación de voz a texto de los correos de voz. Su impacto para los operadores de redes es un aumento del tráfico de la red, mejora de la continuidad de llamadas, tanto para voz como para datos. El éxito objetivo del sistema SpinVox se ha conseguido adoptando un método de diseño que esté orientado primero por la calidad del servicio y en segundo lugar, por la tecnología. El diseño del sistema está basado en un conocimiento detallado de las expectativas de los usuarios del servicio y, lo que es más importante desde una perspectiva de la ingeniería, los puntos fuertes y débiles de la tecnología de ASR. Explotando los puntos fuertes de ASR y subsanando los puntos débiles mediante un control de calidad riguroso, SpinVox es un desarrollo efectivo que cumple los requisitos de diseño prácticos para un sistema de mensajería vocal híbrido no estructurado, independiente del usuario, y a gran escala.

SpinVox ha demostrado su éxito en proporcionar un servicio basado en el procesamiento de la voz concentrando su tecnología de conversión sobre una aplicación objetivo muy específica, que es la conversión del correo de voz. La indicación es que el diseño del sistema que es objetivo para una aplicación muy bien definida es un método más productivo que la búsqueda, aparentemente de carácter ilimitado, para mejoras, cada vez menores, en las medidas del rendimiento brutas de, a modo de ejemplo, los motores de ASR. Este método abre la posibilidad de nuevas áreas de aplicación en las que se podrían desarrollar los componentes de tecnología y el conocimiento del diseño de sistemas de propiedad privada de SpinVox.

SpinVox ha desarrollado un importante conocimiento tecnológico propio como arquitecto de sistemas para aplicaciones basadas en la voz con su propia experiencia que cubre los campos del reconocimiento de la voz, aplicaciones de telecomunicaciones, redes celulares y factores humanos. Las oportunidades para el crecimiento el desarrollo, en tecnologías de mensajería avanzadas, probablemente han de orientarse a permitir la integración de la mensajería vocal y de texto, con lo que se facilitan las operaciones de búsqueda, gestión y archivado de correos de voz con todas las mismas ventajas anteriormente disfrutadas por el correo electrónico y la mensajería de texto de SMS, incluyendo la simplicidad operativa y la documentación automática. Dichos desarrollos están en paralelo con la convergencia de la voz y de los datos en otros sistemas de telecomunicaciones.

El sistema de SpinVox se está desplazando desde lo que, desde el exterior, es un problema independiente de la persona que habla con respecto a un problema dependiente del altavoz, que es un gran discernimiento en la realización del trabajo de la voz en telefonía. La razón es que se está utilizando el hecho de que las llamadas, mensajería y otra comunicación se impulsan por el uso comunitario; a modo de ejemplo, un 80% de los correos de voz procedente de solamente 7 a 8 personas. Mensajes SMS solamente de 5 a 6 personas. Mensajería instantánea IM solamente de 2 a 3 personas. SpinVox utiliza el registro histórico de 'llamada del par' para hacer varias operaciones:

1. Elaborar un perfil de lo que un determinado usuario llamante dice cada vez que llama – un modelo de altavoz dependiendo del usuario que habla – cómo habla el usuario que llama (entonación, etc.);

2. Establecer un modelo de lenguaje de lo que el usuario que llama dice a alguien – un modelo de lenguaje dependiente del usuario que habla – de lo que dice el usuario que llama (palabras, gramática, frases, etc.);

3. En los apartados 1 y 2, estamos realmente construyendo un modelo de lenguaje de cómo A habla a B. Éste es un modelo más perfeccionado que el basado en cómo habla A en general. Es atípico para el tipo de mensajería (esto es, de cómo habla en el correo de voz) y es también atípico de cómo habla a B (p.e., la manera en que una persona

comunica un mensaje a su madre es muy diferente en entonación/gramática/frases/acento/etc., que cuando se lo comunica a su esposa).

4. SpinVox está construyendo modelos de pares de hablante-receptor que van desde la independencia del lenguaje/usuario que habla, en general, al que depende sin ninguna entrada del usuario ni capacitación alguna;

5. SpinVox tiene la capacidad para utilizar el lenguaje de ambas partes entre sí (p.e., cómo puedo efectuar una rellamada y dejar un mensaje) a un perfeccionamiento adicional de las palabras pertinentes (p.e., diccionario), gramática/frase, etc.

En el anexo I siguiente se proporcionan más detalles sobre estos aspectos del sistema de conversión de mensajes vocales de SpinVox.

Anexo I
SpinVox – Sistema de Conversión de Mensajes Vocales

El Sistema de Conversión de Mensajes Vocales (VMCS) de SpinVox se concentra en un solo objetivo: la conversión de mensajes hablados en su equivalente de texto significativo. Este objetivo es único como lo son los métodos y tecnologías avanzadas aquí utilizadas.

Concepto

Un nuevo método de conversión de mensajes vocales en texto utilizando técnicas de reconocimiento automático multietápicas y procesos y técnicas de control de calidad y de garantía de la calidad con asistencia humana. Los elementos automatizados y humanos interaccionan directamente entre sí para generar una realimentación en tiempo real/directa que es básica para que el sistema sea siempre capaz de aprender a partir de datos directos para permanecer en sintonía operativa y para proporcionar una calidad estable. Asimismo, está diseñado para sacar partido de los límites inherentes de AI (ASR) y mejorar, en gran medida, la exactitud mediante el uso de vínculos contextuales, guía humana y datos de lenguajes directamente desde Internet.

Cuestión problemática

Los métodos tradicionales para la conversión de la voz han sido diseñados, en gran medida, al nivel del reconocedor y con la creación de un reconocimiento de voz automático de alta calidad en condiciones de laboratorio, en donde las entradas son muy controladas y garantizan un alto nivel de exactitud.

El problema es que, en el mundo real, el reconocimiento de la voz tiene otros numerosos elementos a superar con:

- Usuarios llamantes aleatorios – cualquiera puede utilizarlo;
- Entrada ruidosa – ruido de fondo y calidad deficiente del altavoz;
- Calidad de transmisión deficiente y variable con pérdida de comprensión y conexiones defectuosas de teléfonos móviles;
- Voz gramaticalmente incorrecta, modismos o expresiones muy localizadas;
- Gramática sensitiva contextual o significado implícito desde un contexto unido entre el creador del mensaje y su destinatario;
- Cambios de contexto dentro de un mensaje – límites del contexto – que invalidan el uso de reglas gramaticales normales

para citar tan solo unos pocos y todos ellos constantemente variando en el tiempo, por lo que la entrada origen real no es un problema definido en el tiempo, sino un problema en constante evolución.

Solución

La clave es definir correctamente el problema: conversión de mensajes vocales en su equivalente en texto significativo.

Esto no significa una transcripción prolija perfecta, en lugar de los elementos más importantes del mensaje presentados en una forma fácilmente entendible. Las medidas de la exactitud son cuantitativas y cualitativas puesto que la calificación última es una precisión de calidad de la voz del usuario en donde el VMCS de SpinVox consigue una valoración constante del 97%.

Existen dos partes principales:

- Utilización de un mecanismo de realimentación directa constante para el sistema de aprendizaje, orientada al operador humano.

- Utilización de información contextual para definir mejor cada problema de conversión.

La utilización de la información contextual ayuda al sistema a estimar mejor la probabilidad de que alguna cosa incluida en un mensaje dada las características de:

- Tipo

- Longitud

- Hora del día

- Geografía

- Contexto del usuario que llama – llamante y destinatario (registro histórico de par-llamada)

- Recientes eventos operativos

- Etc.

y la estructura del lenguaje conocida que es más probable que ocurra en algunos tipos de mensajes – lenguaje natural, según se describe a continuación.

Lenguaje natural

Cuando se analizan mensajes vocales y mensajes de texto hablados, modelos regulares ocurren en qué dice el usuario, cómo lo dice y en qué orden – lenguaje natural. Esto varía claramente por contexto o tipo de mensaje, de modo que para pedir una pizza sería diferente.

A modo de ejemplo, en el correo de voz, cómo se saludan los usuarios puede definirse bien con 35 o un número similar de las expresiones más comunes – “Hola, soy yo”, “Hola, aquí Daniel”, “Qué pasa, ¿Qué estás haciendo?”, “¿Cómo estás compañero?”, “¿Correcto?”, etc. y análogamente, la expresión de despedida puede estar bien definida por expresiones comunes – “De acuerdo, adiós”, “Saludos”, “Saludos colega”, “Gracias ahora, adiós, saludos”, etc…

Evidentemente, diferentes partes de un mensaje hablado tienen significado implícito y por lo tanto, con la utilización de esta clave podemos mejorar la exactitud del reconocimiento utilizando este contexto para seleccionar la clasificación más probable de lo que realmente se dijo.

Su construcción en modelos relacionados, estadísticamente amplios, es lo que se define como nuestro Modelo de Lenguaje Natural, uno para cada lenguaje, incluyendo dialectos dentro de cualquier lenguaje.

Vectores de contexto

El lenguaje natural se suele regular por su contexto, por lo que cuando se convierte el cuerpo de un mensaje, el contexto de lo que se ha dicho puede utilizarse para estimar mejor lo que fue realmente dicho – p.e., una llamada a una línea de reserva contendrá, más probablemente, expresiones relacionadas con la solicitud y algunos nombres específicos de la empresa, producto y precio frente a una llamada a un número de teléfono doméstico, en donde las expresiones amistosas con respecto a los saludos, ‘cómo estás’, ‘llámame más tarde’, etc., son mucho más probables.

Dentro de los mensajes vocales, podemos utilizar vectores de contexto para estimar mejor el probable contenido y el lenguaje natural establecido que se aplica:

- CLI (o cualquier identificador de parte) es un vector de contexto muy potente

    - Puede referirse a la zona geográfica del número probablemente y al idioma/dialecto y nombres reales regionales de más probable utilización.

    - Puede indicarse si el número es un número comercial conocido y por lo tanto, un tipo de mensaje mejor predicho – p.e., las llamadas desde números 0870 son comerciales y por lo tanto, con una alta posibilidad de que se trate de un mensaje de negocios mientras que la gama 07 es de un teléfono móvil personal, por lo que la hora del día motivará a mensajes de un tipo más probable entre de contenido de negocios, personales, sociales u otros.

    - Le permite obtener mejor su número correcto si se comunica dentro del mensaje.

- Es una clave a partir de la que puede establecer un registro histórico y diccionario/gramática conocida – p.e., siempre dice en idioma inglés *'dat's wicked man'* en un acento 'de la calle'.

- Puede construir un sistema de reconocimiento dependiente del usuario que habla – esto es, podemos ajustarle el ASR como un usuario llamante particular y obtener una mucho más alta precisión del reconocimiento, su propio vocabulario, gramática, fraseología, léxico y lenguaje natural general.

- Registro histórico de llamada del par – utilización más profunda de CLI (o cualquier identificador de parte)

  - Puede capacitar el sistema con bastante más exactitud para un registro histórico de mensajes del tipo de llamada del par.

  - Puede capacitar para la voz de la parte A (usuario que llama) haciendo caso omiso de la parte B (destinataria).

  - Puede capacitar para el área de contenido y lenguaje de la parte A que utiliza con la parte B.

  - Puede capacitar múltiples relaciones de las partes A y B y desarrollar el sistema para una más alta precisión y velocidad.

- Hora del día, día de la semana

  - Tasas de tráfico de correo de voz, longitud media de los mensajes y el tipo de contenido varían con la hora del día en cada mercado idiomático, desde mensajes muy de negocios durante las horas de demanda máxima (8 de la mañana a 6 de la tarde) a más personales (7 a 10 de la tarde) a muy sociales (11 de la noche a 1 de madrugada) a muy funcional (2 a 6 de la mañana). Esta situación varía también por el día de la semana siendo un miércoles el día de mayor ocupación y contiene más altos niveles de mensajes de negocios, pero en sábado y domingo presentan un perfil muy distinto de tipo de mensaje (en gran medida mensajes personales de conversación) que necesitan tratarse de forma distinta.

- Números internacionales

  - Realizando un análisis del código de país (p.e., 44, 33, 39, 52, 01) podemos determinar mejor el idioma y el dialecto.

- Datos de clientes disponibles

  - Nombre del cliente, su domicilio y posiblemente su lugar de trabajo.

Contexto implícito entre las partes A y B

Asimismo, existen muchas otras muy importantes pistas que pueden ayudarnos a estimar mejor el probable contenido de un mensaje, en particular los que se relacionan a quién de las dos partes son, cuál es la finalidad probable del mensaje y a dónde o desde dónde se está llamando.

En la conversión de correos de voz a texto y texto hablado, conocemos que el hecho de disponer del número del usuario que llama

- Nos permite estimar mejor cualquier número dejado en el interior del mensaje.

- Establecer el registro histórico de palabras, expresiones, frases, etc., conocidas entre las dos partes.

- Idioma probable (p.e., una llamada desde +33 a +33 probablemente será en francés, pero las llamadas de +33 a +44 pueden tener un 50% de posibilidades de estar en francés).

- Nombres y su ortografía correcta.

Si conoce el registro histórico de las llamadas/mensajes de la parte A y su registro histórico de mensajes para la parte B, podrá establecer un perfil dependiente del usuario que habla y obtener importantes mejoras de su reconocedor y su gramática.

Calidad de conversión

Para resolver este problema, la definición del resultado requerido real es esencial puesto que establece una gran diferencia en su método para resolver cómo convertir los mensajes vocales (correos de voz, mensajes SMS verbales, mensajería instantánea, etc.) a texto y cómo aplicar, de modo óptimo, el recurso de conversión del que dispone.

Cuando alguien nos deja un mensaje de voz, la finalidad es un mensaje, y no un elemento de comunicación escrito formal, por lo que se tolerará una conversión menos exacta en tanto que se transmita correctamente el significado del mensaje.

Además, existe una asimetría, de modo que el depositante del mensaje no está comparando con lo que se dijo con el texto convertido. Con el contexto de quien hace la llamada, el destinatario está leyendo la salida convertida con el objetivo de averiguar de qué trata el mensaje, por lo que el requerimiento es una excelente extracción del mensaje para su conversión y no una conversión prolija (palabra por palabra, expresión por expresión). En realidad, por el contrario, una conversión prolija, a no ser que esté bien dictada, se suele percibir como un mensaje de baja calidad puesto que contiene abundantes elementos indeseados y no elegantes del lenguaje de mensajes vocales (p.e., uhmms, ahhs, repeticiones, deletreos de palabras, etc.).

Por lo tanto, la calidad en este contexto se refiere a la extracción de los elementos importantes de un mensaje.

- Conversión inteligente

En su forma más simple, existen tres elementos claves que proporcionan el máximo del significado y por lo tanto, son esenciales para conseguir una buena calidad del mensaje:

1. De quién es – con gran valor para comprender el significado desde este contexto.

2. Cuál es la finalidad del mensaje – p.e., llámame con urgencia, realización tardía, cambio de planes/tiempos, llámame a este número, solamente decirte hola, etc.

3. Cualesquiera hechos específicos, siendo los más comunes:

    a. Nombres

    b. Números, números telefónicos

    c. Hora

    d. Dirección

Otra información en el mensaje es en gran medida, en soporte de la transmisión de estos elementos claves y suele ayudar a proporcionar un mejor contexto para estos elementos claves.

Variación de la sensibilidad de la calidad dentro del mensaje

Lo que también es muy importante entender es que necesitamos reconocer que cada parte principal de cualquier mensaje tiene una diferente función en la entrega del mensaje y por lo tanto, podemos atribuir otra dimensión de calidad a cada una con el objetivo de conseguirlo durante la conversión.

Los mensajes se pueden desglosar en:

- Saludos (parte superior)

- Mensaje (cuerpo)

- Despedida (final)

El porcentaje de mensajes que contienen cualquier cuerpo es evidentemente una función de la longitud del mensaje depositada, por lo que conocemos que los mensajes cortos (p.e., de menos de 7 segundos) suelen contener solamente un saludo y una despedida. Ante todo, la probabilidad de un cuerpo de mensaje significativo aumenta de forma exponencial. Este hecho nos ayuda también a una mejor estimación de la probable estrategia de conversión que deberíamos utilizar.
Saludos y despedidas

La forma en que alguien le saluda puede clasificarse en unos 50 saludos frecuentemente reconocidos (p.e., “hola a todos”, “Hola, soy yo”, “Hola, ésta es una llamada a X desde Y”, “Hola, simplemente te llamo para….”, etc.). Análogamente, el elemento de ‘despedida’ de un mensaje puede clasificarse en un orden de magnitud similar de despedidas de reconocimiento común (p.e., “Muchas gracias”, “Nos vemos”, “Ta”, “Adiós”, “Te saludo”, “Nos vemos más tarde”, etc.).

Dos cuestiones dictan nuestra exigencia de calidad de la conversión:

1. Los saludos y despedidas son para protocolo de mensajes y suelen contener poco valor para el mensaje principal, por lo que nuestra tolerancia para baja exactitud es alta, a condición de que tengan sentido.

2. Podemos clasificar la amplia mayoría de saludos y despedidas en unas 50 categorías de frecuente reconocimiento cada una.

Por lo tanto, la exigencia de calidad durante un saludo, una acogida o despedida es bastante menor que respecto al contenido en el cuerpo del mensaje, normalmente la cuestión principal de un mensaje o un hecho clave – p.e., llámame al 020 7965 2000.

Cuerpo del mensaje

Por supuesto, el cuerpo del mensaje tiene un más alto requerimiento de calidad, pero es probable que pueda encontrarse, con frecuencia, que contiene modelos regulares del lenguaje naturales que se relacionen con el contexto y que, por lo tanto, podemos aplicarle también un grado de clasificación para ayudarnos mejor a conseguir la respuesta correcta.

Un ejemplo adecuado es:

“Hola Dan, soy John” – Parte superior del mensaje (o saludo)

“Puedes hacerme una rellamada al 0207965200 cuando puedas” – Cuerpo del mensaje

“Muchas gracias compañero, saludos, adiós” – Final del mensaje (o despedida).

En este caso, el cuerpo del mensaje es un elemento bien estructurado del Lenguaje del correo de voz que ha aprendido el sistema de conversión de SpinVox. A continuación, puede desglosarse el mensaje y efectuar una asignación correcta al resultado del cuerpo del mensaje.

Los elementos que se aplican en este caso son:

- Partes A y B conocidas

- El número de teléfono es CLI de John o visto antes en sus llamadas a otros

- Longitud del mensaje – inferior a 10 segundos, por lo que es más probable la expresión común

- Hora del día - horario de trabajo – John normalmente no deja mensajes detallados en el horario de trabajo sino simplemente mensajes simples y cortos.

Sistema de conversión de mensajes vocales de SpinVox

Habiendo indicado correctamente nuestro problema e identificado algunas características muy importantes de la voz y cómo se relaciona con el equivalente de texto, el sistema de SpinVox (véase Figura 2) fue diseñado para obtener la máxima ventaja competitiva de estas características:

Sistema de conversión de mensajes vocales de SpinVox

Este diagrama ilustra las tres etapas claves que nos permiten optimizar nuestra capacidad para convertir correctamente los mensajes vocales (correo de voz, mensaje SMS hablado, mensajes instantáneos, clips de voz, etc.) a texto.

Un concepto clave es que el sistema utilice el término Agente para cualquier recurso de conversión, bien sea humano, bien sea basado en ordenador/máquina.

Pre-procesamiento

Se puede desglosar en dos características:

1. Optimiza la calidad de la señal de audio para nuestro sistema de conversión eliminando ruido, depurando defectos conocidos, normalizando energía de señal/volumen, eliminando secciones en silencio/vacías, etc.

2. Clasifica el tipo de mensaje para un encaminamiento óptimo del mensaje para su conversión o no.

La clasificación del tipo de mensaje se realiza utilizando una gama de ‘Detectores’:

- Idioma

- P.e., inglés del Reino Unido / Estados Unidos / Australia / Nueva Zelanda / Sudáfrica / Canadá y luego, los tipos de dialectos dentro de dichos idiomas (p.e., dentro del Reino Unido – S. East, Cockney, Birmingham, Glasgow, Irlanda del Norte, etc.).

- Nos permite determinar si soportamos el idioma.

- Permite seleccionar qué ruta de conversión utilizar: perfil de QC/QA, reglas de TAT (SLA), qué estrategia de etapas de ASR (motores) cargar y qué estrategia de pos-procesamiento aplicar.

Métodos:

- Identificación de lenguaje estadístico

  - Técnica anterior:

    - varios métodos de identificación de lenguaje automática conocidos

  - Solución de SpinVox:

    - decisión de base sobre contexto: conocimiento sobre registro, localización y registro histórico de llamadas del usuario que llama y del receptor

- Identificación del lenguaje basado en la señal

  - Problema con la técnica anterior:

    - los métodos de alta precisión requieren un reconocimiento de voz con gran vocabulario o al menos reconocimiento fonético, por lo que resulta de alto coste en su obtención y utilización

    - requerimiento para un método fiable y rápido basado exclusivamente en los registros (etiquetados con el idioma pero nada más)

  - Solución de SpinVox:

    1. agrupamiento automático de datos de voz para cada idioma (cuantización vectorial)

    2. combinación de centros de agrupamiento

    3. utilización del modelo estadístico de secuencia de agrupamientos para cada idioma para encontrar una mejor coincidencia

    4. establecer un modelo de relación entre diferencias de calificación entre modelos y la exactitud prevista

    5. combinación de varias versiones de 1-4 (basadas en datos de capacitación variables, métodos de extracción de características, etc.) hasta que se consiga la exactitud deseada.

- Ruido – detector de SNR

  - Si la cantidad de ruido en un mensaje es superior a un determinado umbral, entonces se hace cada vez más difícil detectar correctamente la señal del mensaje y proceder a su conversión. Más significativa es si la relación de la señal a ruido se hace inferior a un determinado nivel, en cuyo caso, tendrá un alto grado de confianza y no siendo capaz de convertir el mensaje.

  - los usuarios de SpinVox valoran el hecho de que cuando reciben una notificación de que el mensaje era inconvertible, el audio origen es tan deficiente que más del 87% del tiempo en que se llama o el texto, la persona vuelve directamente y continúa la ‘conversación’.

- Estimador de calidad de la voz

  - Si la calidad de la voz de alguien es probablemente demasiado baja para la utilización del sistema de conversión o del agente. O el contenido que un usuario habría de escuchar para sí mismo – p.e., alguien les llama con el sonido de cumpleaños feliz.

  - La solución de SpinVox incluye:

    1. encontrar las eliminaciones (paquetes de voz perdidos durante la transmisión) basándose principalmente

en los recuentos de cruces de cero

2. estimar también los niveles de ruido

3. calcular la medida global de la calidad de la voz y utilizar un umbral adaptativo para rechazar los mensajes de la más baja calidad.

- Detector de la operación de descolgar (‘slam-down’)
  - Mensajes en donde alguien llamó, pero no dejó ningún contenido de audio significativo. Normalmente, mensajes cortos con expresiones verbales en segundo plano.
- Detector de llamadas inadvertido
  - Normalmente, una llamada desde la pulsación de un botón de remarcación mientras está en el bolsillo de alguien y dejando un mensaje de estruendo largo sin ningún contenido en audio significativo.
- Mensajes estándar
  - Mensajes pregrabados, más comunes en Estados Unidos, desde un sistema de marcación automática o notificaciones de servicios o llamadas.
- Saludos y despedida
  - Si el mensaje solamente contiene estas partes, en tal caso, podemos utilizar un elemento dedicado de ASR para convertir correctamente estos mensajes.
- Longitud del mensaje y densidad de voz
  - La longitud nos permite estimar inicialmente la probabilidad del tipo de mensaje – p.e., la llamada corta normalmente consiste simplemente en ‘hola, soy X, llámame después por favor’ frente a una llamada larga que contendrá algo más complejo para convertir.
  - La densidad de voz le permitirá ajustar su estimación de la probabilidad de que una longitud de mensaje sea un buen indicador del tipo – baja densidad y mensaje corto es probable que sea simplemente ‘hola, soy X, llámame después, por favor’, pero un mensaje corto, de alta densidad, efectuará un sesgo hacia el usuario necesitando un más alto nivel de recursos de conversión puesto que la complejidad del mensaje será más alta.

Evidentemente, el pre-procesamiento nos permite en algunos casos (p.e., slan-down (descolgar), llamada inadvertida, idioma extraño/no soportado) encaminar inmediatamente el mensaje como clasificado y enviar el texto de notificación correcto al destinatario (p.e., ‘esta persona llamó, pero no dejó ningún mensaje’), ahorrando cualquier uso adicional de valiosos recursos del sistema de conversión.

Reconocimiento de voz automático (ASR)

Se trata de un proceso dinámico. El uso óptimo de recursos de conversión se determina a un nivel de mensaje.

Asumimos la entrada desde la etapa de pre-procesamiento en la clasificación de mensajes y desde cualesquiera vectores de contexto y lo utilizamos para elegir la estrategia de conversión óptima. Esto significa que esta etapa está utilizando la mejor tecnología de ASR para la tarea particular. El motivo es que diferentes tipos de ASR son muy adecuados para tareas específicas (p.e., uno es excelente para saludos, otro para números telefónicos, otro para direcciones en francés).

Esta etapa está diseñada para utilizar una gama de agentes de conversión, sean de ASR o humanos y solamente discierne entre ellos en función de cómo la lógica de conversión está configurada en ese momento. A medida que aprende el sistema, esta situación es adaptada y se pueden utilizar diferentes estrategias, recursos de conversión y secuencias.

Esta estrategia no solamente se aplica al nivel de mensaje completo, sino que también puede aplicarse dentro de un mensaje.

Parte superior ‘n’ final mensaje

Una estrategia consiste en subdividir las secciones de saludos (parte superior), cuerpo y despedida (final) del mensaje, enviarlas a diferentes elementos de ASR que son óptimos para ese elemento de un mensaje. Una vez que se ha concluido esta operación, se ensamblan como un mensaje único.

Encaminamiento de números

Otra estrategia consiste en subdividir cualesquiera elementos simples en donde los números telefónicos, las monedas o el uso obvio de números que se comunican en el mensaje. Estas secciones se envían a un elemento específico de ASR o del agente para su conversión óptima y luego, se vuelven a ensamblar con el resto de los mensajes convertidos.

Encaminamiento de direcciones

Análogamente, la subdivisión de elementos en donde una dirección está siendo comunicada puede enviarse a un elemento específico de ASR o del agente y a un validador de coordenadas de direcciones para garantizar que todos los elementos de una dirección convertida sean reales. A modo de ejemplo, si no puede detectar el nombre de la calle, pero tiene un código postal claro, podrá completar el nombre de calle más probable. La exactitud de encontrar el nombre de la calle se mejora mediante un nuevo procesamiento de la dirección, pero con su nombre de calle estimado *a priori* reajustando sus variables de clasificación de ASR a un conjunto mucho más limitado y constatando si existe, o no, una alta coincidencia.

Encaminamiento de nombres reales

Los nombres reales se renombran para obtener un ASR no fiable. De nuevo, concentrándose en solamente esta parte y aplicando un recurso más especializado, pero informáticamente de mayor coste, podrá estimar mucho mejor el nombre real.

Procesamiento de correo

La etapa de ASR contiene sus propios diccionarios y gramático, pero esto no es suficiente para convertir correctamente las formas más complejas con las que hablamos. ASR está muy orientado a la conversión al nivel de palabras y secuencias muy cortas de palabras para estimar la probabilidad de secuencias de palabras y gramática básica (técnicas de retícula y n-gramas). Un problema es que, desde el punto de vista matemático, cuando amplía el conjunto de palabras intentando estimar las posibles combinaciones más allá de 3 o 4, las permutaciones se hacen tan grandes que su capacidad para captar la correcta disminuye con más rapidez que cualquier ventaja obtenida ampliando el número de palabras secuenciales, por lo que es actualmente una estrategia no fiable.

Un método muy adecuado es buscar frases más amplias o estructuras de sentencias que ocurren en el lenguaje natural. Enfocando el problema desde el nivel de macros, podrá estimar mejores soluciones para errores o palabras/secciones en donde la confianza en ASR sea baja.

Sin embargo, este método también tiene sus puntos débiles. Como se indicó anteriormente, la voz humana contiene mucho ruido, distorsiones y debido a la relación íntima entre las partes A y B es propensa a grandes limitaciones contextuales. Algunas cosas carecen de sentido para alguien distinto de la persona que tiene un amplio conjunto de contexto en el que tienen mucho sentido frases aparentemente aleatorias o palabras de sonidos no fiables.

A modo de ejemplo, la frase en inglés *"see you by the tube at Piccadilly opposite the Trocadero and mine's a skinny mocha when you get here"* no tendría sentido para alguien a no ser que conociera el posible significado de *'tube'*, hayan estado en Londres y conocieran que Piccadilly tiene un edificio llamado el 'Trocadero' muy próximo y entendiera que en Starbucks cercano sirven una bebida llamada 'mocha' y es baja en calorías, es decir, *'skinny'*.

Cuerpos del mundo real - comprobación de contexto

Una solución es buscar un cuerpo muy amplio de palabras inglesas, nombres reales, frases, modismos, expresiones ordinarias para comprobar que su conversión podría haber contenido estas secuencias de palabras.

El problema es que en la voz normal existen grandes combinaciones posibles de los elementos anteriores y de forma crítica, esto carece de ninguna comprobación de contexto en el mundo real. ¿Cómo conoce que las combinaciones de nombres reales Piccadilly, Trocadero, mocha y skinny son válidas, permitiendo sólo buenas conversiones de sus señales de audio origen? Lo único absoluto es una comprobación del mundo real y lamentablemente, por definición, los seres humanos son los únicos, en este momento, capaces de discernir si algo tiene, o no, una validez en el mundo real – lo hacemos después de que todos los ordenadores de programas fijos y bases de datos sean fiables al respecto.

Con la inteligencia al nivel humano, se puede comprobar, con la mayor precisión, si estos elementos aparentemente desconectados tienen cualquier probable contexto en el mundo real. Sin embargo, los seres humanos carecen también de un conocimiento completo de todo lo que tiene un porcentaje significativo de londinenses que no se encontrarían cómodos conociendo si esta frase era probable, o no, habida cuenta de su conocimiento de Piccadilly.

Una solución consiste en utilizar los mayores cuerpos del planeta del conocimiento humano. Los miles de millones de páginas y bases de datos creadas por editores humanos disponibles en Internet. Una simple consulta en cuanto a si la sentencia, o cualquier elemento de su conversión, se cita en Internet le proporciona una prueba del mundo real, muy

cualificada, de si se trata de algo que los seres humanos probablemente han experimentado y registrado y por lo tanto, podría ser real. Por ello, en el ejemplo anterior, encontramos que Google, Yahoo!, MSN y otros importantes motores de búsqueda son capaces de proporcionar bastantes aciertos de páginas con estos elementos por cuanto que tenemos una confianza muy mejorada de que nuestra conversión es correcta, en realidad.

Además, utilizando Internet, podemos, con gran frecuencia, encontrar la ortografía correcta de aproximaciones fonéticas de palabras, nombres reales y nombres de lugares que ASR intenta con nuevas o palabras desconocidas si se atraviesa. Actualmente, esta operación se realiza a través de una programación manual de alto coste y gran consumo de tiempo de los diccionarios de ASR.

La otra ventaja muy valiosa de esta solución es que Internet es un sistema 'vivo' que refleja, con gran exactitud, el lenguaje actual, que es un sujeto dinámico y en evolución y puede variar con un encabezamiento de noticias singular, por lo que no está confiando en actualizar constantemente sus diccionarios de ASR con un subconjunto limitado de lenguaje natural, sino que tiene acceso a probablemente la fuente más actual y de mayor magnitud del lenguaje natural en el planeta.

Ejemplo

SpinVox convierte las señales de audio siguientes:

Mensaje de una persona inglesa

Audio: En inglés la frase *"The cat sat on Sky when Ronaldo scored against Cacá"*

Texto convertido:

The cat sat on sky when Rownowdo/Ron Al Doh/Ronaldow/Ronahldo score against Caka/Caca/Caker.

Problemas:

- *'sat on sky'* es gramáticamente incorrecto, *'you can't sit on 'sky'* en el contexto del diccionario.
- Rownowdo/Ron Al Doh/Ronaldow/Ronahldo son soluciones posibles para un nombre real inusual.
- Caka/Caca/Caker son conjeturas de un nombre real muy inusual.

Búsqueda en Google para elementos difíciles de esta clase indican:

- El cielo es un nombre de marca – puesto número 1 para 'en el cielo'. Por lo tanto, es un nombre real para un objeto, por lo que '*The cat sat on sky*' es posible desde el punto de vista gramatical y válido.
- El primer nombre es más probable Ronaldo simplemente a partir de las comprobaciones ortográficas únicas de todas las versiones (en Google "¿Quería decir: Ronaldo?").
- Ronaldo está en estrecha correlación con la 'Ronaldo scored' como siendo un jugador de fútbol muy famoso y la búsqueda proporciona un gran número de coincidencias exactas para esta frase.
- El segundo nombre es más probablemente Cacá, porque Cacá tiene más aciertos para 'score against Cacá'.
- Además, confiamos en la búsqueda de 'fútbol Cacá'- fútbol es una palabra que se deriva del contexto de 'Ronaldo scored' y obtenemos un gran número de resultados de búsqueda en muy estrecha correlación. Habida cuenta que 'calificado Ronaldo' ha obtenido ya un gran número de búsquedas satisfactorias, confiamos más en que 'Cacá' sea la solución más probable.
- Además, la naturaleza en el mundo real de la iniciación de datos de Google significa que términos que se están utilizando actualmente, términos actuales, obtención de más altas clasificaciones que los términos menos actuales, lo que es esencial para conseguir un reconocimiento de la voz para trabajar con el idioma actual y su contexto.

Gestor de colas de espera

El gestor de colas de espera es responsable de:

- Determinar qué debería suceder a un mensaje vocal en cada etapa – estrategia de conversión.
- Gestión de la decisión de cada etapa automatizada cuando requiere asistencia humana

- si en cualquier etapa de la conversión automatizada, los intervalos de confianza u otras medidas indican que cualquier parte de un mensaje no es suficientemente buena, en tal caso, el gestor de colas de espera lo dirige al agente humano correcto para su asistencia.

- Garantizar nuestro acuerdo de nivel de servicio con cualquier cliente asegurando así que convertimos cualquier mensaje dentro de un tiempo convenido – Tiempo de Ida y Vuelta (TAT)

  - normalmente TAT tiene un valor medio de 3 minutos, un 95% dentro de 10 minutos y un 98% dentro de 15 minutos.

- Toma de decisiones calculando soluciones de compromiso entre el tiempo de conversión y la calidad. Esta es una función de lo que permite el acuerdo SLA, en particular, tratar las demandas máximas de uso de lenguaje anormal o de tráfico imprevisto y el rendimiento hasta la fecha.

Lo anterior se consigue utilizando grandes máquinas de estados que, para cualquier cola de espera de lenguaje dada, puede decidir cómo procesar mejor los mensajes a través del sistema. Interacciona con todas las partes y es el núcleo operativo del SpinVox VMCS.

Aplicación de control de calidad

El Anexo II contiene una descripción más completa de este método de retícula tal como se utiliza dentro de la aplicación de control de calidad de SpinVox.

Según se ilustra en el diagrama de la Figura 2 del sistema de conversión de mensajes vocales (VMCS), los agentes humanos interaccionan con mensajes en varias etapas. Lo hacen utilizando la aplicación de control de calidad.

También utilizan una variante de esta herramienta para inspeccionar, de forma aleatoria, mensajes para garantizar que el sistema está convirtiendo correctamente los mensajes como uno de los problemas con la AI es que es incapaz de cerciorarse de que realmente es exacto.

Una etapa inventiva clave es el uso de agentes humanos para 'guiar' la conversión de cada mensaje. Esto radica en las bases de datos VMCS de SpinVox, que contienen un gran cuerpo de posibles coincidencias, ASR y una introducción humana de algunas palabras para crear soluciones de tecleado predictivas. En su caso extremo, no se requiere ningún agente humano y la conversión es completamente automática.

Cuestión

ASR es solamente aceptable en coincidencias a nivel de palabras. Para convertir un mensaje significativo, frases, sentencias y gramática para mensajería vocal resulta necesario. ASR proporcionará una medida de confianza estadística para cada coincidencia a nivel de palabras y a nivel de frases en donde esté disponible. Es incapaz de utilizar las reglas del lenguaje natural o de contexto de uso para realizar el proceso de una conversión significativa y correcta.

Qué sistemas automatizados son buenos en su ortografía y gramática base – coherencia.

Qué agentes humanos son buenos en cuanto a significado, contexto, lenguaje natural, gramática hablada, gestión de entradas ambiguas y darles sentido. Los agentes humanos tienden a ser incoherentes con la ortografía, gramática y velocidad.

Cuestión de negocios

La utilización de agentes humanos cuesta dinero, por lo que algo se podrá hacer para utilizarlos solamente para asuntos tales como esa materia y por lo tanto, el valor económico, es esencial.

SpinVox VMCS utiliza el concepto de Relación de Conversión de Agente (ACR) - la relación del tiempo que utiliza realmente el agente para procesar un mensaje con respecto a la longitud del mensaje vocal. Algo que reduce la relación ACR y mejora la calidad de conversión del mensaje es un factor comercial tal como una reducción del 1% en ACR que da lugar a por lo menos una mejora del 1% en el margen bruto. De hecho, la sensibilidad es incluso más alta cuando no solamente se reduce el coste directo de las mercaderías vendidas, sino que la sobrecarga de gestión y la disponibilidad operativa del servicio y la escalabilidad del mismo benefician a todos ellos del menor número de agentes humanos requerido.

Solución

Método de retícula: utilizar agente humanos para guiar el sistema para captar el conjunto correcto de palabras, frases, sentencias, mensajes a partir de una lista predeterminada de las opciones más probables.

Las bases de datos de SpinVox VMCS mantienen un registro histórico abundante de datos de mensajes tal como modelos estadísticos grandes (diccionarios y gramática con vectores de contexto que les relacionan) que pueden extraerse en dos formas principales:

Método de la retícula

i. El modelo de lenguaje de VMCS utiliza el contexto (p.e., historia de par de llamada, lenguaje, hora del día, etc. – véase apartado de Vectores de contexto) para captar la conversión más probable (la conversión propuesta) para mostrarla al agente.

ii. Cuando se reproduce el mensaje, el agente selecciona una letra para elegir una alternativa (puede ser simplemente las primeras letras de la palabra correcta) o introduce '*accept*' para aceptar la sección propuesta de texto y desplazarse a la sección siguiente.

iii. A medida que cambien los tipos de agentes, el sistema lo utiliza como entrada para captar la nueva conversión más probable y como realimentación (aprendizaje) de modo que, en la siguiente ocasión, sea más probable que obtenga la coincidencia correcta a la primera vez.

iv. Qué requeriría normalmente un agente para escribir un mensaje completo con los caracteres debidos (p.e., 250), solamente se necesitaría unas pocas pulsaciones de teclas para completar y en tiempo real o más rápido.

v. La salida del agente está restringida ahora a la ortografía correcta, gramática y fraseología o reglas sobre ellas que controlan la calidad y mejor significado del mensaje.

Lo anterior puede presentarse al agente en dos formas principales:

1. Conversión propuesta asistida por ASR

En este caso, ASR se utiliza primero para predecir mejor qué texto debería ser la conversión propuesta para el agente. Utiliza lo que está realmente en la señal de audio del mensaje vocal para reducir el conjunto de posibles opciones de conversión a un mínimo, con lo que se mejora la exactitud y la velocidad del agente.

a. ASR puede utilizarse para la conversión propuesta inicial.

b. ASR puede, entonces, utilizarse continuamente cuando el agente introduce selecciones para reajustar todavía más las restantes secciones de la conversión propuesta.

Técnica anterior: agentes humanos corrigen las transcripciones con elección de las alternativas de palabras.

Problema con la técnica anterior:

- Las correcciones siguen siendo consumidoras de tiempo.

- El motor de ASR podría haber tomado una mejor decisión (más tarde en una expresión oral) si la corrección del usuario hubiera sido conocida durante la decodificación.

2. Tecleado de texto completamente predictivo

De forma similar a lo descrito en el apartado 1 anterior, pero donde no se utiliza ningún ASR para seleccionar la conversión propuesta mostrada a un agente. Esto es diferente para los editores de texto predictivos estándar puesto que se confía en un registro histórico específico (modelos de idiomas VMCS y utilización de vectores de contexto – p.e., historia de pares de llamadas) y funciona al nivel de frase y superior.

Técnica anterior: predecir la palabra más frecuente (lista de alternativas) dada la entrada humana parcial.

Problema con la técnica anterior:

- La palabra más frecuente no suele ser la que desea el usuario.

- Predicciones solamente para una palabra.

En uno u otro caso, los modelos de idiomas de SpinVox VMCS están simplemente capacitados por agentes humanos o por una combinación de ASR y agentes humanos.

En el caso extremo, el sistema está completamente capacitado y es capaz de captar siempre la conversión propuesta correcta por primera vez y solamente requiere asistencia humana para la garantía de la calidad para un muestro aleatorio

y para comprobar si el VMCS se está auto-regulando de forma correcta.

Anexo II – Método de la retícula

Observaciones y supuestos diversos

1. Habida cuenta de la amplitud del vocabulario y de la calidad de audio variable, parece imposible conseguir una exactitud del reconocimiento de voz bastante alta para una conversión completamente automática para más de una pequeña fracción de expresiones orales. Detectar fiablemente esta fracción, esto es, decidir que no se necesita ninguna comprobación humana es un problema de investigación a más largo plazo, de gran interés, pero probablemente no sea una opción realista a corto plazo.

2. Aunque un buen operador tiene un ACR objetivo de 3-4, la media más probable es de 6-8.

3. La corrección de una expresión oral que sea ya un 90% correcta lleva aproximadamente 1.2 (fuente: Investigación Operativa de SpinVox 2005).

4. Un 75% del tiempo de corrección se dedica a encontrar y seleccionar errores (Wald et al).

5. Las listas de selección de palabras (alternativas) reducen el tiempo de escucha (Burke 2006).

6. Los errores tienden a agruparse (Burke 2006).

7. Una reproducción de doble velocidad mantiene la inteligibilidad y los usuarios parecen preferirla después de una corta capacitación (Arons 97).

8. La eliminación de pausas y una reproducción un 50% más rápida proporciona un factor de tiempo real de 1/3 (Arons 97).

9. Según Bain et al 2005, el tecleado normal tiene ACR 6.3 que es igual a ACR para editar la salida de ASR con un 70% de precisión. La transcripción ‘fantasma’ se menciona como ‘viable’ para una subtitulación en directo.

Enfoque

El principal objetivo tiene que ser reducir la relación de conversión de agentes (ACR) utilizando la tecnología de la voz para prestar soporte al agente. Esto puede conseguirse en varias formas:

1. Permitir al agente tomar las decisiones que no podemos proporcionar para equivocarse, esto es, el significado global del mensaje o frases individuales. La máquina puede rellenar los detalles.

2. Ofrecer predicciones mientras que el agente introduce/edita la expresión oral. Esto podría no solamente ahorrar tiempo de tecleado, sino que también ayuda a evitar errores ortográficos.

3. Proporcionar una puesta en mayúsculas (simplificada) y puntuación de forma automática, con el fin de que el agente no necesite tratar estas cuestiones.

Etapas de gestión de llamadas

1. El agente escucha el mensaje a alta velocidad (p.e., 1/2 tiempo real).

2. El agente pulsa un botón para seleccionar la categoría (p.e., “sírvase rellamar”, “solamente rellamada”, ….”general”).

3. En algunos casos, la expresión oral se acepta de inmediato. Esto sucederá si el mensaje sigue un modelo simple definido para la categoría de mensaje, la calidad de la voz era buena y no existe ninguna parte importante, pero fácilmente confundible en el mensaje (p.e., *times* en inglés).

4. El sistema propone una cadena convertida, el agente edita mientras el sistema actualiza continuamente (e instantáneamente) la expresión oral propuesta utilizando predicciones basadas en los resultados de reconocimiento de voz.

5. El agente pulsa una tecla para aceptar la expresión oral tan pronto como la expresión oral visualizada sea correcta.

En la Figura 3 se ilustra, a modo de ejemplo, la etapa de gestión de llamadas 4.

A modo de ejemplo, el agente necesitaría 35 pulsaciones de teclas para editar una expresión oral con 17 palabras y 78 caracteres:

- 15* <accept_word> (p.e., pestaña)

- 14* <accept_char> (p.e., flecha a la derecha)

- 6* entrada normal

- 1* <accept_utterance> (p.e., Introducir).

La mayoría de ellos deben ser muy rápidos porque la misma tecla tiene que pulsarse algunas veces. Solamente 6 de ellas requieren seleccionar una tecla normal.

Conviene señalar que solamente 6 de las 17 palabras (un 35%) eran correctas en la expresión oral (*utterance*) originalmente propuesta por el sistema.

Puesta en práctica

Etapas de procesamiento

1. Un motor de reconocimiento de voz (p.e., HTK) convierte el fichero de voz de expresión oral en una retícula (esto es, gráfico de hipótesis de palabras – un gráfico dirigido, un gráfico cíclico que representa un gran número de posibles secuencias de palabras).

2. La retícula es reclasificada para tener en cuenta la información específica del número de teléfono (par) (p.e., nombres, frases frecuentes en llamadas anteriores, etc.).

3. La retícula se aumenta para permitir una búsqueda muy rápida durante la fase de edición (p.e., la ruta más probable al final de la expresión oral se calcula para cada nodo y el arco que inicia esta ruta se memoriza, se añaden sub-árboles de caracteres a cada nodo que representa puntos de decisión). “Familias”, esto es, varios arcos que difieren solamente en sus tiempos de inicio y final se combinan con algunos límites.

4. Cuando el agente selecciona una categoría concreta (etapa 2 en “etapas de gestión de llamadas”), un modelo de gramática e idioma correspondiente se seleccionan para el análisis sintáctico y la recalificación dinámica. Cuando la categoría es “general”, se utiliza una “gramática” no restringida.

5. Se selecciona la ruta de más alta calificación a través de la retícula que coincide con la gramática de categoría seleccionada (si es adecuada).

6. El resultado encontrado de esta forma será aceptado de inmediato si:

   a. La categoría no es “general”.

   b. La diferencia de calificación es la más alta calificación en ruta no restringida que está dentro de un margen dado. Este margen puede utilizarse como un parámetro para controlar, de forma dinámica, la solución de compromiso entre velocidad y precisión.

   c. Según la gramática utilizada para encontrar la ruta, la expresión oral no contiene partes cruciales que sean fácilmente confundibles (p.e., *times* en inglés).

7. Cuando el usuario acepta palabras o caracteres, el sistema se desplaza a lo largo de la ruta seleccionada a través de la retícula.

8. Cuando se aceptan o introducen caracteres y palabras, cambia su color o fuente de impresión.

9. Cuando el agente teclea algo, el sistema selecciona la ruta de más alta calificación (de nuevo, teniendo en cuenta la gramática actual y posiblemente otra, p.e., información estadística) que se inicia con los caracteres tecleados. Esta nueva ruta se visualiza en este momento.

10. Cuando un agente teclea una palabra no encontrada en la retícula, se comprueba automáticamente por deletreo y se ofrece una corrección si fuera adecuado.

11. Después de que el agente pulse <*accept_utterance*>, el texto se procesa para añadir la puesta en mayúsculas y puntuación, corregir los errores ortográficos, sustituir las palabras de números por dígitos, etc. Esta operación utiliza un sólido analizador probabilístico que utiliza gramáticas semiautomáticamente derivadas de los datos de capacitación.

Reproducción de señales de audio

Los nodos en la retícula contienen información de temporización y por ello, el sistema puede mantener un registro de la parte del mensaje que está editando el agente. El agente puede configurar cuántos segundos tendrá la reproducción del sistema. Si el agente así lo desea, el sistema reproduce la expresión oral (*utterance*) a partir de la palabra antes del nodo actual.

Opciones de perfeccionamiento

Marcar las partes importantes y no importantes

i. Dependiendo de la gramática y categoría pertinente, se resaltan partes específicas del texto de expresión oral visualizada que se consideran cruciales mientras que se sombrean las partes particularmente no importantes (p.e., frases de saludos).

Utilización de clases de frases para partes no importantes

Partes del mensaje se visualizan como clases de frases en lugar de palabras individuales. El agente solamente necesita confirmar la clase mientras que la elección de la clase individual se deja al motor de ASR porque un error en este área no se considera importante. A modo de ejemplo, la clase “HEY”, podría significar "hi, hay, hey, hallo, hello” y el ejemplo anterior podría visualizarse según se ilustra en la Figura 4. En esta versión, la clave <accept_word> aplicada a una frase aceptaría la frase completa. El tecleado de un carácter cambia de nuevo al modo de palabra, p.e., el marcador de la clase de frase se sustituye por palabras individuales.

Visualización de predicción límite

La visualización de las predicciones erróneas podría confundir realmente al agente y podría valer la pena visualizar solamente aquellas (parciales) de las que el sistema tiene una certidumbre relativa. Como alternativa, la confianza relativa en varias predicciones podría ser codificada en colores en alguna manera (véase apartado de “sombreado de confianza”), p.e., algunas inciertas (normalmente alejadas del cursor) se imprimen en un color gris muy claro mientras que las más fiables se muestran más oscuras y en negrillas.

Segmentación de la expresión oral

Periodos de silencio más largos se detectan y utilizan para desglosar el mensaje en segmentos. La interfaz de usuario refleja la segmentación y se asigna una clave extra a *“<accept_segment>”*. Esto permite la confirmación de frases más largas con una sola pulsación de teclas y también la resincronización si el agente escribe una palabra que no extiende la ruta actual a través de la retícula.

Mantenimiento del cursor en el lado izquierdo de la pantalla

Tener una zona amplia, en la parte media de la pantalla, que muestra la frase actual en grandes letras. A medida que prosigue la edición, se desplaza el texto (manteniendo el cursor en la misma posición). Se muestra solamente unas pocas palabras a la izquierda del cursor. A medida que se eliminan palabras de la zona media, se desplazan a la parte superior (tipo de impresión más pequeño, gris). Más frases se visualizan debajo, de nuevo en caracteres pequeños y grises.

Mostrar alternativas de frases

Siempre o después de una pulsación de tecla, mostrar realizaciones de frases alternativas como un menú desplegable a la derecha del cursor y permitir la selección con teclas de flechas. Esto significa que el agente no necesita pensar sobre los primeros caracteres de la palabra correcta que servirían de ayuda para palabras difíciles.

Desplazar el cursor con la voz

A medida que se reproduce el mensaje, la palabra hablada se resalta automáticamente y el cursor se desplaza al principio de la palabra.

Reproducir la zona resaltada

El agente puede seleccionar una zona (p.e., botón de ratón izquierdo y derecho) y el sistema mantiene la reproducción del segmento entre los marcadores hasta que se desplace el agente.

Transcripción ‘fantasma’ para palabras o frases individuales

Las palabras son resaltadas a medida que se reproducen al agente y además de su tecleado, el agente puede simplemente decir la palabra para sustituir la actualmente resaltada (y el resto de la frase). El sistema establece, de forma dinámica, una gramática a partir de los candidatos alternativos en la retícula (palabras y frases) y utiliza ASR para

seleccionar la solución correcta. Esta es una opción técnicamente difícil porque ASR necesita utilizarse desde dentro de la aplicación de QC y los modelos dependientes de la forma de hablar adecuada necesitan capacitarse y seleccionarse en el momento de la ejecución.

Consideraciones sobre la exactitud

Reserva más probable

La exactitud del resultado de más alta calificación, después de la etapa de reconocimiento de voz, se espera que sea bastante baja (p.e., 25%). Por ello, el resultado inicialmente visualizado solamente en raras ocasiones será correcto. IBM informa una tasa de errores de palabras del 28% para el correo de voz en Padmanabhan et al 2002.

Cuando puede identificarse una categoría de la expresión oral (en conjetura: 20% de los casos), la posibilidad de obtener un resultado global correcto debería ser razonablemente alto (p.e., un 70%) si se utiliza el método de la “clase de frase”, esto es, si los errores en las frases exactas utilizadas para la parte superior y la parte inferior son aceptadas y no existe ninguna parte difícil en el mensaje o se pueden verificar utilizando otra información (propietarios de teléfonos, llamadas anteriores). Una conjetura aproximada sería que un porcentaje global de aproximadamente un 10% de expresiones orales podría gestionarse con simplemente una pulsación de tecla (la necesaria para la selección de categoría).

Corrección de errores

Se ha observado que los errores en el reconocimiento de voz tienden a producirse en los agrupamientos, p.e., el número medio de palabras subsiguientes que contienen un error es aproximadamente 2 (TODO: encontrar referencia). Esto suele deberse a:

- Errores de segmentación – la primera palabra incorrecta es más corta o más larga que la correcta y por ello, la siguiente palabra debe ser también errónea.

- La influencia del modelo de idiomas.

- Posiblemente la modelización de coarticulación.

Esta observación motiva la expectativa de que una corrección de una palabra, durante el proceso de edición, corregirá normalmente más de un error en el supuesto de la expresión oral.

En términos muy amplios, el teclado de un carácter limita el número de contendientes para la siguiente palabra en un factor de 1/26. Dos caracteres lo limiten a 1/676 y casi con certeza, debe excluir todas las palabras incorrectas de más alta calificación. Esto motiva otra predicción: un error de ASR debería como media, requerir no más de una pulsación de tecla para su corrección.

Mejor ruta a través de la retícula

Un factor muy importante para el éxito operativo del sistema es el porcentaje de retículas que contienen la ruta correcta aún cuando tenga una calificación comparativamente baja. Si la ruta correcta no está en la retícula, el sistema no será capaz, en algún punto, de seguir una ruta a través de la retícula y por ello, será difícil obtener nuevas predicciones. El sistema podría necesitar esperar a que el agente teclee dos o tres palabras antes de encontrar los puntos adecuados en la retícula restante de nuevo para crear más predicciones.
El tamaño de la retícula y por lo tanto, la posibilidad de obtener la expresión oral correcta, puede controlarse mediante parámetros (número de los denominados *tokens y pruning*) y en teoría, podría incluirse el espacio de búsqueda total. Esto generaría grandes retículas, sin embargo, que no podrían transmitirse al cliente dentro de un marco de tiempo aceptable. Además, tenemos que tratar la ocurrencia ocasional de palabras previamente no vistas que, en consecuencia, no estarían en el vocabulario. Después de unos pocos meses de operación (y por lo tanto, de recogida de datos) parece conseguible una tasa de aproximadamente un 95%.

Si se utiliza la versión de “segmentación de expresión oral” anteriormente descrita, los segmentos proporcionarían puntos fáciles para reiniciar la predicción.

Pos-procesamiento lingüístico

Podría merecer la pena definir una sintaxis de “Mensajes de SpinVox” para cada idioma. Los servicios de mensajes cortos SMSs no suele estar previsto que contengan sentencias adecuadas completas y el intento de añadir mucha puntuación (y frecuentemente hacerlo de forma errónea) podría, en cambio, valer la pena para su uso en raras ocasiones, pero de forma coherente.

Puesta en mayúsculas iniciales

Esta operación es comparativamente fácil en inglés pero más difícil en otros idiomas (p.e., alemán).

Ventajas previstas

1. Al realizar la conversión o edición, el sistema mantiene un registro de dónde, en la expresión oral, el agente está actualmente y por ello, la reproducción de audio se puede controlar mejor.

2. Para una determinada proporción de expresiones orales, en donde el agente solamente necesita determinar la categoría, el ACR podría ser menor que uno (teóricamente 1/3 con reproducción rápida y eliminación de silencio).

3. Un número significativo de mensajes, para los que el rendimiento de ASR es alto, requerirá solamente una comprobación rápida y muy pocas pulsaciones de teclas para realizar correcciones, proporcionando un ACR de aproximadamente 2.

4. La mayoría de los mensajes necesitarán todavía una edición importante. En qué medida estos casos serán ventajosos a partir de las predicciones todavía ha de determinarse.

5. La gestión de la puesta en mayúsculas iniciales y la puntuación deben reducir automáticamente el ACR en un pequeño porcentaje y también mejorar la coherencia.

Cuestiones/ediciones

1. ¿Cuándo la predicción con ediciones controladas por ASR se hacen más consumidoras de tiempo que un tecleado simple? Para sacar partido de las predicciones, el agente necesita leerlas. Si las siguientes palabras se predicen correctamente, su simple aceptación debe ser más rápida que teclearlas pero si la siguiente palabra es errónea, el tiempo adicional requerido para su comprobación simplemente se desperdicia. Por el contrario, el agente necesita escuchar en alguna manera y podría utilizar también el tiempo para comprobar las predicciones.

Combinación de métodos de predicción

Parece prometedor utilizar las predicciones estadísticas como un respaldo operativo para la predicción basada en ASR. Puesto que el modelo de predicción estadística es estático (no dependiente de las llamadas) y puesto que no necesita transmitirse a la aplicación de QC con cada mensaje, puede ser bastante global. Las retículas han de transmitirse para cada mensaje y por lo tanto, no tienen que mantenerse dentro de determinados límites de tamaño y es probable que falten algunas de las hipótesis necesarias.

Los modelos estadísticos y los modelos de predicción basados en ASR se representarían como gráficos y la tarea de combinar las predicciones implicaría atravesar ambos gráficos por separado y luego, elegir la predicción más fiable o combinarlas en función de alguna fórmula de interpolación.

Este método podría extenderse a más gráficos de modelos de predicción, a modo de ejemplo, basados en pares de llamadas.

Predicciones estadísticas

Estas predicciones están basadas en modelos de idiomas de n-gramas. Estos modelos memorizan las probabilidades condicionales de secuencias de palabras. A modo de ejemplo, un modelo de 4-gramas podría memorizar la probabilidad de la palabra “to” después del contexto de tres palabras “I am going”, inglés. Estos modelos pueden ser muy grandes y formas eficientes de memorizarlo se requiere que también permitan una rápida generación de predicciones.

Puesta en práctica

Los modelos de n-gramas suelen memorizarse en una estructura de gráficos en donde cada nodo representa un contexto (palabras ya transcritas) y cada enlace saliente se anota con una palabra y la correspondiente probabilidad condicional.

Puesto que habrá siempre palabras que nunca fueron encontradas (o en muy raras ocasiones) después de un determinado contexto, pero que son todavía requeridas en el tiempo de ejecución, el modelo necesita una manera de tratar las palabras previamente no vistas en un contexto dado. Esto se consigue con una “copia de reserva” para el contexto más corto correspondiente. En nuestro ejemplo, si la palabra “to” no hubiera sido observada después de “I am going” en inglés, el modelo buscaría “to” después de “am going”. Si no fue observada nunca, buscaría el nodo de contexto para “going” y por último, en el nodo de contexto vacío en donde se representan todas las palabras en el vocabulario. Esta “copia de seguridad” se pone en práctica añadiendo un enlace especial a cada nodo de contexto que apunta al nodo con el correspondiente contexto más corto y se anota con “back-off penalty” que puede interpretarse como el (logaritmo de) la masa de probabilidad no distribuida a través de todos los demás enlaces que salen del nodo.

La probabilidad logarítmica global de “to” después de “I am going” podría, a modo de ejemplo, calcularse como back_off (“I am going”) + back_off (“am going”) + link_prob (“to” @ nodo de contexto “going”).

Expansión de gráficos de palabras

Sería de alto coste desde el punto de vista del cálculo, buscar la palabra (enlace) más probable que comience con una secuencia de caracteres dada cada vez que el usuario pulsa una tecla. Una forma de acelerar esta operación se basa en ampliar el gráfico de palabras en un gráfico de caracteres en donde los enlaces salientes, en cada nodo, se clasifiquen por probabilidad decreciente. Conviene señalar que el número máximo de enlaces salientes, en cada nodo, es el número de caracteres en el idioma más dos para el enlace de reserva de seguridad y el enlace de la palabra final. Puesto que la búsqueda a través de esta lista requeriría, como máximo, 100 comparaciones de caracteres para el idioma inglés con el coste previsto menor que aproximadamente 50 comparaciones teniendo en cuenta que las palabras más probables se intentarían en primer lugar.

Lo anterior ignora el coste de búsqueda de palabras no encontradas en el nodo de contexto actual. Cuando esta circunstancia se requiere, podría ser mejor aceptar que las predicciones no pueden generarse con gran rapidez y utilizar el enlace de seguridad normal para la búsqueda en los nodos de contexto de reserva de seguridad. La alternativa de memorizar enlaces de reserva en cada nodo de caracteres requeriría demasiado espacio de memoria.

La expansión desde el gráfico de palabras al gráfico de caracteres puede ponerse en práctica en la forma siguiente:

1. Para cada nodo de contexto (nivel de palabra):
    - clasificar todos los enlaces salientes por su probabilidad (decreciente).
    - Para cada enlace (en orden):
        - establecer el puntero para el nodo actual (palabra)
        - para cada carácter:
            - si es ya un enlace anotado con este carácter, establecer el puntero al nodo al que apunta el enlace
            - en cualquier otro caso: añadir un nuevo enlace al nodo al que apunta el puntero y crear un nuevo nodo como destino para el enlace. Poner el puntero hacia este nuevo nodo.
        - Añadir un nuevo enlace para el puntero objetivo, apuntando al destino del enlace de palabra actual.

Después de la expansión, todos los enlaces de palabras (incluyendo sus probabilidades) se pueden suprimir, exceptuados para los enlaces de reserva de seguridad. Conviene señalar que esto permitirá encontrar siempre la predicción de frase más probable, pero no la lista de las menos probables. Si esta última se requiere (más adelante), la secuencia en la que se realizó la expansión de caracteres tendría que memorizarse en alguna manera.

Predicción

Tomando un identificador de nodo de palabra, un identificador de nodo de carácter, la subcadena de palabras actual y un carácter como entrada, el método de “predicción” sería:

1. Ir al nodo de carácter [character_node_id]
2. Encontrar el enlace anotado con el carácter de entrada (utilizar la búsqueda lineal en enlaces clasificados por probabilidad).
3. Si se encuentra, seguir el enlace y comenzando desde su nodo objetivo, seguir el primer enlace dejando cada nodo hasta que se alcance alguna condición de parada. En cada transición, añadir el carácter encontrado en el enlace a la cadena de resultado. Reenviar el identificador del nodo del nodo objetivo inicial y la cadena resultado.
4. En otros casos: utilizar el enlace de reserva de seguridad desde los nodos de palabras [word_node_id] y la cadena de palabras actual para encontrar predicciones en los nodos de reserva de seguridad. Esta operación no está prevista para encontrar predicciones en tiempo real si el usuario teclea con rapidez.

Anexo III

Conceptos básicos

Los siguientes conceptos están cubiertos. Cada concepto básico A – I puede combinarse con cualquier otro concepto básico en una puesta en práctica.

El siguiente texto describe también varios subsistemas que, *inter alia* ponen en práctica características de los conceptos básicos. Estos subsistemas no necesitan estar separados entre sí. A modo de ejemplo, un subsistema puede ser parte de otro de los subsistemas. Tampoco los subsistemas tienen que ser discretos en alguna otra manera, las funciones de puesta en práctica de códigos de un subsistema pueden formar parte del mismo programa informático que las funciones de puesta en práctica de códigos de otro subsistema.

Concepto básico A

Un sistema de mensajería vocal, independiente del usuario e independiente del dispositivo, a gran escala que convierte mensajes vocales no estructurados en texto para su presentación visual en una pantalla; el sistema comprende (i) subsistemas puestos en práctica por ordenador y también (ii) una conexión de red para operadores humanos que proporcionen la transcripción y el control de la calidad; estando el sistema adaptado para optimizar la eficacia de los operadores humanos incluyendo, además:

3 subsistemas básicos, a saber (i) un extremo frontal de pre-procesamiento que determina una estrategia de conversión adecuada; (ii) uno o más recursos de conversión y (iii) un subsistema de control de calidad.

Otras características:

- Los recursos de conversión incluyen uno o más de lo siguiente: uno o más motores de ASR; recursos de procesamiento de señal; los operadores humanos.
    - Los recursos de procesamiento de la señal optimizan la calidad de la señal de audio para conversión realizando una o más de las funciones siguientes: eliminación del ruido, depuración de defectos conocidos, normalización de energía de señal/volumen, eliminación de secciones vacías/en silencio.
- Los operadores humanos realizan pruebas de garantía de la calidad aleatorias sobre mensajes convertidos y proporcionan una realimentación informativa al extremo frontal de pre-procesamiento y/o los recursos de conversión.

Concepto básico B

Vectores de contexto

Un sistema de mensajería vocal, independiente del usuario e independiente del dispositivo, a gran escala, que convierte mensajes vocales no estructurados en texto para su presentación visual en una pantalla; el sistema que comprende (i) subsistemas puestos en práctica por ordenador y también (ii) una conexión de red para operadores humanos que proporcionan transcripción y control de la calidad; estando el sistema adaptado para optimizar la eficacia de los operadores humanos incluyendo, además:

un subsistema de contexto puesto en práctica por el ordenador adaptado para utilizar información sobre el contexto de un mensaje o una parte de un mensaje para mejorar la exactitud de la conversión.

Otras características:

- la información de contexto se utiliza para limitar el vocabulario utilizado en cualquier motor de ASR o para perfeccionar los procesos de búsqueda o de coincidencia utilizados por el motor de ASR.
- La información de contexto se utiliza para seleccionar un recurso de conversión particular una combinación de recursos de conversión, tales como un motor de ASR particular.
- La información del contexto incluye uno o más de los identificadores ID del usuario que llama, identificador ID del destinatario, si el usuario que llama o el destinatario es una empresa u otro tipo de entidad clasificable, o no, idioma específico del usuario que llama, registro histórico de pares de llamadas; hora de la llamada; día de la llamada; geo-referencia u otros datos de localización del usuario que llamada o del usuario llamado; datos de PIM (datos de gestión de información personal, incluyendo agenda de direcciones, diario) del usuario que llama o del usuario llamado; el tipo de mensajes, incluyendo si el mensaje es un correo de voz, texto hablado, un mensaje instantáneo, una entrada del denominado *blog*, un correo electrónico, un memorándum o una nota; longitud del mensaje; información descubrible utilizando un cuerpo de conocimiento en línea; datos de presencia; densidad de la voz del mensaje; calidad de la voz del mensaje.
- El subsistema de contexto incluye un subsistema de confianza del reconocedor que determina automáticamente el nivel de confianza asociado con una conversión de un mensaje específico o parte de un mensaje, utilizando la información de contexto.

- El subsistema de contexto incluye o está conectado a un subsistema de confianza del reconocedor que determina llamante el nivel de confianza asociado con una conversión de un mensaje específico, o parte de un mensaje, utilizando la salida de uno o más motores de ASR.

    - El subsistema de confianza del reconocedor puede ponderar, de forma dinámica, cómo utiliza la salida de diferentes motores de ASR dependiendo de su precisión o eficacia probable.

- El conocimiento del contexto de un mensaje se extrae por un subsistema y se envía, en sentido directo, a un subsistema de flujo abajo que utiliza esa información de contexto para mejorar el rendimiento de la conversión.

    - Un subsistema de flujo abajo es un subsistema de control y de garantía y/o supervisión de la calidad.

Concepto básico C

Registro histórico de pares de llamadas

Un sistema de mensajería vocal, independiente del usuario e independiente del dispositivo, a gran escala, que convierte mensajes vocales no estructurados en texto para su presentación visual en una pantalla; el sistema que comprende (i) subsistemas puestos en práctica por ordenador y también (ii) una conexión de red para operadores humanos que proporcionan transcripción y control de la calidad; estando el sistema adaptado para optimizar la eficacia de los operadores humanos incluyendo, además:

un subsistema de pares de llamadas puesto en práctica por ordenador, adaptado para utilizar información de registro histórico de pares de llamadas para mejorar la precisión de la conversión.

Otras características:

- El registro histórico de pares de llamadas permite al sistema ser independiente del usuario pero para adquirir en el transcurso del tiempo, sin capacitación del usuario explícita, datos dependientes del usuario que permiten mejorar el rendimiento de la conversión.

- El registro histórico de pares de llamadas está asociado con un par de números, incluyendo los números asociados con teléfonos móviles, teléfonos fijos, direcciones IP, direcciones de correo electrónico o direcciones únicas proporcionadas por una red.

- El registro histórico de pares de llamadas incluye información relacionada con uno o más de: idioma o dialecto probablemente utilizado; país llamado desde o llamado a; zonas temporales; hora de la llamada; día de la llamada; frases específicas utilizadas; lenguaje específico del usuario que llama; entonación; datos de PIM (datos de gestión de información personal, incluyendo agenda de direcciones, diario).

- Un subsistema de modelo de idioma dinámico puesto en práctica por ordenador, adaptado para construir un modelo de lenguaje dinámico utilizando uno o más de: dependencia del usuario que llama; dependencia de la llamada del par; dependencia del usuario llamado.

    - el usuario que llama es cualquier que deposita un mensaje vocal, haciendo caso omiso de que intente realizar una llamada vocal y el usuario llamado es cualquiera que lee el mensaje convertido, haciendo caso omiso de si tenían previsto recibir una llamada vocal.

- Un subsistema de perfil personal puesto en práctica por ordenador adaptado para construir un perfil personal de un usuario que llama para mejorar la precisión de la conversión.

    - el perfil personal incluye palabras, frases, gramática o entonación del usuario que llama.

Concepto básico D

Taxonomía de mensajes de 3 partes

Un sistema de mensajería vocal, independiente del usuario e independiente del dispositivo, a gran escala, que convierte mensajes vocales no estructurados en texto para su presentación visual en una pantalla; el sistema que comprende (i) subsistemas puestos en práctica por ordenador y también (ii) una conexión de red para operadores humanos que proporcionan transcripción y control de la calidad; estando el sistema adaptado para optimizar la eficacia de los operadores humanos incluyendo, además:

un subsistema de selección de límites puesto en práctica por ordenador, adaptado para procesar un mensaje buscando los límites entre secciones del mensaje que transmiten diferentes tipos de contenido o transmiten diferentes tipos de mensaje.

Otras características:

- El subsistema de selección de límites puesto en práctica por ordenador analiza una o más de las partes componentes siguientes: una parte de saludo; una parte de cuerpo; una parte de despedida.

    - Diferentes estrategias de conversión se aplican a cada parte, siendo la estrategia aplicada óptima para la conversión de esa parte.

    - Diferentes partes del mensaje tienen diferentes exigencias de calidad y un subsistema de evaluación de la calidad aplica diferentes niveles estándar a las diferentes partes.

    - Un estimador de la calidad de la voz detecta límites entre secciones del mensaje que transmiten diferentes tipos de contenido o transmiten diferentes tipos de mensaje.

        - Se detectan o infieren límites en zonas en el mensaje en donde modifica la densidad de la voz.

        - Se detectan límites o se infieren en una pausa en el mensaje.

        - Se infieren límites como surgiendo en una proporción predefinida del mensaje.

            - Se infiere un límite de saludos en aproximadamente un 15% de la longitud del mensaje completa.

Concepto básico E

Extremo frontal de pre-procesamiento

Un sistema de mensajería vocal, independiente del usuario e independiente del dispositivo, a gran escala, que convierte mensajes vocales no estructurados en texto para su presentación visual en una pantalla; el sistema que comprende (i) subsistemas puestos en práctica por ordenador y también (ii) una conexión de red para operadores humanos que proporcionan transcripción y control de la calidad; estando el sistema adaptado para optimizar la eficacia de los operadores humanos incluyendo, además:

un subsistema de extremo frontal de pre-procesamiento puesto en práctica por ordenador que determina una estrategia de conversión adecuada utilizada para convertir los mensajes vocales.

Otras características:

- El extremo frontal de pre-procesamiento optimiza la calidad de la señal de audio para conversión realizando una o más de las funciones siguientes: eliminación de ruido, depuración de defectos conocidos, normalización de energía de la señal/volumen, eliminación de secciones vacías/ en silencio y clasifica el tipo de mensaje para un encaminamiento óptimo del mensaje para su conversión, o no.

- El extremo frontal de pre-procesamiento determina el idioma que se está utilizando por el usuario que llama, en función de uno o más de lo siguiente: conocimiento sobre registro, localización y registro histórico de llamadas del usuario que llama y/o del receptor.

- El extremo frontal de pre-procesamiento selecciona un motor de ASR particular para convertir un mensaje o parte de un mensaje.

    - Diferentes recursos de conversión, tales como motores de ASR, se utilizan para diferentes mensajes.

    - Los operadores humanos son tratados como motores de ASR.

    - El extremo frontal de pre-procesamiento utiliza o está conectado con un subsistema de confianza del reconocedor para determinar automáticamente el nivel de confianza asociado con una conversión de un mensaje específico, o parte de un mensaje, y un recurso de conversión particular, tal como un motor de ASR, se desarrolla luego dependiente de ese nivel de confianza.

- La estrategia de conversión implica la selección de una estrategia de conversión a partir de un conjunto de estrategias de conversión que incluyen lo siguiente: (i) mensajes para los que se comprueba automáticamente una confianza de conversión de ASR que es suficientemente alta mediante un subsistema de evaluación de la calidad para su conformidad con las normas de calidad; (ii) mensajes para los que la confianza en la conversión de ASR no es suficientemente alta se encaminan a un operador humano para su comprobación y, si fuera necesario, su corrección; (iii) mensajes para los que la confianza en la conversión de ASR es muy baja se indican como no convertibles y el usuario es informado de la recepción de un mensaje no convertible.

Concepto básico F

Gestor de colas de espera

Un sistema de mensajería vocal, independiente del usuario e independiente del dispositivo, a gran escala, que convierte mensajes vocales no estructurados en texto para su presentación visual en una pantalla; el sistema que comprende (i) subsistemas puestos en práctica por ordenador y también (ii) una conexión de red para operadores humanos que proporcionan transcripción y control de la calidad; estando el sistema adaptado para optimizar la eficacia de los operadores humanos incluyendo, además:

un subsistema de gestor de colas de espera puesto en práctica por ordenador que gestiona, de forma inteligente, la carga y las llamadas en recursos que se requieren para garantizar que los tiempos de entrega de mensajes convertidos cumplan un estándar predefinido.

Otras características:

- El subsistema del gestor de colas de espera determina lo que debería suceder a un mensaje vocal en cada etapa de procesamiento a través del sistema.
    - Si en cualquier etapa de la conversión automatizada, los intervalos de confianza u otros mensajes indican que cualquier parte de un mensaje no es suficientemente buena, en tal caso, el gestor de colas de espera lo dirige al operador humano correcto para su asistencia.
- El subsistema del gestor de colas de espera toma decisiones calculando las soluciones de compromiso entre el tiempo de conversión y la calidad.
- El subsistema del gestor de colas de espera utiliza máquinas de estados que, para cualquier cola de espera de idioma dada, puede decidir cómo procesar mejor los mensajes a través del sistema.

Concepto básico G

Retícula

Un sistema de mensajería vocal, independiente del usuario e independiente del dispositivo, a gran escala, que convierte mensajes vocales no estructurados en texto para su presentación visual en una pantalla; el sistema que comprende (i) subsistemas puestos en práctica por ordenador y también (ii) una conexión de red para operadores humanos que proporcionan transcripción y control de la calidad; estando el sistema adaptado para optimizar la eficacia de los operadores humanos incluyendo, además:

un subsistema de retícula puesto en práctica por ordenador que genera una retícula de posibles secuencias de palabras o frases y permite a un operador humano guiar un subsistema de conversión mostrando una o más palabras o frases convertidas candidatos desde las retícula y permitiendo al operador seleccionar esa palabra o frase candidato o, introducir uno o más caracteres para una palabra o frase convertida diferente, para iniciar operativamente el subsistema de conversión para proponer una palabra o frase alternativa.

Otras características:

- El subsistema de conversión recibe entradas desde un subsistema que gestiona la información del registro histórico de pares de llamadas.
- El subsistema de conversión recibe entradas desde recursos de conversión.
- El subsistema de conversión recibe entradas desde un subsistema de contexto que tiene conocimiento del contexto de un mensaje.
- El subsistema de conversión aprende, a partir de las entradas de operadores humanos, las palabras que probablemente corresponden a un modelo sonoro.
- El operador humano está obligado a gestionar solamente una tecla única para aceptar una palabra o frase.
- El subsistema de conversión proporciona automáticamente las mayúsculas iniciales y la puntuación.
- El subsistema de conversión puede proponer números candidatos, nombres reales, direcciones de la web, direcciones de correo electrónico, direcciones físicas, información de localización u otras coordenadas.

- El subsistema de conversión diferencia automáticamente entre partes del mensaje que probablemente sean importantes y las que probablemente sean no importantes.

- Las partes no importantes del mensaje se confirman por el operador como pertenecientes a una clase propuesta por el subsistema de conversión y luego son convertidas exclusivamente por un motor de ASR de máquina.

- El operador humano puede hablar la palabra correcta al sistema de conversión, que luego la transcribe automáticamente.

Concepto básico H

Cuerpo en línea

Un sistema de mensajería vocal, independiente del usuario e independiente del dispositivo, a gran escala, que convierte mensajes vocales no estructurados en texto para su presentación visual en una pantalla; el sistema que comprende (i) subsistemas puestos en práctica por ordenador y también (ii) una conexión de red para operadores humanos que proporcionan transcripción y control de la calidad; estando el sistema adaptado para optimizar la eficacia de los operadores humanos incluyendo, además:

un subsistema de búsqueda puesto en práctica por ordenador que analiza un mensaje convertido con respecto a un cuerpo de conocimiento en línea.

Otras características:

- El cuerpo de conocimiento en línea es Internet, según se accede por un motor de búsqueda.

- El cuerpo de conocimiento en línea es una base de datos de motor de búsqueda, tal como Google.

- El análisis del mensaje convertido permite que la exactitud de la conversión sea evaluada por un operador humano y/o un subsistema de confianza del reconocedor.

- El análisis del mensaje convertido permite la resolución de ambigüedades en el mensaje por un operador humano y/o un motor de ASR.

Concepto básico I

Detectores

Un sistema de mensajería vocal, independiente del usuario e independiente del dispositivo, a gran escala, que convierte mensajes vocales no estructurados en texto para su presentación visual en una pantalla; el sistema que comprende (i) subsistemas puestos en práctica por ordenador y también (ii) una conexión de red para operadores humanos que proporcionan transcripción y control de la calidad; estando el sistema adaptado para optimizar la eficacia de los operadores humanos incluyendo, además:
un subsistema detector puesto en práctica por ordenador que está adaptado para detectar operaciones de descolgar.

Otras características:

- El detector de operaciones de descolgar se pone en práctica como parte de un extremo frontal de pre-procesamiento.

Otros detectores que pueden utilizarse también:

- Un subsistema de detector puesto en práctica por ordenador que se ajusta para detectar diferentes lenguajes hablados tales como inglés, español, francés, etc.

    - El detector de idiomas puede detectar cambios en la parte de idiomas a través de un mensaje.

    - El detector de idiomas puede utilizar entradas desde un subsistema que tiene información del registro histórico de pares de llamadas que registran cómo ocurrieron los cambios en el idioma en mensajes anteriores.

- Un subsistema detector puesto en práctica por ordenador que está adaptado para estimar la calidad de la voz.

    - El estimador de la calidad de la voz encuentra desvanecimientos de la voz, estima los niveles de ruido y calcula una medida global de la calidad de la voz y utiliza un umbral adaptativo para rechazar los mensajes de más baja calidad.

- Un subsistema detector puesto en práctica por ordenador que está adaptado para detectar operaciones de descolgar.

    - El detector de operaciones de descolgar se pone en práctica como parte de un extremo frontal de pre-procesamiento.

- Un subsistema detector puesto en práctica por ordenador que está adaptado para detectar llamadas inadvertidas.

    - El detector de llamadas inadvertidas se pone en práctica como parte de un extremo frontal de pre-procesamiento.

- Un subsistema detector puesto en práctica por ordenador que está adaptado para detectar y convertir mensajes pre-registrados.

- Un subsistema detector puesto en práctica por ordenador que está adaptado para detectar y convertir números hablados.

- Un subsistema detector puesto en práctica por ordenador que está adaptado para detectar y convertir direcciones habladas.

- Un subsistema detector puesto en práctica por ordenador que está adaptado para detectar y convertir nombres reales, números, direcciones de la web, direcciones de correo electrónico, direcciones físicas, información de localización y otras coordenadas.

Tipos de mensajes

- El mensaje es un correo de voz previsto para un teléfono móvil y el mensaje vocal se convierte en texto y se envía a ese teléfono móvil.

- El mensaje es un mensaje vocal previsto para un servicio de mensajería instantánea y el mensaje vocal se convierte a texto y se envía a un servicio de mensajería instantánea para una presentación visual en una pantalla.

- El mensaje es un mensaje vocal previsto para un blog de la web y el mensaje vocal se convierte a texto y se envía a un servidor para su presentación visual como parte del blog de la web.

- El mensaje es un mensaje vocal previsto para convertirse a formato de texto y enviarse como un mensaje de texto.

- El mensaje es un mensaje vocal previsto para convertirse a formato de texto y enviarse como un mensaje de correo electrónico.

- El mensaje es un mensaje vocal previsto para convertirse a formato de texto y enviarse como una nota o memorándum, por correo electrónico o texto, a un emisor del mensaje.

Otros elementos de la cadena de valor

- Una red de telefonía móvil que está conectada al sistema de cualquier reivindicación precedente.

- Un teléfono móvil cuando visualiza un mensaje convertido por el sistema de cualquier reivindicación precedente.

- Una pantalla de presentación visual de ordenador cuando visualiza un mensaje convertido por el sistema de cualquier reivindicación precedente.

- Un método para proporcionar mensajería vocal, que comprende la etapa de un usuario que envía un mensaje vocal a un sistema de mensajería según se establece en cualquier reivindicación precedente.

REIVINDICACIONES

1. Un sistema de mensajería vocal a gran escala, independiente del usuario e independiente del dispositivo, que permite convertir un mensaje vocal no estructurado en texto para una presentación visual en una pantalla; caracterizado porque el sistema comprende (i) subsistemas puestos en práctica por ordenador, así como (ii) una conexión de red para proporcionar una transcripción y un control de calidad a operadores humanos; estando el sistema adaptado para optimizar la eficacia de los operadores humanos comprendiendo, además:

un subsistema de retícula puesta en práctica por ordenador para generar una retícula de posibles secuencias de frases o palabras y para permitir a un operador humano guiar un subsistema de conversión presentando una o más palabras o frases convertidas candidatas a partir de la retícula y para permitir al operador seleccionar la palabra o la frase candidata o bien, introduciendo uno o varios caracteres para una palabra convertida diferente, iniciar operativamente el subsistema de conversión para proponer una palabra o frase alternativa.

2. El sistema según la reivindicación 1, en donde el subsistema de retícula está configurado para recibir entradas procedentes de un subsistema que gestiona la información de registro histórico de llamada del par.

3. El sistema según la reivindicación 1, en donde el subsistema de retícula está configurado para recibir entradas procedentes de recursos de conversión.

4. El sistema según la reivindicación 1, en donde el subsistema de retícula está configurado para recibir entradas procedentes de un subsistema de contexto que tiene conocimiento del contexto de un mensaje.

5. El sistema según la reivindicación 1, en donde el subsistema de retícula está configurado para aprender, a partir de las entradas del operador humano, palabras o frases probables que corresponden a un modelo sonoro.

6. El sistema según la reivindicación 1, en donde el operador humano debe seleccionar solamente una sola tecla para aceptar una palabra o una frase.

7. El sistema según la reivindicación 1, en donde el subsistema de retícula está configurado para proporcionar automáticamente mayúsculas iniciales y signos de puntuación.

8. El sistema según la reivindicación 1, en donde el subsistema de retícula está configurado para proponer números, nombres reales, direcciones web, direcciones de correo electrónico, direcciones físicas, información de localización u otras coordenadas candidatas.

9. El sistema según la reivindicación 1, en donde el subsistema de retícula está configurado para realizar automáticamente la distinción entre las partes del mensaje que son susceptibles de ser importantes y las que son susceptibles de no tener importancia.

10. El sistema según la reivindicación 1, en donde las partes sin importancia del mensaje son confirmadas por el operador, como perteneciente a una clase propuesta por el subsistema de retícula y a continuación, se convierten únicamente por un motor de reconocimiento vocal ASR del aparato (1).

11. El sistema según la reivindicación 1, en donde el operador humano puede pronunciar la palabra correcta en el destino del sistema de conversión, que está configurado para su transcripción automática más adelante.

12. El sistema según la reivindicación 3, en donde los recursos de conversión analizan una palabra o una frase convertida con respecto a un cuerpo de conocimiento en línea.

13. El sistema según la reivindicación 12, en donde el cuerpo de conocimiento en línea es Internet, accesible por un motor de búsqueda.

14. El sistema según la reivindicación 13, en donde el cuerpo de conocimiento en línea es una base de datos de motor de búsqueda.

15. El sistema según una cualquiera de las reivindicaciones precedentes, en donde el mensaje es uno de los mensajes siguientes:

(a) un correo de voz destinado a un teléfono móvil y el sistema está configurado para convertir el mensaje vocal en texto y enviar el mensaje vocal a ese teléfono móvil o

(b) un mensaje vocal destinado a un servicio de mensajería instantánea y el sistema está configurado para convertir el mensaje vocal en texto y enviar el mensaje vocal a un servicio de mensajería instantánea para su presentación visual en una pantalla o

(c) un mensaje vocal destinado a un servicio web y el sistema está configurado para convertir el mensaje vocal en texto y enviar el mensaje vocal a un servidor para una presentación visual como parte del servicio web.

16. El sistema según cualquier reivindicación precedente, en donde el mensaje es uno de los mensajes siguientes:

(a) un mensaje vocal destinado a convertirse al formato de texto y enviarse bajo la forma de mensaje de texto o

(b) un mensaje vocal destinado a convertirse al formato de texto y enviarse en tanto como mensaje de correo electrónico o

(c) un mensaje vocal destinado a convertirse al formato de texto y enviarse bajo la forma de una nota o de un memorándum, por correo electrónico o texto, a un expedidor del mensaje.

17. Un método que permite proporcionar un sistema de mensajería vocal a gran escala, independiente del usuario e independiente del dispositivo, que convierte un mensaje vocal no estructurado en texto para una presentación visual en una pantalla, caracterizado por cuanto que el sistema comprende: (i) subsistemas puestos en práctica por ordenador así como (ii) una conexión de red para proporcionar una transcripción y un control de calidad a operadores humanos; optimizando el método la eficacia de los operadores humanos comprendiendo las etapas de:

un subsistema de retícula puesto en práctica por ordenador, que genera una retícula de posibles secuencias de palabras o frases y que permite a un operador humano guiar un subsistema de conversión presentándole una o más palabras o frases convertidas candidatas a partir de la retícula y que permite al operador seleccionar la palabra o la frase candidata o, introduciendo uno o varios caracteres para una palabra o una frase convertida diferente, para iniciar operativamente el subsistema de conversión para proponer una palabra o frase alternativa.

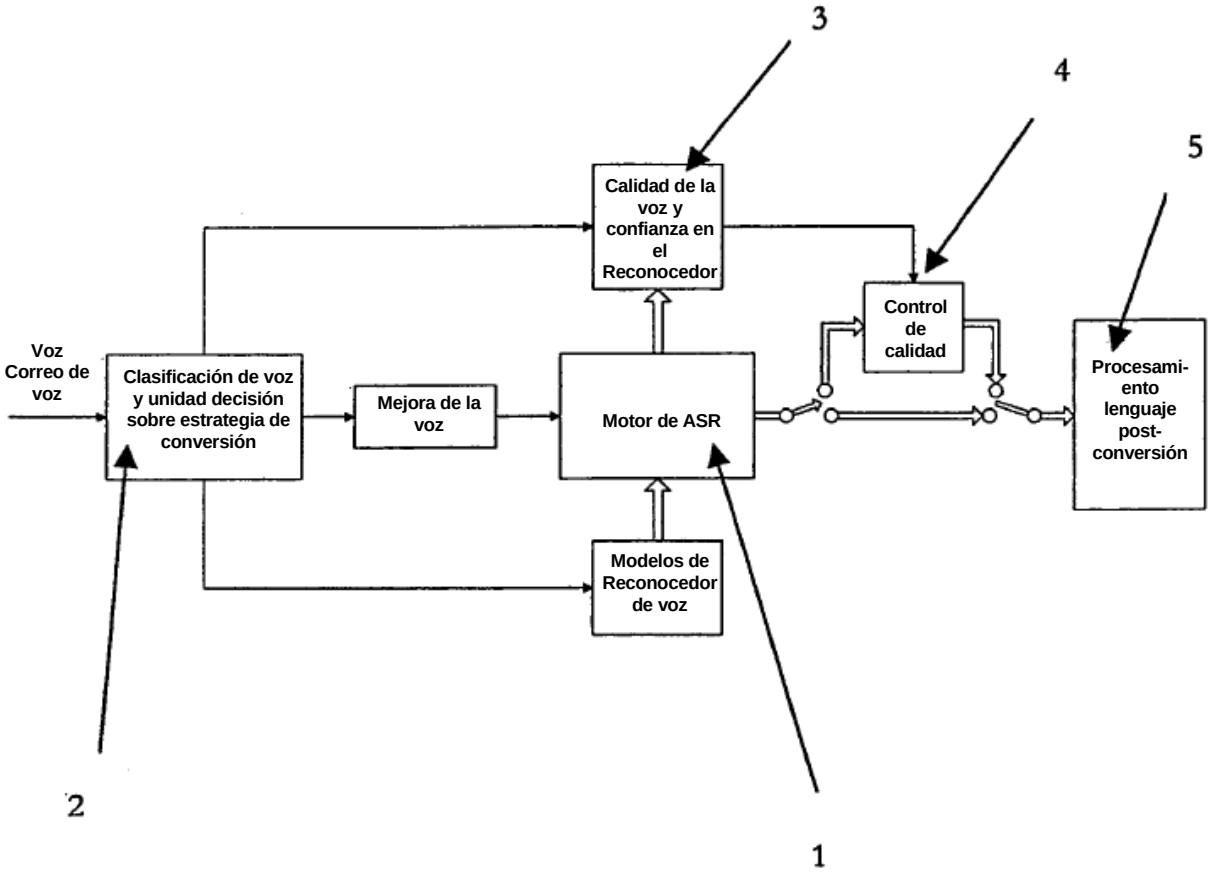
3
4
5
Calidad de la voz y confianza en el Reconocedor
Control de calidad
Procesami-ento lenguaje post-conversión
Voz Correo de voz
Clasificación de voz y unidad decisión sobre estrategia de conversión
Mejora de la voz
Motor de ASR
Modelos de Reconocedor de voz
2
1

Figura 1

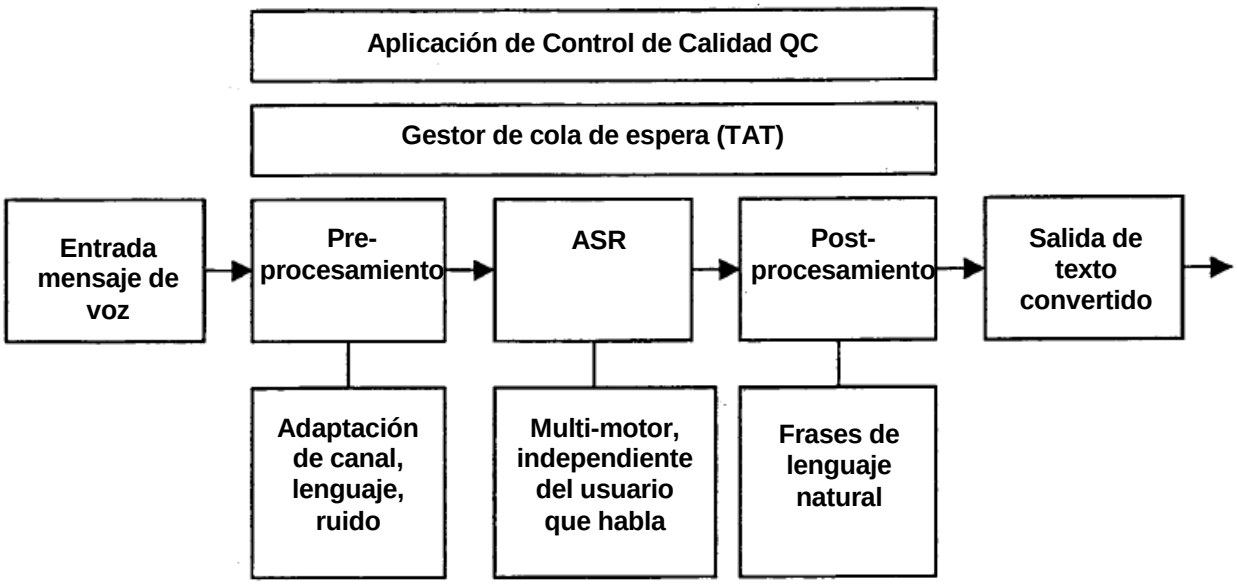
Aplicación de Control de Calidad QC
Gestor de cola de espera (TAT)
Entrada mensaje de voz
Pre-procesamiento
ASR
Post-procesamiento
Salida de texto convertido
Adaptación de canal, lenguaje, ruido
Multi-motor, independiente del usuario que habla
Frases de lenguaje natural

Figura 2

Corregir: hi jonathan i will be in the stag and hounds at seven forty see you soon andy

Salida en pantalla | hi john it's tam i will be into stagecoach after four to meet you soon amy

Entrada: *<accept_word>*

Salida: **hi** | john it's tam i will be into stagecoach after four to meet you soon amy

Entrada: *3 * <accept_char>*

Salida: **hi jo** | hn it's tam i will be into stagecoach after four to meet you soon amy

Entrada: n

Salida: **hi jon** | athan i will be into stagecoach after four to meet you soon amy

Entrada: *4 * <accept_word> 3 * <accept_char>*

Salida: **hi jonathan i will be in** | to stagecoach after four to meet you soon amy

Entrada: *<space>*

Salida: hi jonathan i will be in | the stadium from after four to meet you soon amy

Entrada: *<accept_word> 4 * <accept_char>*

Salida: **hi jonathan i will be in the sta** | dium from after four to meet you soon amy

Entrada: *g*

Salida: **hi jonathan i will be in the stag** | ecoach after four to meet you soon amy

Entrada: *"*

Salida: **hi jonathan i will be in the stag** | and hounds after four to meet you soon amy

Entrada: *3 * <accept_word> 2 * <accept_char>*

Salida: **hi jonathan i will be in the stag and hounds a** | fter four to meet you soon amy

Figura 3

Entrada: *t*

Salida: **hi jonathan i will be in the stag and hounds at** | seven forty see you soon amy

Entrada: *6 * <accept_word> 2 * <accept_cbar>*

Salida: **hi jonathan i will be in the stag and hounds at seven forty see you soon a** | my

Entrada: *n*

Salida: **hi jonathan i will be in the stag and hounds at seven fourty see you soon a** | ndy

Entrada: *<accept_utterance>*

Salida en pantalla: | HEY john it's tam i will be into stagecoach after four to SOON amy

Figura 4