



República Federativa do Brasil
Ministério da Economia
Instituto Nacional da Propriedade Industrial

(11) PI 0607655-6 B1

(22) Data do Depósito: 24/01/2006

(45) Data de Concessão: 20/08/2019



* B R P I 0 6 0 7 6 5 5 B 1 *

(54) Título: MÉTODO E SISTEMA PARA FORMAR UMA MENSAGEM DE VOZ

(51) Int.Cl.: G10L 21/00; G10L 15/00.

(30) Prioridade Unionista: 16/02/2005 GB 0503162.0.

(73) Titular(es): INTERNATIONAL BUSINESS MACHINES CORPORATION.

(72) Inventor(es): STEVEN GROEGER; STUART JOHN HAYTON; JAMIE LISTER; TIMOTHY DAVID POULTNEY.

(86) Pedido PCT: PCT EP2006050403 de 24/01/2006

(87) Publicação PCT: WO 2006/087256 de 24/08/2006

(85) Data do Início da Fase Nacional: 16/08/2007

(57) Resumo: MÉTODO E APARELHO PARA A EDIÇÃO DE MENSAGEM DE VOZ. Esta invenção relaciona-se a um método e a um aparelho para a edição de mensagem de voz. Particularmente, esta se relaciona a um método e a um aparelho para juntar seções de caixas postais de mensagem. São descritos um sistema, método e produto de programa de computador para formar uma mensagem de voz em um sistema de resposta de voz interativo (IVR) falado por um usuário compreendendo: gravar uma primeira mensagem de voz, em que o usuário fala a primeira mensagem de voz, mas causa um erro em uma última porção da primeira mensagem de voz; gravar uma segunda mensagem de voz, em que o usuário fala a última porção outra vez sem o erro; determinar pontos de junção na primeira e na segunda mensagem de voz; e juntar a primeira e a segunda mensagem de voz nos pontos de junção por meio de que a primeira e a segunda mensagem de voz emendada são uma mensagem de voz contínua que inclui a última porção, mas não inclui o erro.

"MÉTODO E SISTEMA PARA FORMAR UMA MENSAGEM DE VOZ".

[0001] Esta invenção diz respeito a um método e aparelho para edição de mensagem de voz. Em particular, refere-se a um método e aparelho para unir seções de correio de voz.

Antecedentes

[0002] Um sistema de resposta de voz interativa (IVR) é um computador ligado a uma rede de telefonia e proporcionando a integração entre os dois. A rede de telefonia pode ser um simples velho sistema telefônico, como uma linha ou uma rede de telefonia comutada ou rede de telefonia de pacotes comutados como uma rede de voz sobre IP (VoIP). Uma IVR normalmente executa um aplicativo que controla a interação de um usuário e da IVR. Esse aplicativo de voz é um aplicativo de correio de voz para controlar a interação entre o usuário e o computador, em sua própria parte ou de outra interação de voz.

[0003] Uma aplicação telefonia de correio de voz requer a gravação de mensagens de áudio diretamente em um sistema IVR através do telefone. Este método de gravação pode ser complicado quando se trabalha com longos trechos de fala, especialmente se for feito um erro, enquanto da uma leitura de uma passagem. Muitas vezes a pronuncia errada de uma palavra ou tossir durante uma longa passagem cria dados áudio indesejados no segmento. Tais dados de áudio indesejados (chamados de artefatos abaixo) resultam em ter que re-gravar a mensagem inteira que custa tempo e dinheiro.

[0004] Seria útil ter um sistema de mensagem de voz que não exige uma completa regravação da voz da mensagem.

Sumário da Invenção

[0005] De acordo com um primeiro aspecto da presente invenção, é fornecido um método de formação de uma mensagem de voz em um sistema de resposta de voz interativa (IVR) falada por um usuário que inclui: a gravação de uma primeira mensagem de voz onde o usuário fala a primeira mensagem de voz, mas comete um erro na última parte da primeira mensagem de voz, grava uma segunda mensagem de voz, onde o usuário fala a última porção mais uma vez sem o erro; determina o ponto de divisão na primeira e segunda mensagem de voz, e encaixa a primeira e a segunda mensagem de voz nos pontos de em que a primeira e segunda mensagem de voz é uma mensagem de voz contínua, incluindo a última porção, mas não incluindo o erro.

[0006] Esse sistema une uma segunda mensagem de voz e uma primeira mensagem de voz e elimina a necessidade de re-gravar a totalidade da primeira mensagem.

[0007] Vantajosamente, a determinação de pontos de união compreende determinar regiões de dados de voz comum na primeira e na segunda mensagem onde os pontos de união são pontos correspondentes nas regiões de dados de voz comum.

[0008] Na concretização preferida, a determinação das regiões comuns e os pontos de união compreende as seguintes etapas: desempenhar reconhecimento de fonema na primeira e na segunda mensagem para adquirir primeiro e segundo fonemas de reconhecimento correspondentes; determinar regiões de fonemas comuns no primeiro e segundo fonemas de reconhecimento; e determinar as regiões correspondentes na primeira e segunda mensagens de voz.

[0009] Em outra concretização, a determinação da região e ponto de união comum é realizada através da comparação do sinal de energia ao longo do tempo da primeira e segunda mensagem de voz e para determinar os padrões de energia mais semelhantes, o ponto de união é um ponto baixo no sinal que se presume ser um espaço entre duas palavras.

[0010] Usando-se o reconhecimento de fala com base em fonema para determinar o ponto de união significa que o ponto de união irá situar-se entre dois silêncios em uma frase, mas não utiliza a totalidade dos recursos de processamento necessária para a identificação de palavras completas.

[0011] O uso da sequência de fonemas permite uma excelente comparação entre as mensagens e permite também uma determinação exata do espaço entre duas palavras para o ponto de união. O uso do reconhecimento por fonema significa que o reconhecimento não está restrito a uma gramática de palavra.

[0012] No entanto, em outra concretização, o reconhecimento de fala com base em palavra poderia melhorar a precisão da determinação do ponto de união enquanto usando mais poder de processamento. Além disso, o uso do reconhecimento de fala com base na palavra permite que pontos de união sejam precisamente localizados entre as palavras. Considerando que o reconhecimento com base em fonema depende dos silêncios entre os fonemas.

[0013] Por exemplo, chamador pressiona uma tecla quando um erro é cometido, enquanto a gravação de uma mensagem de áudio é feita. Pressionar uma tecla faz a

cessação da primeira gravação, avisa o usuário para iniciar a leitura da mensagem antes do erro gravado, e inicia uma segunda gravação. Ao aplicar um mecanismo de reconhecimento de fala para a primeira e segunda mensagens gravadas, é possível estimar a localização do erro na primeira gravação e unir as duas sequências de áudio juntas para remover o erro. Isto irá acelerar bastante e aumentar a facilidade de utilização da gravação de segmentos de áudio longos diretamente em uma IVR.

[0014] Nesta implementação, o aplicativo IVR registrará os primeiros segmentos de áudio e o usuário sinalizará o erro pressionando, por exemplo, a tecla * para gerar um sinal digital. Após a recepção do tom digital da tecla *, a IVR incitará o chamador para começar a falar antes de o erro ter sido feito. Quando a gravação suplementar terminar, ambas as gravações serão submetidas ao reconhecimento de fala. O texto retornado a partir do mecanismo de reconhecimento é comparado para ver onde está a sobreposição, e as amostras de tempo para as frases correspondentes em cada gravação serão coletadas através do mecanismo de reconhecimento de fala. As duas gravações serão então unidas com base nessas amostras de tempo.

Descrição dos Desenhos

[0015] Concretizações da invenção serão agora descritas, através de exemplo, apenas com referência aos desenhos anexos, em que:

A Figura 1 mostra um sistema IVR de acordo com a concretização preferida;

A Figura 2 mostra sinais de fala de exemplo que estão sendo processados pela concretização preferida.

Descrição das Concretizações

[0016] Referindo-se à Figura 1, é mostrado um sistema de resposta de voz interativa (IVR) 10 conectado através de um comutador de telefonia (PSTN) 12 a um chamador 14. O chamador 14 é uma dos muitos chamadores que poderão ser conectados a IVR 10. A IVR 10 inclui: aplicação de gravação de áudio 16 e unidade de reconhecimento de voz 18. A aplicação de gravação de áudio 16 é um programa de controle para realizar as etapas do método 102, 104, 106, 108, 110, 112 e 114. A unidade de reconhecimento de voz 18 compreende um programa de controle para executar as etapas de reconhecimento de fala 111 e 113. Nesta descrição, o usuário é referido como um chamador umz vez que as chamadas IVR são normalmente consideradas recebidas a partir do usuário, no entanto, a chamada pode ser realizada com o usuário sendo a parte chamada.

[0017] A aplicação de gravação de áudio 16 é ativada quando um chamador se liga à IVR e faz uma solicitação para deixar uma mensagem de voz.

[0018] A etapa 102 da aplicação de gravação de áudio desencadeia a IVR reproduzir uma solicitação para a pessoa que ligou para solicitar o registro da primeira mensagem de voz. Nesta concretização, o prompt mostra "Por favor, grave passagem de áudio X após o tom, pressione *, se você cometer um erro e # quando você acabar". O chamador entende que, quando um erro foi feito na gravação, então ele deve pressionar a tecla *. Se nenhum erro é feito, então, ele

compreende que pressione a tecla # no final da mensagem de voz.

[0019] Na etapa 104, o aplicativo de gravação de áudio registra a voz do chamador, como uma primeira mensagem de voz e monitora para uma tecla pressionada. Se for pressionada a tecla #, então o final da gravação foi recebido sem um erro e o processo avança para a etapa 106. Se a tecla * é pressionada, em seguida, um erro foi recebido e o processo avança para a etapa 108.

[0020] Na etapa 106, a aplicação de gravação de áudio armazena a gravação feita na etapa 104, como uma primeira mensagem de voz e se prepara para uma outra mensagem de voz a ser gravada, devolvendo o controle para a etapa 102. Nesta fase, o chamador também pode fechar o aplicativo de gravação de áudio se satisfeito com o áudio gravado. A etapa 106 também tem uma entrada de registro 114, neste caso, a gravação é uma gravação unida feita na etapa 114.

[0021] Na etapa 108, o aplicativo de gravação de áudio dirige a IVR para reproduzir uma solicitação para a pessoa que ligou para solicitar o registro da segunda mensagem de voz. Na concretização preferida, o prompt mostra "Favor começar a falar da frase antes do erro que você cometeu, pressione # para terminar." O chamador entende que ele deve começar falando de um ponto na frase antes de o erro ser cometido e depois para terminar a gravação deve pressionar # no final. A chamada de voz é gravada como a segunda mensagem de voz.

[0022] Na etapa 110, ambas a primeira e a segunda voz são apresentadas à unidade de reconhecimento de voz 18

e primeiro e segundo textos de reconhecimento são retornados. Cada texto de reconhecimento inclui uma sequência de rótulos de texto correspondendo à mensagem de voz.

[0023] Na etapa 111, a unidade de reconhecimento de voz processa a primeira e a segunda mensagem de voz. Na concretização preferida, apenas o reconhecimento de voz parcial é realizado e os rótulos de texto devolvidos são os fonemas correspondentes aos dados de voz. Em outra concretização, texto legível por humanos é retornado.

[0024] Na etapa 112, a aplicação de gravação de áudio localiza os rótulos de texto correspondentes na primeira e segunda mensagens de texto. As posições dos rótulos correspondentes na primeira e na segunda mensagens de texto são inseridos em uma consulta para a unidade de reconhecimento de voz. Usando a posição dos rótulos correspondentes, a unidade de reconhecimento de voz localiza a primeira amostra de tempo na primeira mensagem de voz e uma segunda amostra de tempo na segunda mensagem de voz. Estas representam as amostras de tempo dos pontos de união na primeira e segunda mensagens de voz.

[0025] Na etapa 113, a unidade de reconhecimento de voz recebe a amostra de tempo consulta e retorna a primeira e a segunda amostra de tempos. Em outra concretização, a amostra de tempo para os rótulos de texto pode ser fornecida pela unidade de reconhecimento de voz, ao mesmo tempo que os rótulos de texto.

[0026] Na etapa 114, a primeira e a segunda amostra de tempos localizadas são usadas para unir o primeiro e o segundo segmentos de voz. O processo, em seguida, retorna à

etapa 106 em que o segmento de voz unido é gravado e o aplicativo de áudio sai ou se prepara para a próxima passagem.

[0027] A concretização preferida relacionada com uma IVR de correio de voz, mas essa solução poderia ser utilizada em outras aplicações de voz como ditado de voz pessoal.

[0028] Um exemplo da expressão, sinal de discurso e transcrição fonética envolvidos na união de duas mensagens de voz é mostrado na Figura 2. O usuário fala a primeira mensagem de voz Discurso A "O serviço atual está indisponível, tente mais tarde", que é mostrado no topo da Figura 2. Diretamente abaixo do discurso A na Figura 2 está a gravação A - um exemplo da amplitude do sinal contra o tempo correspondendo às palavras no discurso A. Diretamente abaixo da gravação A está uma transcrição fonética representando cadeia de fonema derivada da gravação A pela unidade de reconhecimento de voz. A escala de tempo em milisegundos é mostrada ao longo do eixo x sob transcrição fonética A. Para um determinado texto, ou sequência de fonema ou sinal, a unidade reconhecimento de voz pode retornar a amostra de tempo de início e a amostra de tempo final.

[0029] O usuário fala a segunda mensagem de voz Discurso B "por favor, tente novamente mais tarde" mostrado sob o acesso de escala tempo na Figura 2. Diretamente abaixo do Discurso B está a Gravação B - um exemplo do sinal em função do tempo correspondente às palavras do Discurso B. Diretamente abaixo da Gravação B, a transcrição fonética B representando a sequência de fonemas derivada da Gravação B pela unidade de reconhecimento de voz. A escala de tempo em

milissegundos, ao longo do eixo x representa apenas a escala de tempo relativa para a transcrição fonética B.

[0030] As seqüências de fonemas resultante são mostradas a seguir:

Transcrição	fonética	A
"thequrreentservisizunavaylablpleeztriyayter"		

Transcrição fonética B "pleeztriyagaynlayter"

[0031] O objetivo do alinhamento é o de encontrar onde a segunda mensagem começa na primeira mensagem. A concretização preferida encontra a maior subsequência entre as duas - neste caso, a subsequência 'pleeztriy'. Um usuário da presente concretização terá tendência a dizer uma frase semelhante ou idêntica como uma referência para alinhar contra e, em seguida, dizer alguma coisa diferente (a correção). O método preferido é o de manter o áudio da primeira mensagem de voz até o ponto de união (mesmo antes da primeira ocorrência dos fonemas semelhantes para a segunda mensagem de voz) e juntar todo o áudio a partir da segunda mensagem de voz.

[0032] A concretização preferida funciona bem com bom reconhecimento de voz, precisão de reconhecimento imperfeita irá introduzir erros na forma de inserções, repetições, substituições e supressões de fonemas. Neste caso, um algoritmo de correspondência mais complicado pode ser usado que leva em conta a possibilidade de reconhecimento de erros.

[0033] Um outro exemplo é descrito quando o reconhecimento de fala completo é utilizado e este exemplo não é ilustrado.

[0034] Um usuário pode ler: "... Dois aumentos em três meses teria derramado combustível sobre o fogo de expectativas de novos aumentos para vir, potencialmente ****". Neste exemplo **** representa uma tosse, um outro ruído produzido em erro, ou qualquer outro erro. A IVR registra isso como uma primeira mensagem de voz.

[0035] O usuário tecla: * e continua a ler: "as expectativas de novos aumentos para vir, potencialmente reduzindo à libra a níveis prejudiciais face ao dólar...". A IVR registra esta leitura como uma segunda mensagem de voz.

[0036] O usuário tecla: # para terminar a gravação da segunda mensagem de voz.

[0037] As duas mensagens de voz são retornadas como mensagens de texto de um mecanismo de reconhecimento de fala:

[0038] Primeira mensagem de texto: "... Dois aumentos em três meses teria derramado combustível sobre o fogo de expectativas de novos aumentos para vir, potencialmente #####", onde o ##### representa as letras devolvidas pelo reconhecedor representando o erro.

[0039] Segunda mensagem de texto: "expectativas de novos aumentos para vir, potencialmente reduzindo a níveis prejudiciais a libra frente ao dólar ..."

[0040] Os textos das mensagens de voz são tratados posteriormente para determinar a sobreposição de partes:

[0041] Primeira mensagem: "... Dois aumentos em três meses teria derramado combustível sobre o fogo de expectativas de novos aumentos para vir, potencialmente #####"

[0042] Segunda mensagem: "expectativas de novos aumentos para vir, potencialmente reduzindo a níveis prejudiciais a libra frente ao dólar ..." onde as partes sobrepostas estão sublinhadas.

[0043] Amostras de tempos (em segundos) nas mensagens de voz correspondentes ao início e final da frase correspondente no primeiro segmento de texto, e a segunda mensagem de texto são adquiridos.

Primeira mensagem de voz: 05:06:43 - 05:09:90 segundos

Segunda mensagem de voz: 00:02.81 - 00:05:27 segundos

[0044] A primeira e a segunda mensagens de voz são unidas com base em amostras de tempos adquiridas: mensagem de voz Final = Primeira mensagem de voz (00:00:00 - 05:06:43) + segunda mensagem de voz (00:02:81 - Fim de áudio)

[0045] A aplicação de gravação de áudio da concretização preferida é parte de um sistema de mensagens no servidor IVR e um sistema telefônico. No entanto, em uma concretização alternativa, aplicação de gravação de áudio pode ser parte do telefone ou dispositivo cliente e interage para deixar uma mensagem com uma mensagem completada em um sistema de mensagens em um servidor IVR. Nesta concretização alternativa, o cliente pode precisar fazer o download do programa de gravação de áudio a partir de um servidor antes da execução.

[0046] Em resumo, esta especificação refere-se a um método e aparelho para editar mensagem de voz. Em particular, refere-se a um método e aparelhos para unir seções de correio de voz. É descrito um sistema, método e produto de programa de computador formando uma mensagem de voz em um sistema de

resposta de voz interativo (IVR) falada por um usuário que inclui: a gravação de uma primeira mensagem de voz onde o usuário fala a primeira mensagem de voz, mas comete um erro em uma última parte da primeira mensagem de voz, gravação de uma segunda mensagem de voz onde o usuário fala a última porção mais uma vez sem o erro; determinação do ponto de união na primeira e segunda mensagem de voz, e unir a primeira e a segunda voz no ponto de união em que a primeira e a segunda mensagem de voz é uma mensagem de voz contínua, incluindo a última porção, mas não incluindo o erro.

REIVINDICAÇÕES

1. Método para formar uma mensagem de voz falada por um usuário em um sistema de resposta de voz interativo (IVR) **caracterizado pelo** fato de que compreende:

gravar uma primeira mensagem de voz, em que o usuário fala a primeira mensagem de voz, mas causa um erro em uma última porção da primeira mensagem de voz;

gravar uma segunda mensagem de voz, em que o usuário fala a última porção outra vez sem o erro;

determinar automaticamente pontos de junção na primeira e na segunda mensagem de voz; e

juntar a primeira e a segunda mensagem de voz nos pontos de junção por onde a primeira e a segunda mensagem de voz unidas são uma mensagem de voz contínua que inclui a última porção, mas não inclui o erro;

em que a determinação dos pontos de junção compreende a determinação de regiões de dados de voz comuns na primeira e na segunda mensagem, em que os pontos de junção são pontos correspondentes dentro das regiões de dados de voz comuns.

2. Método, de acordo com a reivindicação 1, **caracterizado pelo** fato de que a determinação das regiões comuns e dos pontos de junção compreende as seguintes etapas:

executar o reconhecimento de voz na primeira e na segunda mensagem para adquirir primeiro e o segundo texto de reconhecimento correspondente;

determinar regiões de texto comum no primeiro e segundo texto de reconhecimento; e

determinar as regiões correspondentes na primeira e segunda mensagem de voz.

3. Método, de acordo com a reivindicação 2, **caracterizado pelo** fato de que o texto de reconhecimento compreende uma sequência de fonema e não uma palavra da linguagem de modo a reduzir a quantidade de processamento usada para determinar o ponto de junção.

4. Método, de acordo com a reivindicação 1, **caracterizado pelo** fato de que a determinação da região comum e o ponto de junção é executada pela comparação da energia do sinal ao longo do tempo da primeira e da segunda mensagem de voz e determinação dos padrões de energia mais similares, o ponto de junção é um ponto baixo no sinal que é suposto como sendo um espaço entre duas palavras.

5. Sistema para formar uma mensagem de voz falada por um usuário para um sistema de resposta de voz interativo (IVR) **caracterizado pelo** fato de que compreende:

meio para gravar uma primeira mensagem de voz, em que o usuário fala a primeira mensagem de voz, mas causa um erro em uma última porção da primeira mensagem de voz;

meio para gravar uma segunda mensagem de voz, em que o usuário fala a última porção outra vez sem o erro;

meio para determinar automaticamente os pontos de junção na primeira e na segunda mensagem de voz; e

meio para juntar a primeira e a segunda mensagem de voz nos pontos de junção por meio de que a primeira e a segunda mensagem de voz emendada são uma mensagem de voz contínua que inclui a última porção, mas não inclui o erro;

em que o meio para determinação dos pontos de junção compreende meio para determinação de regiões de dados de voz comuns na primeira e na segunda mensagem, em que os pontos

de junção são pontos correspondentes dentro das regiões de dados de voz comuns.

6. Sistema, de acordo com a reivindicação 5, **caracterizado pelo** fato de que o meio para determinação das regiões comuns e dos pontos de junção compreende:

meio para executar o reconhecimento de voz na primeira e na segunda mensagem para adquirir primeiro e segundo textos de reconhecimento correspondentes;

meios para determinar as regiões do texto comum no primeiro e no segundo texto de reconhecimento; e

meios para determinar as regiões correspondentes na primeira e na segunda mensagem de voz.

7. Sistema, de acordo com a reivindicação 6, **caracterizado pelo** fato de que o texto de reconhecimento compreende uma sequência de fonemas e não uma palavra da linguagem de modo a reduzir a quantidade de processamento usada para determinar o ponto de junção.

8. Sistema, de acordo com a reivindicação 5, **caracterizado pelo** fato de que compreende meio para determinação da região comum e o ponto de junção é executado pelo meio para comparação da energia do sinal ao longo do tempo da primeira e da segunda mensagem de voz e meio para determinação dos padrões de energia mais similares, o ponto de junção é um ponto baixo no sinal que é suposto como sendo um espaço entre duas palavras.

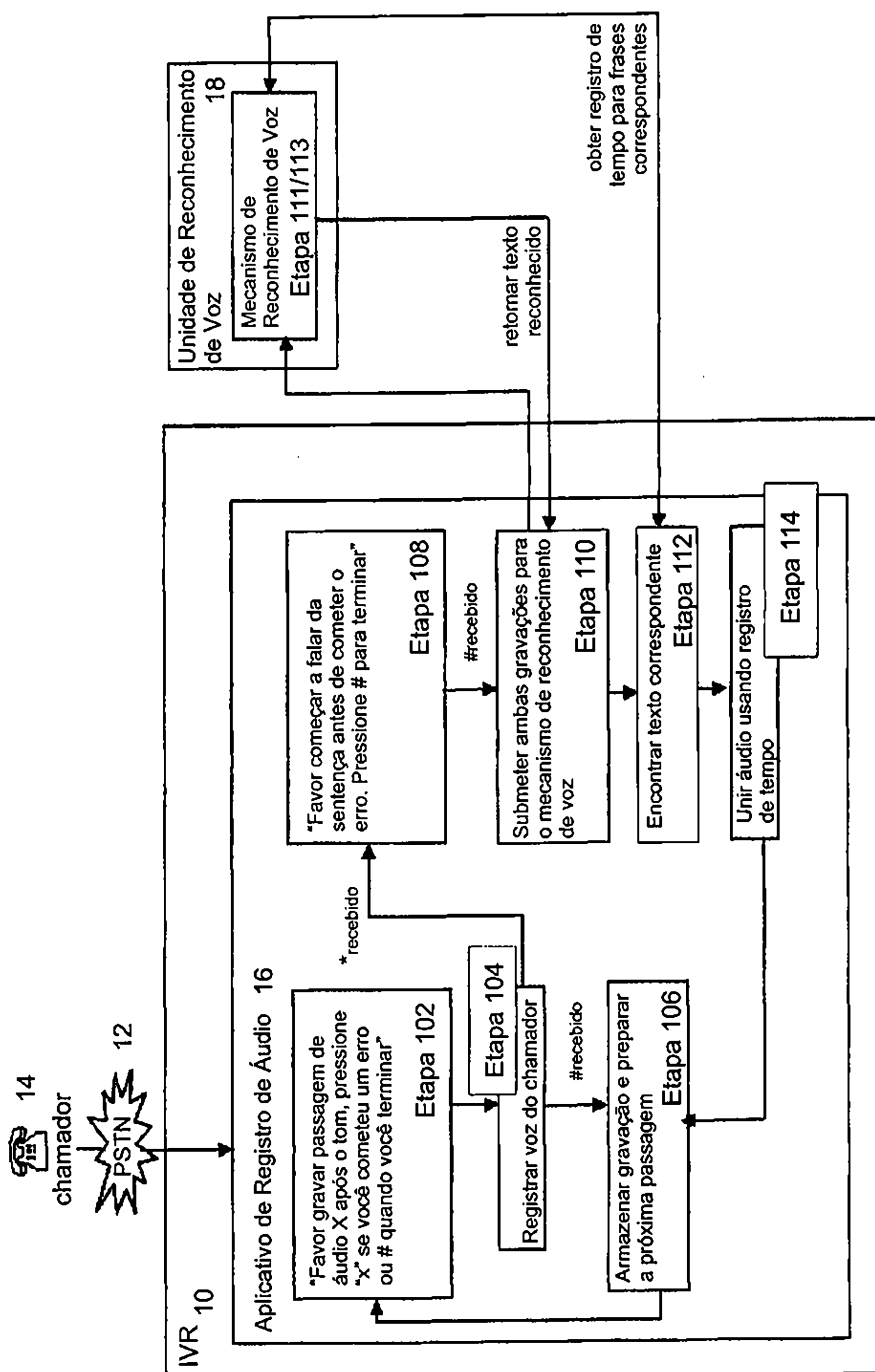


FIGURA 1

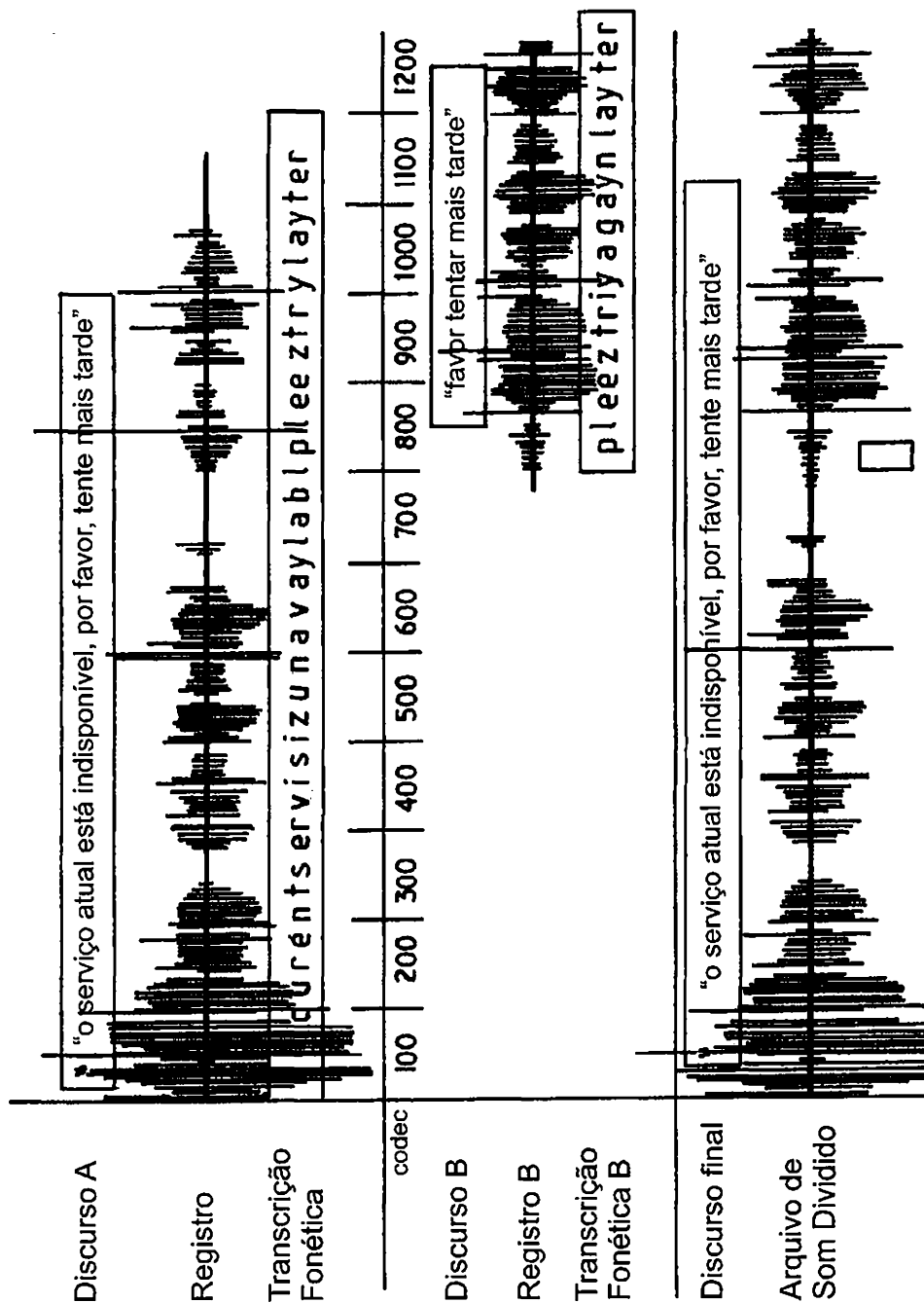


FIGURA 2