ROYAUME DE BELGIQUE

MINISTERE DES AFFAIRES ECONOMIQUES

BREVET D'INVENTION

NUMERO DE PUBLICATION : 1011892A3

NUMERO DE DEPOT : 09800314

Classif. Internat. : G10L

Date de délivrance le : 01 Février 2000

Le Ministre des Affaires Economiques,

Vu la Convention de Paris du 20 Mars 1883 pour la Protection de la propriété industrielle;

Vu la loi du 28 Mars 1984 sur les brevets d'invention, notamment l'article 22;

Vu l' arrêté royal du 2 Décembre 1986 relatif à la demande, à la délivrance et au maintien en vigueur des brevets d'invention, notamment l'article 28;

Vu le procès verbal dressé le 27 Avril 1998 à 14H25 à l 'Office de la Propriété Industrielle

ARRETE:

ARTICLE 1.- Il est délivré à : MOTOROLA INC.
1303 East Algonquin Road, SCHAUMBURG ILLINOIS 60196(ETATS-UNIS D'AMERIQUE)

représenté(e)(s) par : VOSSWINKEL Philippe, GEVERS & VANDER HAEGHEN, Rue de Livourne 7, -B 1060 BRUXELLES.

un brevet d' invention d' une durée de 20 ans, sous réserve du paiement des taxes annuelles, pour : METHODE, DISPOSITIF ET SYSTÈME POUR GENERER DES PARAMETRES DE SYNTHESE VOCALE A PARTIR D'INFORMATIONS COMPRENANT UNE REPRESENTATION EXPLICITE DE L'INTONATION.

INVENTEUR(S) : Corrigan Gerald E., 1948 W. Farwell Avenue, Chicago 60626 (US);Massey Noel, 2370 John Rolfe Drive, Apt.C., Schaumburg, Illinois 60194 (US); Karaali Orhan, 5 Juniper Road, Rolling Meadows, Illinois 60056, (US)

PRIORITE(S) 22.05.97 US USA 861751

ARTICLE 2.- Ce brevet est délivré sans examen préalable de la brevetabilité de l'invention, sans garantie du mérite de l'invention ou de l'exactitude de la description de celle-ci et aux risques et périls du(des) demandeurs(s).

Bruxelles, le 01 Février 2000
PAR DELEGATION SPECIALE :

PETIT M.
Conseiller adjoint

"Méthode, dispositif et système pour générer des paramètres de synthèse vocale à partir d'informations comprenant une représentation explicite de l'intonation"

Domaine de l'invention

La présente invention concerne des systèmes générant des paramètres utilisés dans la synthèse vocale, et plus particulièrement le codage de l'intonation dans des systèmes de codeur générant des paramètres utilisés dans la synthèse vocale.

Contexte de l'invention

Comme la figure 1, référence 100, le montre, pour convertir du texte en paroles, des systèmes statistiques (102) convertissent de manière typique des représentations linguistiques (phonétiques) de texte en paramètres caractérisant des formes d'onde vocale. Le système statistique illustré sur la figure 1 utilise deux systèmes statistiques (réseaux neuraux) (110 et 118). Un réseau neural (110) convertit les descriptions linguistiques de phones vocaux et de leurs contextes en durées de segment pour des segments de formes d'onde vocale associés aux phones. Le deuxième réseau neural (118) convertit les descriptions linguistiques de séquences vocales (parties de formes d'onde vocale survenant pendant de brèves périodes de temps) en descriptions acoustiques (120) des séquences. Celles-ci sont alors converties en formes d'onde vocale (124) en utilisant un vocodeur (122).

Un problème posé par les approches statistiques basées sur l'information linguistique extraite du texte réside dans le fait que le texte ne contient pas suffisamment d'information pour générer correctement la prosodie (rythme et intonation) dans les formes d'onde vocale. On sait que, pour tout texte donné, il existe une multitude de contours intonatifs et de configurations rythmiques qui peuvent être générés pour ce texte. Les conversions effectuées par les systèmes statistiques sont contrôlées par des descriptions statistiques de données d'entraînement consistant de manière typique en un ensemble de

vecteurs comprenant des entrées possibles dans le système et des sorties souhaitées quand les entrées possibles sont présentées au système. Dans les systèmes statistiques utilisés pour convertir du texte en parole, les données d'entraînement sont générées de manière typique par l'analyse du langage naturel. Étant donné que les contours intonatifs et les configurations rythmiques ne peuvent être prédits à partir de l'information linguistique extraite de textes, les systèmes statistiques tendent à produire une prosodie qui rend moyennement les contours et rythmes possibles. Cette reproduction moyenne des contours et des rythmes peut rendre la parole moins compréhensible.

C'est pourquoi il faut une méthode et un dispositif pour améliorer la performance de génération de la prosodie d'un système convertissant du texte en parole.

Brève description des dessins

La figure 1 est une représentation schématique d'un système convertissant du texte en parole comprenant deux systèmes statistiques pour générer des paramètres vocaux que la technique connaît.

La figure 2 est une représentation schématique d'un système pour générer des paramètres vocaux conformément à la présente invention.

La figure 3 est une représentation schématique du processus d'entraînement d'un réseau neural conformément à la technologie existante.

La figure 4 est une représentation schématique du processus d'entraînement d'un réseau neural conformément à la présente invention.

La figure 5 est une représentation schématique d'un système convertissant du texte en parole englobant la présente invention.

La figure 6 est un schéma fonctionnel d'une réalisation des

étapes conformément à la méthode de la présente invention.

La figure 7 est une représentation schématique d'une réalisation préférée d'un dispositif en conformité avec la présente invention.

Description détaillée d'une réalisation préférée

La présente invention prévoit une méthode, un dispositif et un système pour améliorer la prosodie exprimée par un système générant des paramètres vocaux en intégrant une représentation explicite de la prosodie comme entrée dans le système générant des paramètres vocaux. Cette amélioration donne un rythme et une intonation plus réaliste à la parole, rendant la signification de la parole plus compréhensible que la parole dans des systèmes convertissant du texte en parole dans l'art antérieur.

Dans une réalisation préférée, le système générant des paramètres vocaux de la présente invention est un réseau neural produisant une série de paramètres ou des vecteurs de paramètre consistant en un ou plusieurs paramètres qui décrivent un aspect prédéterminé de la forme d'onde vocale. Comme la figure 2, référence 200, le montre, certains paramètres peuvent être les descriptions acoustiques des séquences (220) ou les durées des segments (206). La figure 2 illustre une réalisation préférée d'un dispositif dans un système conforme à la présente invention. Le dispositif donne, en réponse à des informations phonétiques / linguistiques prédéterminées (222) et à des informations prosodiques prédéterminées (224), une génération efficace de la description acoustique des séquences (220), c'est-à-dire des paramètres vocaux prosodiquement améliorés. Le dispositif comprend: A) une unité de calcul de la durée des segments (202), couplée pour recevoir l'information phonétique / linguistique et à une unité de génération de la prosodie (204), pour convertir l'information phonétique / linguistique en durées de segment (206), B) l'unité de génération de la prosodie (204), couplée pour recevoir l'information prosodique

prédéterminée, utiliser l'information prosodique prédéterminée afin de générer la sortie prosodique (208), C) une unité de génération de la description des séquences (210), couplée à une unité de calcul de la durée des segments et à l'unité de génération de la prosodie et afin de recevoir l'information phonétique / linguistique, pour utiliser les durées de segment (206), la sortie prosodique (208) et l'information phonétique / linguistique (222) afin de générer des descriptions linguistiques / prosodiques des séquences (212), et D) une unité de calcul des descriptions acoustiques (214), couplée à une unité de génération de la description des séquences (210), pour calculer les descriptions acoustiques des séquences (220) pour l'information prosodique prédéterminée (224) et l'information phonétique / linguistique (222) afin de générer des paramètres vocaux qui fournissent une reproduction prosodique fiable. Les paramètres conviennent de manière typique pour être utilisés comme entrée dans un synthétiseur ou un générateur de formes d'onde (216) comme un vocodeur, qui génère une forme d'onde vocale (218).

Dans une réalisation préférée, le dispositif générant les paramètres vocaux est un réseau neural produisant une série de vecteurs de paramètre consistant en un ou plusieurs paramètres décrivant un quelconque aspect de la forme d'onde vocale. Sur la figure 1, les paramètres pourraient être les descriptions acoustiques des séquences (120) ou les durées des segments (112). La figure 7, référence 700, est une représentation schématique d'une réalisation préférée d'un dispositif conforme à la présente invention. Comme dans les systèmes existants générant des paramètres vocaux, cette invention utilise un système statistique (702) pour convertir l'information comprenant l'information linguistique (704) provenant du texte en certains paramètres décrivant un quelconque aspect de la parole à générer (706). Cependant, contrairement aux systèmes existants générant des paramètres vocaux, cette invention comprend une unité de

détermination de la prosodie (712) qui convertit l'information (710) sur le style et la netteté de l'expression en une description prosodique (708) de la parole à générer. Cette description prosodique est également fournie au réseau neural pour déterminer les paramètres vocaux.

Dans une réalisation préférée, le dispositif fournit, en réponse à l'information comprenant l'information linguistique et à au moins un des éléments: information sur le style et information sur la netteté, une génération efficace de paramètres vocaux prosodiquement améliorés. Le dispositif comprend: A) une unité de détermination de la prosodie (712), couplée pour recevoir au moins un des éléments, information sur le style et information sur la netteté, générant l'information prosodique, décrivant le rythme et l'intonation de la parole à générer, et B) un système statistique (702), qui dans la réalisation préférée est un réseau neural, couplé pour recevoir l'information comprenant l'information linguistique et prosodique, pour fournir des paramètres décrivant des parties de formes d'onde vocale.

L'information (710) sur le style et la netteté d'expression peut être fournie par un modèle de dialogue ou être introduite par l'usager. D'un autre côté, la détermination du style et de la netteté peut être faite à l'avance pour différents types de phrase, comme des déclarations, des questions ou des ordres.

De manière typique, comme la figure 3, référence 300, le montre, la méthode pour générer des réseaux neuraux pour la synthèse vocale a consisté à entraîner un réseau neural à prédire les paramètres vocaux pour une partie particulière de la forme d'onde vocale. La parole (302) est tout d'abord enregistrée et stockée dans une base de données audio (304). La parole est tout d'abord étiquetée phonétiquement et syntaxiquement (unité d'étiquetage phonétique / syntaxique, 306) et l'information de l'étiquette est stockée dans une base de données phonétiques (308). (D'un autre côté, l'étiquetage phonétique / syntaxique peut se faire manuellement.) La base de données audio (304) et / ou la

base de données phonétiques (308) sont alors traitées par l'unité de calcul des sorties souhaitées (310) pour extraire des paramètres décrivant une quelconque partie de la parole enregistrée, utilisée comme sortie d'entraînement (312). Pour chaque valeur de sortie d'entraînement, la base de données phonétiques (308) est traitée par l'unité de génération des entrées (314) pour produire les entrées d'entraînement (316) pour un réseau neural (318) qui génère les paramètres vocaux (320). Le réseau neural est alors entraîné pour générer une bonne approximation de la sortie d'entraînement en réponse à l'entrée d'entraînement, en utilisant un quelconque critère de qualité, comme la différence quadratique moyenne minimale entre la sortie du réseau neural et la sortie d'entraînement

Le réseau neural dans la réalisation préférée de la présente invention est généré comme la figure 4 le montre. La parole (402) est tout d'abord enregistrée et stockée dans une base de données audio (404). Dans un premier temps, la parole est étiquetée phonétiquement et syntaxiquement (unité d'étiquetage phonétique / syntaxique, 406) et l'information de l'étiquette est stockée dans une base de données phonétiques (408). La parole est alors étiquetée prosodiquement (unité d'étiquetage prosodique, 410), indiquant la prosodie effective de la parole enregistrée, pour créer une base de données prosodiques (412). (D'un autre côté, soit l'étiquetage phonétique / syntaxique soit l'étiquetage prosodique soit les deux peuvent se faire manuellement.) La base de données audio (404) ou la base de données phonétiques (408) ou les deux sont alors traitées (unité de calcul des sorties souhaitées, 414) pour extraire des paramètres décrivant une quelconque partie de la parole enregistrée, utilisée comme sortie d'entraînement (416). Pour chaque valeur de sortie d'entraînement, la base de données phonétiques (408) et la base de données prosodiques (412) sont traitées (unité de génération des entrées, 418) pour produire une entrée d'entraînement (420) pour un

réseau neural (422). Le réseau neural est alors entraîné pour générer une bonne approximation de la sortie d'entraînement en réponse à l'entrée d'entraînement, en utilisant un critère prédéterminé de qualité comme la différence quadratique moyenne minimale entre la sortie du réseau neural et la sortie d'entraînement. Cela est similaire au processus utilisé pour l'entraînement dans la technologique existante, sauf que la parole est étiquetée prosodiquement pour générer la base de données prosodiques (412) qui est utilisée dans la génération des entrées (unité de génération d'entrées, 418).

La figure 5, référence 500, est une représentation schématique d'un système convertissant du texte en parole englobant la présente invention. Celui-ci est similaire au système convertissant du texte en parole de la figure 1, sauf qu'il y a des informations prosodiques prédéterminées (508) qui sont introduites en plus du texte (504) qui est utilisé pour déterminer la prosodie de la parole générée. Le texte (504) est introduit dans une unité de conversion du texte en linguistique (506) qui fournit une description (phonétique) linguistique (510) à une unité de calcul de la durée des segments (514) qui est de manière typique un réseau neural. L'unité de conversion des calculs de la durée des segments (514) utilise la description (phonétique) linguistique (510) et la sortie prosodique (512) pour sortir les durées de segment (516). L'information prosodique prédéterminée (528) peut venir d'un modèle de dialogue ou de sélections d'usager, ou être simplement un ensemble de décisions arbitraires préalables concernant la façon dont la prosodie sera générée. Cette information ainsi que l'information provenant du texte est utilisée dans la génération de la prosodie (unité de génération de la prosodie, 508) pour produire la sortie de la prosodie (512) fournie par l'unité de calcul de la durée des segments (514). Cette information est aussi fournie à l'unité de génération de la description des séquences (518) pour produire les descriptions linguistiques / prosodiques des séquences (520) fournies à l'unité de calcul des descriptions

acoustiques (522). L'unité de calcul de la durée des segments (514) et l'unité de calcul des descriptions acoustiques (522) sont des exemples de la présente invention. La sortie de l'unité de calcul des descriptions acoustiques (522) est alors introduite dans une unité de génération de formes d'onde (vocodeur, 524) qui génère une forme d'onde vocale (526).

L'information prosodique prédéterminée (528) peut être l'information sur le style et la netteté d'expression qui est fournie par un modèle de dialogue ou introduite par l'usager. D'un autre côté, la détermination du style et de la netteté peuvent se faire à l'avance pour différents types de phrases comme des déclarations, des questions et des ordres.

Les systèmes générant des paramètres vocaux (514 et 522) peuvent être chacun un réseau neural, une arborescence de décisions, un algorithme génétique ou la combinaison de deux de ceux-ci ou plus.

Les méthodes existantes de synthèse statistique génèrent la prosodie de la parole en n'utilisant que l'information phonétique et l'information qui peut être extraite du texte. Étant donné que ces méthodes ont tendance à aplanir les contours intonatifs et les configurations rythmiques qui peuvent survenir dans la parole, produisant des variations prosodiques qui ne sont pas claires, la méthode de la présente invention a été mise au point pour générer la prosodie en utilisant une représentation explicite de la prosodie. Par conséquent, la présente invention fournit une intonation et un rythme plus naturels à la parole générée.

Comme les étapes représentées sur la figure 6, référence 600, la méthode de la présente invention fournit, en réponse à l'information comprenant l'information linguistique et au moins un des éléments: information sur le style et information sur la netteté, une génération efficace des paramètres vocaux prosodiquement améliorés.

Dans une réalisation, la méthode comprend les étapes de A) détermination (602) de l'information prosodique qui décrit le rythme et l'intonation de la parole à générer, à partir d'au moins un des éléments: information sur le style et information sur la netteté, et B) d'utilisation (604) d'un système statistique pour convertir l'information comprenant l'information linguistique et l'information prosodique en paramètres vocaux, qui décrivent des parties de formes d'onde vocale.

Les paramètres vocaux sont de manière typique des paramètres convenant pour être utilisés comme entrées dans un synthétiseur / générateur de formes d'onde qui peut être utilisé pour synthétiser la parole. D'un autre côté, les paramètres vocaux peuvent être des durées de segment.

Le logiciel qui met en œuvre la présente invention peut être intégré dans un microprocesseur ou un processeur de signaux numériques (DSP). D'un autre côté, la méthode peut être mise en œuvre par un circuit intégré spécifique à une application (ASIC). La méthode peut également être mise en œuvre par une combinaison de microprocesseur, de DSP et d'ASIC.

L'information prosodique prédéterminée comprend en général au moins un de: A) les emplacements des terminaisons de mot et un degré de disjonction entre les mots, B) les emplacements des accents toniques et une forme des accents toniques, et C) les emplacements des limites marquées dans un contour de tonalité et une forme des limites. Un format de l'information prosodique prédéterminée est de manière typique un de: A) l'information décrivant une proximité d'événements prosodiques marqués par rapport à des séquences définies entourant une séquence pour laquelle des paramètres du codeur sont générés, B) l'information décrivant un temps séparant des événements prosodiques marqués d'une séquence pour laquelle les paramètres du codeur sont générés, C) l'information décrivant un temps séparant certains événements prosodiques marqués d'autres

événements prosodiques marqués, D) l'information décrivant un certain nombre d'événements prosodiques d'un type dans une période de temps séparant un événement prosodique d'un autre type et une séquence pour laquelle les paramètres du codeur sont générés, E) l'information décrivant un certain nombre d'événements prosodiques d'un type survenant dans une période de temps séparant deux événements prosodiques d'un autre type, et F) au moins deux des formats A-E. Si un choix est fait, les durées de segment peuvent être utilisées comme entrée supplémentaire dans une unité de génération des paramètres des séquences acoustiques.

Les paramètres vocaux peuvent être sélectionnés pour être des durées de segments vocaux associés à des phones, et la méthode peut être mise en œuvre en utilisant un logiciel et / ou un matériel comme décrit ci-dessus. Dans cette mise en œuvre, l'information prosodique prédéterminée et le format de l'information prosodique prédéterminée sont comme décrits ci-dessus.

Si un choix est fait, au moins l'une des unités, unité de calcul de la durée des segments et unité de calcul des descriptions acoustiques, peut être un réseau neural, une unité d'arborescence des décisions ou une unité qui utilise un algorithme génétique.

Dans la réalisation préférée de l'invention, l'information prosodique utilisée provient du système d'étiquetage TOBI mis au point par Silverman et al. (Silverman K. et al. "TOBI: A standard for Labeling English Prosody", Proc. ICSLP 92, pp. 867-870, Banff, Octobre 1992). Le moment où chaque mot se termine est marqué conjointement avec un nombre (l'indice de coupure) qui indique le degré de disjonction entre le mot marqué et le mot suivant. Les tonalités sont aussi marquées avec un inventaire de symboles pour les accents toniques sur les syllabes et les marques des limites intonatives.

Il est possible d'avoir recours à diverses représentations pour l'information prosodique fournie comme entrée au système

statistique. Dans une représentation temporisée d'un réseau neural, l'entrée consiste en une série de vecteurs, chaque vecteur représentant le contexte linguistique pendant une quelconque période de temps d'échantillonnage près de la partie de la forme d'onde pour laquelle le système génère des paramètres vocaux (la partie actuelle de la forme d'onde vocale), dans ce cas, différentes entrées peuvent indiquer la présence d'un ton ou d'une marque d'indice de rupture dans la période de temps d'échantillonnage, ou la proximité d'une marque par rapport à la période de temps d'échantillonnage. Comme solution de rechange à la représentation temporisée, l'entrée peut indiquer les distances entre la partie actuelle de la forme d'onde vocale et les événements qui sont marqués dans l'information prosodique. De plus, les distances entre ces événements peuvent être fournies comme entrée. Les distances entre deux éléments peuvent être mesurées comme la période de temps les séparant, ou comme comptages du nombre d'événements d'un autre type séparant les deux éléments. Par exemple, une entrée peut représenter le temps entre la partie actuelle de la forme d'onde vocale et l'occurrence précédente d'une marque de limite intonative tandis qu'une autre peut indique le nombre d'accents toniques avec baisse (un rétrécissement de la plage de tonalité) qui survient depuis la limite. Quand les paramètres vocaux sont des représentations acoustiques de séquences vocales convenant pour être utilisées comme entrée dans un synthétiseur de formes d'onde, la réalisation préférée utilise une combinaison de ces trois techniques. Quand les paramètres vocaux sont des durées de segment, la réalisation préférée n'utilise pas une représentation temporisée pour représenter l'information prosodique.

Le logiciel mettant en œuvre la méthode peut être intégré dans un microprocesseur ou dans un processeur de signaux numériques. D'un autre côté, un circuit intégré spécifique à une application peut mettre en œuvre la méthode, ou une combinaison de plusieurs de ces mises en œuvre peut être utilisée. Dans une réalisation

préférée, le système générant les paramètres vocaux utilise des réseaux neuraux. D'un autre côté, le système générant des paramètres vocaux peut utiliser des unités d'arborescence des décisions ou des algorithmes génétiques.

Les parties des formes d'onde vocale décrites peuvent être des segments associés à des éléments phonétiques comme des phones, et les paramètres générés peuvent être les durées de ces segments.

Les parties des formes d'onde vocale décrites peuvent être des séquences vocales et les paramètres générés peuvent être une représentation acoustique de ces séquences.

Le logiciel mettant en œuvre la méthode peut être intégré dans un microprocesseur ou dans un processeur de signaux numériques. D'un autre côté, un circuit intégré spécifique à une application peut mettre en œuvre la méthode, ou une combinaison de plusieurs de ces mises en œuvre peut être utilisée. Dans une réalisation préférée, le système générant les paramètres vocaux utilise des réseaux neuraux. D'un autre côté, le système générant les paramètres vocaux peut utiliser des unités d'arborescence des décisions ou des algorithmes génétiques.

La présente invention peut être mise en œuvre par un dispositif destiné à fournir, en réponse à l'information comprenant l'information phonétique et prosodique, une génération efficace des paramètres vocaux. Le dispositif comprend un système statistique, couplé pour recevoir l'information comprenant l'information phonétique et prosodique, destiné à fournir des paramètres décrivant des parties de formes d'onde vocale.

Les parties de formes d'onde vocale décrites peuvent être des segments associés à des éléments phonétiques comme des phones, et les paramètres générés peuvent être les durées de ces segments.

Les parties de formes d'onde vocale décrites peuvent être des séquences vocales, et les paramètres générés peuvent être une

représentation acoustique de ces séquences.

À nouveau, le système générant les paramètres vocaux peut comprendre des réseaux neuraux, des unités d'arborescence des décisions, ou des unités qui utilisent un algorithme génétique.

Le dispositif de la présente invention peut être mis en œuvre dans un système convertissant du texte en parole, un système de synthèse vocale ou un système de dialogue.

La présente invention peut être réalisée dans d'autres formes spécifiques sans s'écarter de son esprit ou de ses caractéristiques essentielles. Les réalisations décrites ne sont à considérer à tous égards que comme des exemples et comme non restrictives. La portée de l'invention est, par conséquent, indiquée dans les revendications annexées plutôt que dans la description susdite. Tous les changements qui relèvent du sens et du champ d'équivalence des revendications sont à inclure dans son champ d'application.

Figure 1 - art antérieur

102 Système statistique convertissant du texte en parole

104 Texte

106 Conversion de texte en linguistique

108 Description (phonétique) linguistique

110 Calcul de la durée des segments (réseau neural)

112 Durées des segments

114 Génération de la description des séquences

116 Description linguistique des séquences

118 Calcul des descriptions acoustiques (système neural)

120 Description acoustique des séquences

122 Génération de formes d'onde (vocodeur)

124 Forme d'onde vocale

Figure 2

202 Unité de calcul de la durée des segments

204 Unité de génération de la prosodie

206 Durées des segments

208 Sortie prosodique

210 Unité de génération de la description des séquences

212 Description prosodique / linguistique des séquences

214 Unité de calcul des descriptions acoustiques

216 Unité de génération de formes d'onde

218 Forme d'onde vocale

220 Description acoustique des séquences

Figure 3

302 Parole

304 Base de données audio

306 Unité d'étiquetage phonétique / syntaxique

308 Base de données phonétiques

310 Unité de calcul des sorties souhaitées

312 Sortie d'entraînement

314 Unité de génération des entrées

316 Entrée d'entraînement

318 Réseau neural

320 Paramètres vocaux

Figure 4

402 Parole

404 Base de données audio

406 Unité d'étiquetage phonétique / syntaxique

408 Base de données phonétiques

410 Unité d'étiquetage prosodique

412 Base de données prosodiques

414 Unité de génération des sorties souhaitées

416 Sortie d'entraînement

418 Unité de génération des entrées

420 Entrée d'entraînement

422 Réseau neural

424 Paramètres vocaux avec information prosodique

Figure 5

502 Système statistique convertissant du texte en parole

504 Texte

506 Unité de conversion de texte en linguistique

508 Unité de génération de la prosodie

510 Description (phonétique) linguistique

512 Sortie prosodique

514 Unité de calcul de la durée des segments (réseau neural)

516 Durées des segments

518 Unité de génération de la description des séquences

520 Description linguistiques / prosodique des séquences

522 Unité de calcul des descriptions acoustiques (réseau neural)

524 Unité de génération de formes d'onde (vocodeur)

526 Forme d'onde vocale

528 Information prosodique prédéterminée

Figure 6

Au moins un des éléments: information sur le style et information sur la netteté

602 Déterminer l'information prosodique qui décrit le rythme et l'intonation de la parole à générer à partir d'au moins un des éléments: information sur le style et information sur la netteté

604 Utiliser un système statistique pour convertir l'information comprenant l'information linguistique et l'information prosodique en paramètres vocaux qui décrivent des parties de formes d'onde vocale

Paramètres vocaux avec information prosodique

Figure 7

702 Système statistique

704 Entrée linguistique

706 Paramètres vocaux

708 Description prosodique

710 Information sur le style / la netteté

712 Unité de détermination de la prosodie

REVENDICATIONS

1. Méthode pour fournir, en réponse à l'information comprenant l'information linguistique et au moins l'une des éléments: information sur le style et information sur la netteté, une génération efficace des paramètres vocaux prosodiquement améliorés comprenant les étapes de:

1A) détermination de l'information prosodique qui décrit le rythme et l'intonation de la parole à générer, à partir d'au moins un des éléments: information sur le style et information sur la netteté, et

1B) utilisation d'un système statistique pour convertir l'information comprenant l'information linguistique et l'information prosodique en paramètres vocaux, qui décrivent des parties de formes d'onde vocale.

2. Méthode selon la revendication 1, dans laquelle les paramètres vocaux sont des paramètres convenant pour être utilisés comme entrée dans un synthétiseur / générateur de formes d'onde, et, si sélectionné, où au moins un de 2A-2D:

2A) comprenant en outre la fourniture de paramètres vocaux à un synthétiseur / générateur de formes d'onde pour synthétiser la parole,

2B) dans laquelle un de:

2B1) le logiciel mettant en œuvre la méthode est intégré dans un microprocesseur,

2B2) le logiciel mettant en œuvre la méthode est intégré dans un processeur de signaux numériques,

2B3) la méthode est mise en œuvre par un circuit intégré spécifique à une application, et

2B4) la méthode est mise en œuvre par une combinaison d'au moins deux de 2B1-2B3, et

2C) dans laquelle l'information prosodique comprend au moins un de:

2C1) les emplacements des terminaisons de mot et le degré de disjonction entre les mots,

2C2) les emplacements des accents toniques et la forme des accents toniques, et

2C3) les emplacements des limites marquées dans le contour de la tonalité et la formes des limites,

2D) et si sélectionné à l'étape 2C, dans laquelle le format de l'information prosodique est un de:

2D1) l'information décrivant la proximité d'événements prosodiques marqués par rapport à des séquences définies entourant la séquence pour laquelle les paramètres du codeur sont générés,

2D2) l'information décrivant le temps séparant des événements prosodiques marqués de la séquence pour laquelle les paramètres du codeur sont générés,

2D3) l'information décrivant le temps séparant certains événements prosodiques marqués d'autres événements prosodiques marqués,

2D4) l'information décrivant le nombre d'événements prosodiques d'un type dans la période de temps séparant un événement prosodique d'un autre type et la séquence pour laquelle les paramètres du codeur sont générés,

2D5) l'information décrivant le nombre d'événements prosodiques d'un type survenant dans la période de temps séparant deux événements prosodiques d'un autre type, et

2D6) au moins deux de 2D1-2D5.

3. Méthode selon la revendication 1 dans laquelle les paramètres vocaux sont des durées de segments vocaux associés à des phones, et si sélectionné, au moins un de 3A-3C:

3A) comprenant en outre la fourniture des durées des

segments comme entrée supplémentaire à une méthode de génération des paramètres de séquences acoustiques,

3B) dans laquelle un de 3B1-3B4:

3B1) le logiciel mettant en œuvre la méthode est intégré dans un microprocesseur,

3B2) le logiciel mettant en œuvre la méthode est intégré dans un processeur de signaux numériques,

3B3) la méthode est mise en œuvre par un circuit intégré spécifique à une application, et

3B4) la méthode est mise en œuvre par une combinaison d'au moins deux de 3B1-3B3, et

3C) dans laquelle l'information prosodique comprend au moins un de 3C1-3C3:

3C1) les emplacements des terminaisons de mot et le degré de disjonction entre mots,

3C2) les emplacements des accents toniques et la forme des accents toniques, et

3C3) les emplacements des limites marquées dans le contour de la tonalité et la forme des limites, et si sélectionné, dans laquelle le format de l'information prosodique est un de 3C3a-3C3f:

3C3a) l'information décrivant la proximité d'événements prosodiques marqués par rapport à des segments définis entourant le segment pour lequel la durée est calculée,

3C3b) l'information décrivant le nombre de segments séparant des événements prosodiques marqués du segment pour lequel la durée est calculée,

3C3c) l'information décrivant le nombre de segments séparant des événements prosodiques marqués d'autres événements prosodiques marqués,

3C3d) l'information décrivant le nombre d'événements prosodiques d'un type quelconque dans les segments

séparant un événement prosodique d'un autre type et le segment pour lequel la durée est calculée,

3C3e) l'information décrivant le nombre d'événements prosodiques d'un type survenant dans les segments séparant deux événements prosodiques d'un autre type, et

3C3f) au moins deux de 3C3a—3C3e.

4. Méthode selon la revendication 1 dans laquelle au moins un de 4A-4C:

4A) le système statistique est un réseau neural,

4B) le système statistique est une unité d'arborescence des décisions, et

4C) le système statistique est une unité qui utilise un algorithme génétique.

5 Dispositif pour fournir, en réponse à l'information comprenant l'information linguistique et au moins un des éléments: information sur le style et information sur la netteté, une génération efficace de paramètres vocaux prosodiquement améliorés comprenant:

5A) une unité de détermination de la prosodie, couplée pour recevoir au moins un des éléments, information sur le style et information sur la netteté, générant l'information prosodique, décrivant le rythme et l'intonation de la parole à générer, et

5B) un système statistique, couplé à l'information comprenant l'information linguistique et prosodique, pour fournir des paramètres décrivant des parties de formes d'onde vocale.

6. Dispositif selon la revendication 5 dans lequel au moins un de 6A-6C:

6A) les paramètres vocaux sont des paramètres convenant pour être utilisés comme entrée dans un synthétiseur / générateur de formes d'onde, et si sélectionné, un de 6A1-6A2:

6A1) dans lequel le dispositif est un des éléments:

6A1a) un microprocesseur,

6A1b) un processeur de signaux numériques,

6A1c) un circuit intégré spécifique à une application, et

6A1d) une combinaison d'au moins deux de 6A1a-6A1c,

6A2) dans lequel l'information prosodique comprend au moins un de 6A2a-6A2c:

6A2a) les emplacements des terminaisons de mot et le degré de disjonction entre mots,

6A2b) les emplacements des accents toniques et la forme des accents toniques, et

6A2c) les emplacements des limites marquées dans le contour de la tonalité et la forme des limites, et si sélectionné pour l'étape 6A2, dans lequel le format de l'information prosodique est un de 6A2d-6A2i:

6A2d) l'information décrivant la proximité d'événements prosodiques marqués par rapport à des segments définis entourant le segment pour lequel la durée est calculée,

6A2e) l'information décrivant le nombre de segments séparant des événements prosodiques marqués du segment pour lequel la durée est calculée,

6A2f) l'information décrivant le nombre de segments séparant des événements prosodiques marqués d'autres événements prosodiques marqués,

6A2g) l'information décrivant le nombre d'événements prosodiques d'un type quelconque dans les segments séparant un événement prosodique d'un autre type et le segment pour lequel la durée est calculée,

6A2h) l'information décrivant le nombre d'événements prosodiques d'un type survenant dans les segments séparant deux événements prosodiques d'un autre type, et

6A2i) au moins deux de 6A2d-6A2h,

6B) dans lequel l'information prosodique comprend au moins un de 6B1-6B3:

6B1) les emplacements des terminaisons de mot et le degré de disjonction entre mots,

6B2) les emplacements des accents toniques et la forme des accents toniques, et

6B3) les emplacements des limites marquées dans le contour de la tonalité et la forme des limites, et dans lequel si sélectionné, un format de l'information prosodique est un de 6B4-6B9,

6B4) l'information décrivant la proximité d'événements prosodiques marqués par rapport à des séquences définies entourant la séquence pour laquelle les paramètres du codeur sont générés,

6B5) l'information décrivant le temps séparant des événements prosodiques marqués de la séquence pour laquelle les paramètres du codeur sont générés,

6B6) l'information décrivant le temps séparant des événements prosodiques marqués d'autres événements prosodiques marqués,

6B7) l'information décrivant le nombre d'événements prosodiques d'un type quelconque dans la période de temps séparant un événement prosodique d'un autre type et la séquence pour laquelle les paramètres du codeur sont générés,

6B8) l'information décrivant le nombre d'événements prosodiques d'un type survenant dans la période de temps séparant deux événements prosodiques d'un autre type, et

6B9) au moins deux de 6B4-6B8,

6C) dans lequel les paramètres vocaux sont des durées de segments vocaux associés à des phones, et si sélectionné, comprenant en outre une unité de génération des paramètres des

séquences acoustiques couplée pour recevoir les durées des segments vocaux.

7. Dispositif selon la revendication 5 dans lequel au moins un de 7A-7E:

7A) le système générant les paramètres vocaux fournit en outre les paramètres vocaux à un synthétiseur / générateur de formes d'onde pour synthétiser la parole,

7B) dans lequel le dispositif est un de 7B1-7B4:

7B1) un microprocesseur,

7B2) un processeur de signaux numériques,

7B3) un circuit intégré spécifique à une application, et

7B4) une combinaison d'au moins deux de 7B1-7B3,

7C) dans lequel le système statistique est un réseau neural,

7D) dans lequel le système statistique est une unité d'arborescence des décisions, et

7E) dans lequel le système statistique est une unité qui utilise un algorithme génétique.

8. Système convertissant du texte en paroles / système de synthèse vocale / système de dialogue ayant au moins un dispositif pour fournir, en réponse à l'information comprenant l'information linguistique, et au moins un des éléments: information sur le style et information sur la netteté, une génération efficace de paramètres vocaux, chaque dispositif comprenant:

8A) une unité de détermination de la prosodie, couplée pour recevoir au moins une des éléments: information sur le style et information sur la netteté, qui génère l'information prosodique qui décrit le rythme et l'intonation de la parole à générer, et

8B) un système statistique, couplé pour recevoir

l'information comprenant l'information linguistique et l'information prosodique, pour générer des paramètres vocaux qui décrivent des parties de formes d'onde vocale.

9. Système convertissant du texte en parole / système de synthèse vocale / système de dialogue selon la revendication 8 dans lequel au moins un de 9A-9G:

9A) le dispositif fournit des paramètres vocaux qui sont des paramètres convenant pour être utilisés comme entrée dans un codeur de formes d'onde, et si sélectionné, dans lequel le dispositif qui produit les paramètres vocaux qui sont des paramètres convenant pour être utilisés comme entrée dans un codeur de formes d'onde fournit en outre les paramètres vocaux à un synthétiseur de formes d'onde pour synthétiser la parole,

9B) dans lequel le dispositif est un des éléments:

9B1) un microprocesseur,

9B2) un processeur de signaux numériques,

9B3) un circuit intégré spécifique à une application, et

9B4) une combinaison d'au moins deux de 9B1-9B3,

9C) dans lequel l'information prosodique comprend au moins un de 9C1-9C3:

9C1) les emplacements des terminaisons de mot et le degré de disjonction entre mots,

9C2) les emplacements des accents toniques et la forme des accents toniques, et

9C3) les emplacements des limites marquées dans le contour du ton et la forme des limites, et si pris en considération, pour l'étape 9C, dans lequel le format de l'information prosodique est un de 9C3a-9C3f:

9C3a) l'information décrivant la proximité

d'événements prosodiques marqués par rapport à des séquences définies entourant la séquence pour laquelle les paramètres du codeur sont générés,

9C3b) l'information décrivant le temps séparant des événements prosodiques marqués de la séquence pour laquelle les paramètres du codeur sont générés,

9C3c) l'information décrivant le temps séparant des événements prosodiques marqués d'autres événements prosodiques marqués,

9C3d) l'information décrivant le nombre d'événements prosodiques d'un type quelconque dans la période de temps séparant un événement prosodique d'un autre type et la séquence pour laquelle les paramètres du codeur sont générés,

9C3e) l'information décrivant le nombre d'événements prosodiques d'un type survenant dans la période de temps séparant deux événements prosodiques d'un autre type, et

9C3f) au moins deux de 9C3a-9C3e,

9D) dans lequel le dispositif est un de 9D1-9D4:

9C1) un microprocesseur,

9C2) un processeur de signaux numériques,

9C3) un circuit intégré spécifique à une application, et

9C4) une combinaison d'au moins deux de 9C1-9C3,

9E) dans lequel au moins un des dispositifs pour fournir une génération efficace de paramètres vocaux prosodiquement améliorés est un réseau neural,

9F) dans lequel au moins un des dispositifs pour fournir une génération efficace de paramètres vocaux prosodiquement améliorés est une unité d'arborescence de décisions, et

9G) dans lequel au moins un des dispositifs pour fournir

une génération efficace de paramètres vocaux prosodiquement améliorés est une unité qui utilise un algorithme génétique.

10. Système convertissant du texte en parole / système de synthèse vocale / système de dialogue selon la revendication 8 dans lequel au moins un des dispositifs génère des paramètres vocaux qui sont des durées de segments vocaux associés à des phones, et si sélectionné, au moins un de 10A-10B:

10A) le dispositif qui génère des durées de segments fournit en outre les durées de segments comme entrée supplémentaire à un dispositif de génération des paramètres de séquences acoustiques, et

10B) l'information prosodique comprend au moins un des éléments:

10B1) les emplacements des terminaisons de mot et le degré de disjonction entre mots,

10B2) les emplacements des accents toniques et la forme des accents toniques, et

10B3) les emplacements des limites marquées dans le contour de la tonalité et la forme des limites, et si en outre sélectionné à l'étape 10B, dans lequel le format de l'information prosodique est un de 10B3a-10B3f:

10B3a) l'information décrivant la proximité d'événements prosodiques marqués par rapport à des segments entourant le segment pour lequel la durée est calculée,

10B3b) l'information décrivant le nombre de segments séparant des événements prosodiques marqués du segment pour lequel la durée est calculée,

10B3c) l'information décrivant le nombre de segments séparant des événements prosodiques marqués d'autres événements prosodiques marqués,

10B3d) l'information décrivant le nombre

d'événements prosodiques d'un type dans les segments séparant un événement prosodique d'un autre type et le segment pour lequel la durée est calculée,

10B3e) l'information décrivant le nombre d'événements prosodiques d'un type survenant dans les segments séparant deux événements prosodiques d'un autre type, et

10B3f) au moins deux de 10B3a-10B3e.

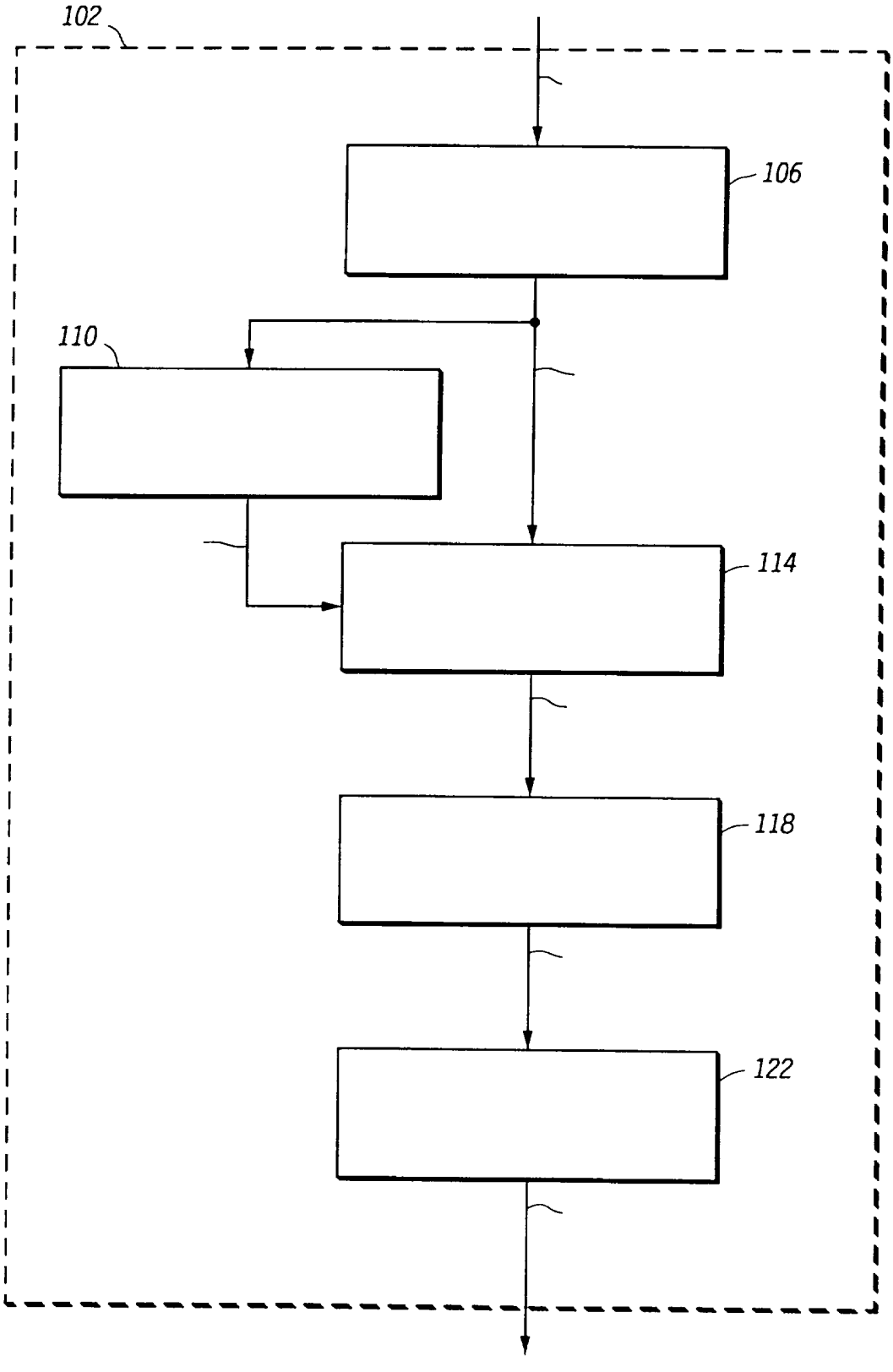

FIG.1

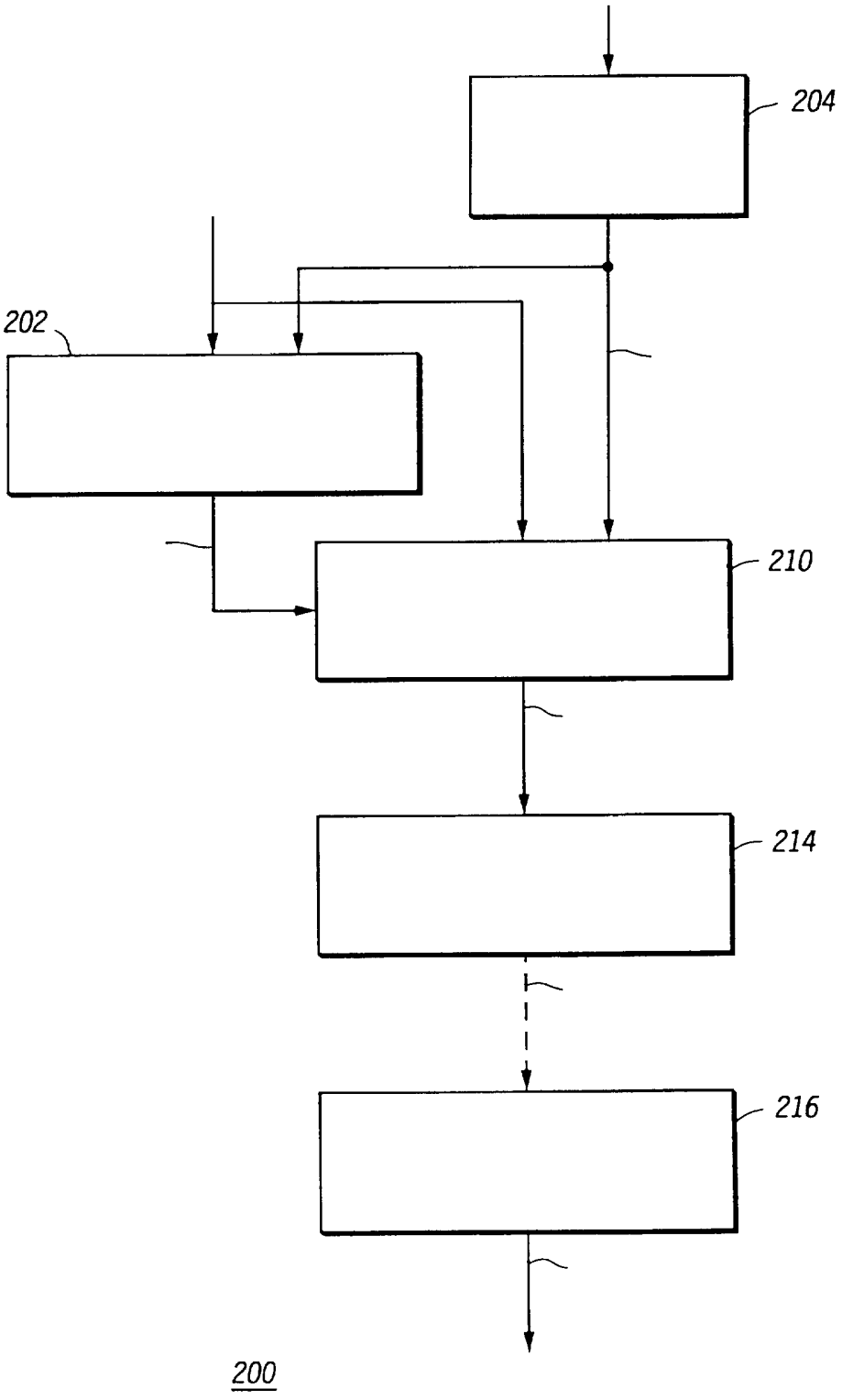

FIG.2

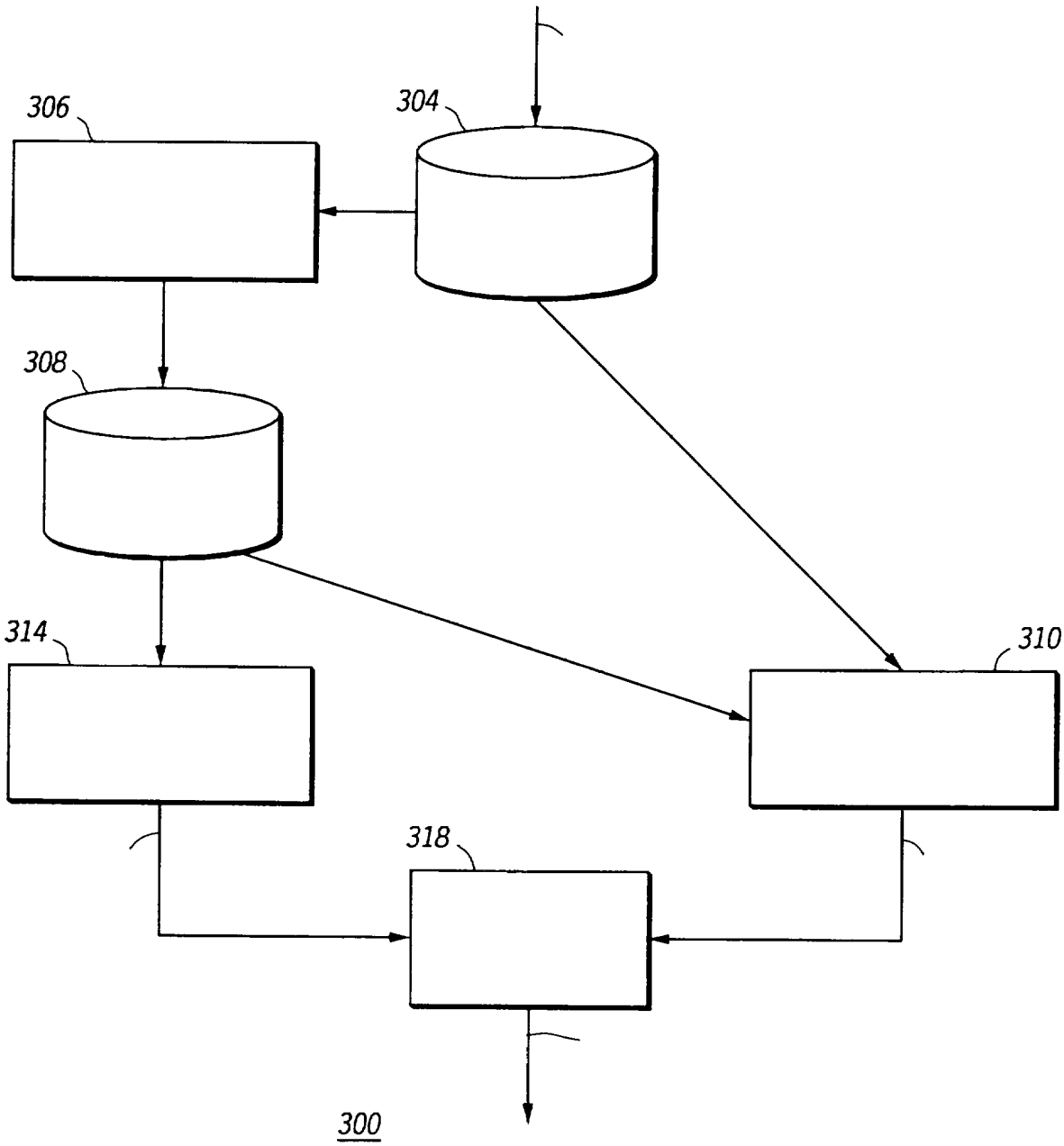

FIG.3

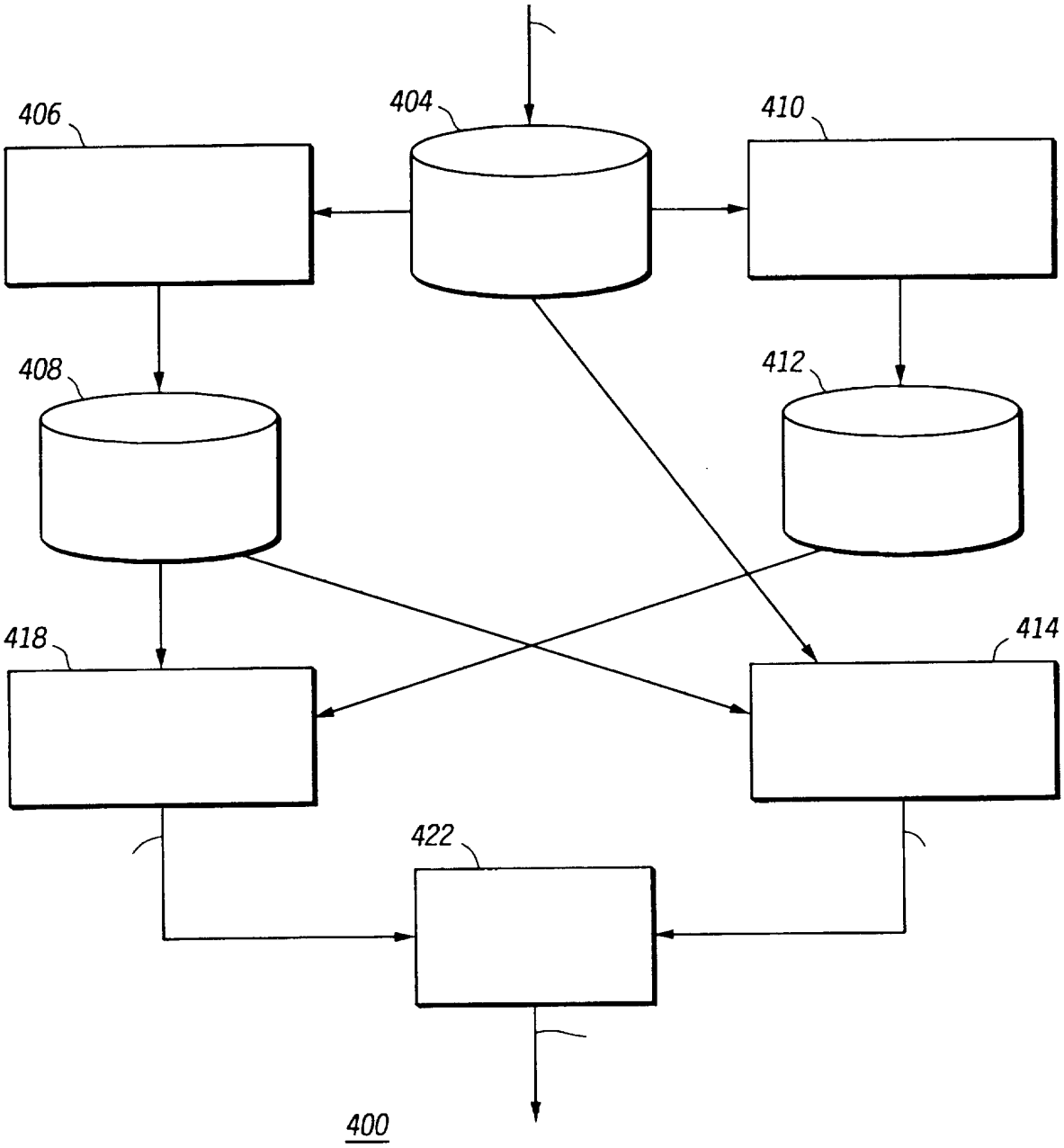

FIG.4

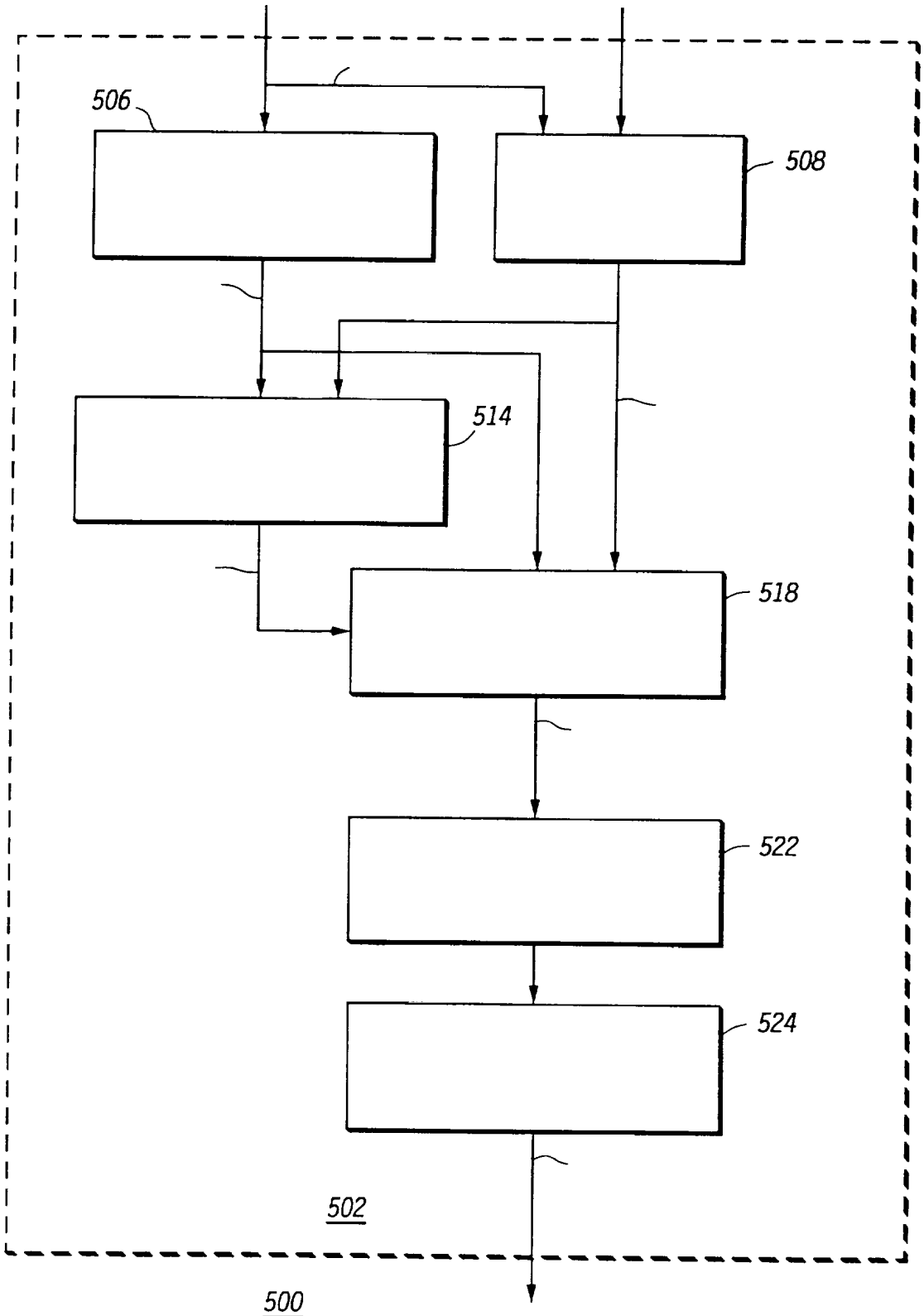

FIG.5

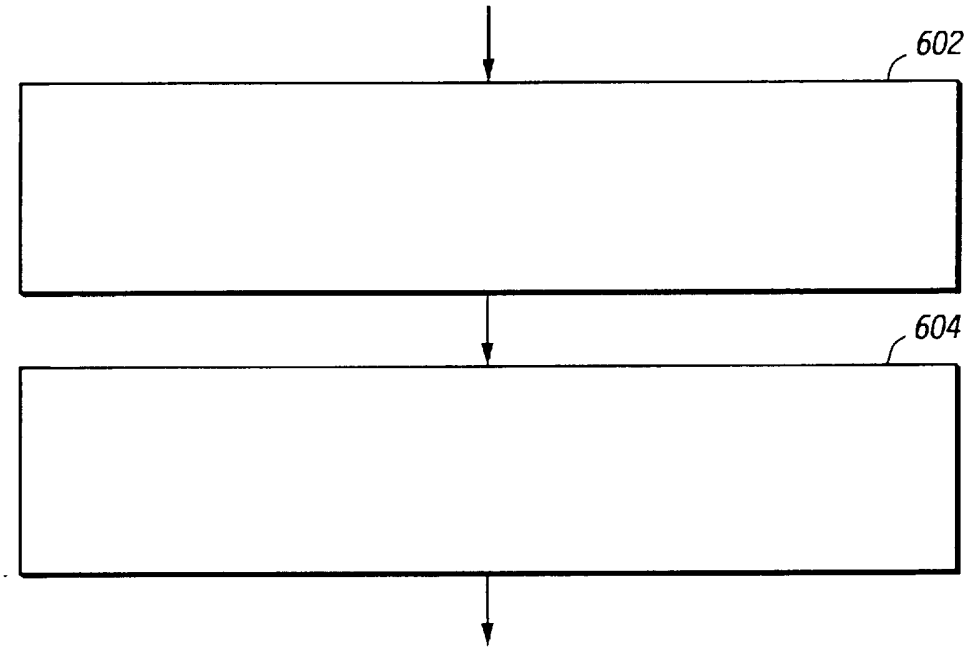

FIG. 6

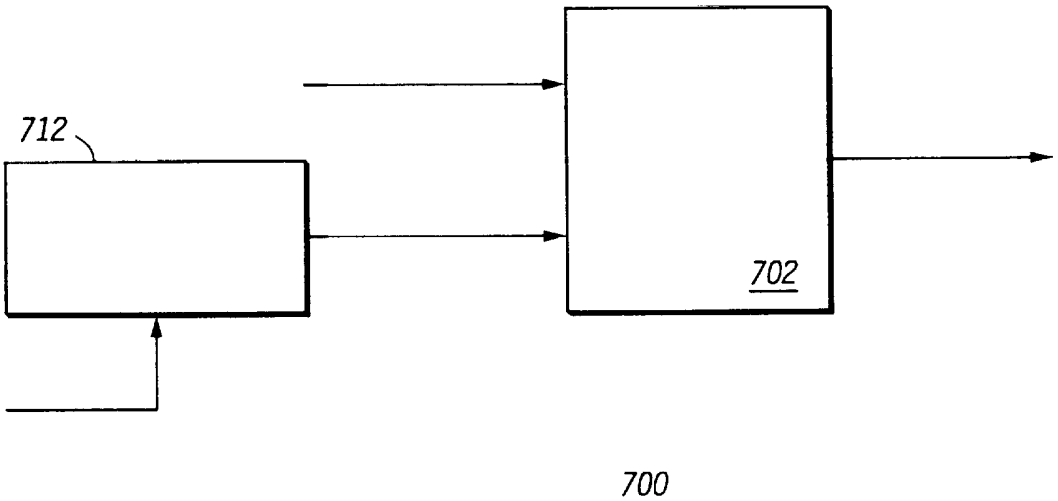

FIG. 7

Numero de la demande nationale

Office européen des brevets

RAPPORT DE RECHERCHE

établi en vertu de l'article 21 § 1 et 2 de la loi belge sur les brevets d'invention du 28 mars 1984

BO 7032
BE 9800314

DOCUMENTS CONSIDERES COMME PERTINENTS

Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes	Revendication concernée	CLASSEMENT DE LA DEMANDE (Int.Cl.6)
X	XUEDONG HUANG ET AL: "Whistler: a trainable text-to-speech system" PROCEEDINGS ICSLP 96. FOURTH INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING (CAT. NO.96TH8206), PROCEEDING OF FOURTH INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING. ICSLP '96, PHILADELPHIA, PA, USA, 3-6 OCT. 1996, pages 2387-2390 vol.4, XP002101886 ISBN 0-7803-3555-4, 1996, New York, NY, USA, IEEE, USA * le document en entier *	1,5,8	G10L5/04
A		2-4,6,7, 9,10	

X	LARREUR D ET AL: "LINGUISTIC AND PROSODIC PROCESSING FOR A TEXT-TO-SPEECH SYNTHESIS SYSTEM" PROCEEDINGS OF THE EUROPEAN CONFERENCE ON SPEECH COMMUNICATION AND TECHNOLOGY (EUROSPEECH), PARIS, SEPT. 26 - 28, 1989, vol. 1, no. CONF. 1, 26 septembre 1989, pages 510-513, XP000209680 TUBACH J P;MARIANI J J * abrégé * * alinéa 2.1 * * alinéa 4.1.2 * * alinéa 4.2 *	1,5,8	DOMAINES TECHNIQUES RECHERCHES (Int.Cl.6) G10L
A		2-4,6,7, 9,10	

	-/--		

Date d'achèvement de la recherche	Examinateur
4 mai 1999	Wanzeele, R

CATEGORIE DES DOCUMENTS CITES

X : particulièrement pertinent à lui seul
Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie
A : arrière-plan technologique
O : divulgation non-écrite
P : document intercalaire

T : théorie ou principe à la base de l'invention
E : document de brevet antérieur, mais publié à la date de dépôt ou après cette date
D : cité dans la demande
L : cité pour d'autres raisons

& : membre de la même famille, document correspondant

EPO FORM 1503 03.82 (P04C48)

Office européen des brevets

RAPPORT DE RECHERCHE

établi en vertu de l'article 21 § 1 et 2 de la loi belge sur les brevets d'invention du 28 mars 1984

BO 7032
BE 9800314

DOCUMENTS CONSIDERES COMME PERTINENTS

Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes	Revendication concernée	CLASSEMENT DE LA DEMANDE (Int.Cl.6)
X	LOPEZ-GONZALO E ET AL: "DATA-DRIVEN JOINT FO AND DURATION MODELING IN TEXT TO SPEECH CONVERSION FOR SPANISH" PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, SIGNAL PROCESSING (ICASSP), SPEECH PROCESSING 1. ADELAIDE, APR. 19 - 22, 1994, vol. 1, 19 avril 1994, pages I-589-I-592, XP000529432 INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS * alinéa 2 * * alinéa 4 *	1,5,8	
A		2-4,6,7, 9,10	

X,P	EP 0 833 304 A (MICROSOFT CORP) 1 avril 1998 * le document en entier *	1,5,8	

A	BLACK A W ET AL: "Generating F/sub 0/ contours from ToBI labels using linear regression" PROCEEDINGS ICSLP 96. FOURTH INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING (CAT. NO.96TH8206), PROCEEDING OF FOURTH INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING. ICSLP '96, PHILADELPHIA, PA, USA, 3-6 OCT. 1996, pages 1385-1388 vol.3, XP002101887 ISBN 0-7803-3555-4, 1996, New York, NY, USA, IEEE, USA * alinéa 1 * * alinéa 2 * * alinéa 3 *	1,5,8	DOMAINES TECHNIQUES RECHERCHES (Int.Cl.6)

	-/--		

1

Date d'achèvement de la recherche	Examinateur
4 mai 1999	Wanzeele, R

CATEGORIE DES DOCUMENTS CITES

X : particulièrement pertinent à lui seul
Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie
A : arrière-plan technologique
O : divulgation non-écrite
P : document intercalaire

T : théorie ou principe à la base de l'invention
E : document de brevet antérieur, mais publié à la date de dépôt ou après cette date
D : cité dans la demande
L : cité pour d'autres raisons

& : membre de la même famille, document correspondant

EPO FORM 1503 03.82 (P04C48)

Numero de la demande nationale

Office européen des brevets

RAPPORT DE RECHERCHE

établi en vertu de l'article 21 § 1 et 2 de la loi belge sur les brevets d'invention du 28 mars 1984

BO 7032
BE 9800314

DOCUMENTS CONSIDERES COMME PERTINENTS

Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes	Revendication concernée	CLASSEMENT DE LA DEMANDE (Int.Cl.6)
A	WO 95 30193 A (MOTOROLA INC) 9 novembre 1995 * revendications 1-8 * -----	1,5,8	

DOMAINES TECHNIQUES RECHERCHES (Int.Cl.6)

Date d'achèvement de la recherche	Examinateur
4 mai 1999	Wanzeele, R

CATEGORIE DES DOCUMENTS CITES

X : particulièrement pertinent à lui seul
Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie
A : arrière-plan technologique
O : divulgation non-écrite
P : document intercalaire

T : théorie ou principe à la base de l'invention
E : document de brevet antérieur, mais publié à la date de dépôt ou après cette date
D : cité dans la demande
L : cité pour d'autres raisons

& : membre de la même famille, document correspondant

1

EPO FORM 1503 03.82 (P04C48)

ANNEXE AU RAPPORT DE RECHERCHE RELATIF A LA DEMANDE DE BREVET BELGE NO.

BO 7032
BE 9800314

La présente annexe indique les membres de la famille de brevets relatifs aux documents brevets cités dans le rapport de recherche visé ci-dessus.
Lesdits membres sont contenus au fichier informatique de l'Office européen des brevets à la date du
Les renseignements fournis sont donnés à titre indicatif et n'engagent pas la responsabilité de l'Office européen des brevets.

04-05-1999

Document brevet cité au rapport de recherche	Date de publication	Membre(s) de la famille de brevet(s)	Date de publication
EP 0833304 A	01-04-1998	JP 10116089 A	06-05-1998
WO 9530193 A	09-11-1995	AU 675389 B	30-01-1997
		AU 2104095 A	29-11-1995
		CA 2161540 A	09-11-1995
		CN 1128072 A	31-07-1996
		EP 0710378 A	08-05-1996
		FI 955608 A	22-11-1995
		JP 8512150 T	17-12-1996
		US 5668926 A	16-09-1997

EPO FORM P0463