Vocollect

#1 in the Business of Voice™

Choix de technologie de reconnaissance de la parole pour l'entrepôt

Août 2010

Un livre blanc de Vocollect

Table de Matières

Résumé

Le présent document fait dans un premier temps le point sur certains éléments fondamentaux des systèmes vocaux par ordinateur, avant d'aborder les choix technologiques qui doivent être pris en fonction des besoins des opérateurs amenés à utiliser ce type de systèmes en entrepôt et de calculer le coût des erreurs commises par les modules de reconnaissance vocale dans ce type d'environnement. En ce qui concerne le choix de la technologie — système mono-locuteur (avec formation) ou multi-locuteurs (sans formation) —, nous démontrerons que le temps consacré à la formation du module de reconnaissance vocale est généralement rentabilisé en quelques jours grâce aux gains de performances qu'il apporte.

Qu'attend-on d'un système de reconnaissance vocale utilisé en entrepôt?

Avant d'entrer dans les détails techniques, nous allons nous intéresser aux attentes des opérateurs qui utilisent un module de reconnaissance vocale en entrepôt. Dans un monde parfait, le principal objectif serait le suivant : « comprendre immédiatement et correctement tout ce que l'utilisateur souhaite que vous entendiez et rien de qu'il ne souhaite pas entendre ». Comme nous le verrons par la suite, il est impossible d'atteindre cet idéal mais nous allons tenter de nous en approcher au plus près. Compte tenu de cette contrainte, un système vocal en entrepôt doit :

- Fonctionner efficacement dans un large éventail d'environnements professionnels pouvant passer rapidement de « très calme » à « très bruyant ».
- Fonctionner efficacement avec le plus large éventail de profils d'utilisateurs, indépendamment de leur sexe, de leur langue maternelle, de leur accent, des modèles vocaux, etc.
- Réagir "instantanément" aux ordres de l'opérateur afin d'éliminer à la fois les coûts et la frustration que provoquent les retards.
- Minimiser le "coût total d'utilisation", y compris le temps perdu en préparant le système avant son utilisation, et à cause des erreurs et des retards générés par le module pendant que l'opérateur travaille.

Comment fonctionnent les systèmes vocaux?

Les systèmes vocaux informatisés associent des « modèles ». La séquence des événements (sous une forme très simplifiée) est la suivante:

- Le module de reconnaissance charge un ensemble de « modèles de références sonores » qui représentent soit des mots, soit des parties de mots (« phonèmes ») dont l'application sait que l'utilisateur va les prononcer. Les spécialistes utilisent l'expression « vocabulaire actif » pour désigner la liste des mots que l'utilisateur peut prononcer à tout moment, et le terme « vocabulaire » pour désigner la liste de tous les mots que l'opérateur peut prononcer lorsqu'il utilise l'application.

- L'application transmet alors au module de reconnaissance un son inconnu qui représente soit un mot ou une série de mots (une « expression ») prononcés par l'opérateur, soit un son « étranger », voire une expression contaminée par des sons étrangers.
- Le module de reconnaissance vocale "classe" le son inconnu et calcule la meilleure association possible entre le son inconnu et une série d'un ou plusieurs modèles de référence. Par exemple, si les modèles de référence représentent des chiffres, le module de reconnaissance vocale peut indiquer que le son inconnu correspond à la suite de chiffres 123. Le module de reconnaissance vocale « calcule » également la proximité du son inconnu par rapport aux modèles de référence. Si le « score » est médiocre, l'application peut alors décider que le son inconnu correspond probablement à un bruit extérieur et non à une expression prononcée par l'opérateur. Dans ce cas, l'application ignorera l'information signalée.

Qu'est ce qui rend la reconnaissance vocale si complexe?

La reconnaissance vocale (qu'elle soit humaine ou informatique) serait un problème relativement simple à résoudre si tout le monde s'exprimait de façon identique et homogène. Mais ce n'est pas le cas. Les modes d'expression sont comme les flocons de neige, uniques. Ainsi, Monsieur A prononcera le mot « un » différemment de Monsieur B. Pire encore, si Monsieur A répète un mot plusieurs fois de suite, il sera à chaque fois légèrement différent. La tâche des modules de reconnaissance vocale est en outre compliquée par d'autres facteurs :

- Lorsque nous parlons sans interruption entre deux mots, la façon dont nous prononçons chaque mot est affectée par le mot qui précède et par le mot qui suit. Ce phénomène appelé la co-articulation affecte également les sons à l'intérieur des mots, de sorte qu'un son donné sera prononcé différemment en fonction du son qui le précède et du son qui le suit.
- Les expressions peuvent être corrompues par les bruits de fond.
- L'application peut transmettre au module de reconnaissance vocale un bruit extérieur qui ne contient aucun son prononcé par l'utilisateur.
- Les sons générés par l'utilisateur peuvent ne pas être correctement acheminés au module de reconnaissance vocale (par exemple, dans le cas d'une liaison téléphonique).

Lorsqu'un système de reconnaissance vocale commet une erreur, ce qui arrive aux hommes comme aux machines, elle peut être de trois types : insertion, suppression et substitution.

	Exemple	
Type d'erreur	**L'opérateur dit:**	**Le module de reconnaissance vocale pense que l'opérateur a dit:**
Insertion	<rien>	Un
	Un cinq trois	Un cinq neuf trois
Suppression	Un	<rien>
	Un cinq trois	Un trois
Substitution	Un cinq trois	Un neuf trois
	Cinq	Neuf

Comme on peut l'imaginer, optimiser les performances d'un système de reconnaissance en fonction de ces trois types d'erreur n'est pas simple. Par exemple, « régler » le module de reconnaissance pour qu'il soit moins susceptible aux erreurs d'insertion peut se traduire par une augmentation du risque de suppressions.

Simplifier le problème

Pour concevoir un équipement intégrant un module de reconnaissance vocale, l'objectif est de minimiser les erreurs en faisant en sorte que le problème auquel est confronté le module soit le plus simple possible. Il existe plusieurs façons d'y parvenir:

- En limitant le vocabulaire du module de reconnaissance — Dans un système de dictée, les contraintes sont minimes : le locuteur peut dire pratiquement ce qu'il veut, quand il veut. Dans une application industrielle, si le système demande à l'utilisateur d'entrer une quantité, nous pouvons simplifier la tâche du module de reconnaissance en lui apprenant à reconnaître exclusivement les séries de chiffres.
- En insistant sur un environnement de travail qui limite le bruit de fond — Ceci est raisonnable pour un système de dictée mais difficile à mettre en œuvre dans le cas d'un système industriel conçu pour être utilisé dans un entrepôt ou une usine. Dans ce type d'environnement, le mieux que nous puissions faire est de minimiser le bruit de fond au moyen de microphones à « annulation de bruit » et en intégrant des algorithmes dans le module de reconnaissance pour faire en sorte de minimiser les effets de la « contamination sonore ».
- En permettant au système d'utiliser ce qu'il sait de l'utilisateur — Une fois de plus, chaque système pourra utiliser cette connaissance en fonction de ses propres capacités.
    - o De façon réaliste, un système téléphonique, tel que ceux qu'utilisent les compagnies aériennes, ne peut exiger des utilisateurs qu'ils s'identifient; les transactions sont si courtes que le système ne peut apprendre que très peu de choses sur les habitudes vocales de l'utilisateur.

  - Avant qu'ils puissent l'utiliser, un système de dictée peut demander aux nouveaux utilisateurs de lui parler, généralement en lisant un ou plusieurs scripts figés totalisant de cinq à quinze minutes. Ceci permet au module de reconnaissance d'enregistrer des informations concernant le « type de voix » de l'utilisateur (grave ou aigue, par exemple) et son accent.
  - Un système doté d'un "vocabulaire réduit", tel que ceux utilisés en entrepôt peut exiger des nouveaux utilisateurs qu'ils prononcent plusieurs mots précis utilisés dans leurs tâches. Le système peut alors créer des « modèles vocaux » spécifiques à chaque utilisateur pour chacun des mots de ce vocabulaire. On dit d'un système qui utilise ce type de connaissance à propos des utilisateurs qu'il est « mono-locuteur » ou « entièrement formé ».
- Certains systèmes de reconnaissance exigent des utilisateurs qu'ils prononcent des « mots d'ancrage » avant — parfois même avant et après — chaque mot (par exemple : "start 1 2 3 stop"). S'ils peuvent améliorer certains aspects des performances d'un module de reconnaissance vocale médiocre, les mots d'ancrage augmentent de façon considérable le nombre de mots à prononcer, ce qui pénalise considérablement la productivité. Chez Vocollect, nous avons fait en sorte que notre système de reconnaissance vocale assure des performances optimales sans recourir à des mots d'ancrage.
- Les systèmes de reconnaissance les plus avancés apprennent à connaître l'utilisateur pendant qu'il travaille. Ils utilisent ces connaissances pour améliorer les performances. Vocollect appelle cette technologie « reconnaissance adaptative ». La société l'incorpore dans ses produits depuis 2006 et continue à l'améliorer.

Comment choisir un module de reconnaissance vocale pour entrepôt?

Avant de concevoir un système de reconnaissance vocale pour entrepôts, plusieurs décisions s'imposent : certaines sont faciles à prendre, d'autres méritent réflexion. Il convient ainsi :

- De limiter le vocabulaire en fonction de la tâche à accomplir ;
- D'utiliser des équipements (microphones, par exemple) et des algorithmes permettant de minimiser le bruit de fond ;
- D'éviter l'utilisation de mots d'ancrage ;
- D'adapter le système à l'utilisateur.

La dernière grande décision concerne le mode d'utilisation du système : mono ou multi-locuteurs. Examinons les caractéristiques des applications en entrepôt qui ont un impact sur le choix de la technologie :

- Un vocabulaire restreint et figé – cette caractéristique, absente de nombreuses autres applications vocales, permet d'utiliser un système mono-locuteur;

- Un grand nombre de transactions par utilisateur et un taux de transactions élevé — comme nous le verrons ci-après, ces exigences font de la précision de la reconnaissance et de la rapidité de réponse deux critères déterminants, dans la mesure où les erreurs et les retards peuvent s'accumuler rapidement;
- Des employés qui parlent plusieurs langues, différentes de la langue nationale – une large couverture linguistique, des niveaux de langage et des formes linguistiques est requise;
- Un fonctionnement adapté à tous les utilisateurs – aucun autre mode de saisie des données n'est disponible;
- Expressions courtes et mots courts, prononcés dans des environnements bruyants — les expressions et les mots courts peuvent entraîner des erreurs d'insertion, ce qui implique une étanchéité optimale aux bruits de fond;
- Modification des modèles vocaux et des bruits de fond – les modèles vocaux des utilisateurs changent pour de multiples raisons : par exemple, à cause de la fatigue qui augmente au fil de la journée.

Le tableau suivant dresse la liste des principaux objectifs de conception applicables aux modules de reconnaissance vocale utilisés dans des entrepôts, en indiquant pour chacun s'il est convient au mode mono-locuteur ou multi-locuteurs.

Objectif	Système "mono-locuteur"	Système "multi-locuteurs"
Minimise le délai de formation avant utilisation		✔
Maximise la précision et la productivité de l'opérateur	✔	
Fonctionne indépendamment de la langue	✔	
Imperméabilité à l'accent, au timbre de la voix, au sexe du locuteur, etc.	✔	
Maximise le rejet des bruits de fond	✔	
Maximise le niveau de satisfaction de l'utilisateur	✔	
Maximise les avantages de la reconnaissance adaptative	✔	

Calculer l'impact des coûts

Tout calcul visant à choisir un système de reconnaissance vocale doit évidemment comparer le coût de la formation par rapport au gain de performances enregistré. S'il est impossible de mesurer immédiatement les avantages ou le coût de la satisfaction des utilisateurs, nous pouvons estimer assez rapidement le coût de la formation préalable à l'utilisation des systèmes, ainsi que celui des erreurs enregistrées en cours d'utilisation.

Nous nous appuyons sur les éléments suivants:

Coût de l'opérateur (salaire et prestations):	20 dollars de l'heure
Durée d'utilisation du système vocal:	8 hr/jour, 360 jours par an
Rythme des transactions (nb de préparations par heure) :	200
Mots prononcés par transaction:	4
Durée des erreurs de reconnaissance:	3,5 secondes
Temps de formation avant utilisation:	20 minutes

Notes sur ces hypothèses:

- L'utilisation du système sera inférieure si l'entrepôt fonctionne 5 jours par semaine et non 7 ; mais elle peut être sensiblement supérieure lorsque plusieurs équipes se succèdent.
- Le rythme des transactions correspond à une tâche de prélèvement standard. Pour les autres tâches, le rythme des transactions peut être plus élevé ou plus faible.
- Le nombre de mots prononcés (ici quatre) représente la quantité minimale pour toute opération en entrepôt : il s'agit d'une préparation standard (« sans exception ») où l'opérateur prononce trois chiffres de contrôle pour confirmer l'emplacement de la préparation et un chiffre unique pour confirmer la quantité prélevée.
- Vocollect a mesuré, à l'aide d'observations, le temps que représente une erreur de reconnaissance. Le délai de récupération en cas d'erreur peut être supérieur au chiffre indiqué (3,5 secondes), certains opérateurs expérimentés étant capables de continuer à travailler tout en rattrapant une erreur.
- Le temps de formation pré-utilisation correspond à un système vocal d'entrepôt de Vocollect.

Sur la base de ces hypothèses, nous pouvons calculer :

Le nombre de mots prononcés par jour : 200 x 8 x 4 = 6 400

Depuis plusieurs années, Vocollect consacre beaucoup de temps et d'énergie à la création d'une base de données intégrant de nombreuses heures de modèles vocaux fournis par de nombreux opérateurs qui utilisent nos systèmes vocaux en entrepôt. Nous nous appuyons sur cette base de données pour calculer au mieux les taux d'erreurs que les utilisateurs peuvent enregistrer en conditions réelles. Dès que des améliorations sont apportées à nos algorithmes de reconnaissance vocale, nous les testons par rapport à cette base de données avant de vérifier les performances en procédant à des tests sur le terrain. L'utilisation d'un ensemble de données spécifiquement enregistrées pour des applications en entrepôt nous fournit une excellente corrélation entre les améliorations effectuées en laboratoire et les résultats issus du terrain.

S'il est démontré de façon tangible par Vocollect, il est encore plus difficile de mesurer l'impact des mauvaises performances du système de reconnaissance vocale sur le niveau de satisfaction des utilisateurs, sur les performances d'ensemble et sur le mauvais traitement infligé au système. Nous considérons ces problèmes comme réels et importants, ce qui nous amène non seulement à chercher en permanence à améliorer les performances de notre système de reconnaissance, mais également à nous concentrer sur les caractéristiques des produits et des services, au-delà du cadre des algorithmes du module de reconnaissance, par exemple les questions liées à la conception des casques et à la formation des utilisateurs.

Résultats des tests

Vocollect a récemment procédé à une campagne de tests en s'appuyant sur sa base de données internes pour établir une comparaison entre son module de reconnaissance vocale mono-locuteur et d'autres systèmes multi-locuteurs concurrents (y compris le module le plus couramment déployé dans les systèmes de reconnaissance vocale concurrents utilisés en entrepôts).

Les résultats de ces tests indiquent que pour une utilisation en entrepôt, l'augmentation du taux d'erreurs entre un système mono-locuteur et un système multi-locuteurs représente plusieurs points de pourcentage. Ainsi, pour les utilisateurs possédant un accent modéré ou fort, le taux d'erreur passe de 6 % à 20 %. Comme le montre le graphique ci-dessous, l'augmentation de coût par unité vocale ainsi enregistrée — sur la base des hypothèses indiquées ci-précédemment — peut aisément dépasser 1 000 dollars par an, même dans le cas d'équipes uniques, voire atteindre plusieurs milliers de dollars par an avec plusieurs équipes.

Il est important de noter que l'analyse fournie dans le présent document s'applique uniquement à la reconnaissance vocale en entrepôt. Vocollect, par exemple, a mis au point et déployé un module de reconnaissance multi-locuteurs pour environnements médicaux où les exigences sont sensiblement différentes. Mais nous restons intimement convaincus que notre module de reconnaissance vocale mono-locuteur est le produit idéal pour les entrepôts.

À chaque augmentation du nombre d'erreurs de 1 % — c'est à dire à chaque fois que le nombre d'erreurs commises par le système de reconnaissance vocale passe de 1 à 2 par 100 mots —, on obtient:

Augmentation du nombre d'erreurs par jour = 6 400 x 1 % = 64

Temps perdu par jour = 64 x 3,5 = 224 secondes = 3,7 minutes

Temps perdu par an = 3,7 x 360 / 60 = 22,4 heures

Coût par an = 22,4 x 20 $ = 450 dollars

Notons que les erreurs peuvent appartenir aux trois catégories décrites précédemment, les systèmes de reconnaissance multi-locuteurs étant tout particulièrement sensibles aux erreurs d'insertion.

Par conséquent, si l'utilisation d'un module de reconnaissance vocale mono-locuteur abaisse le taux d'erreur de seulement 1 %:

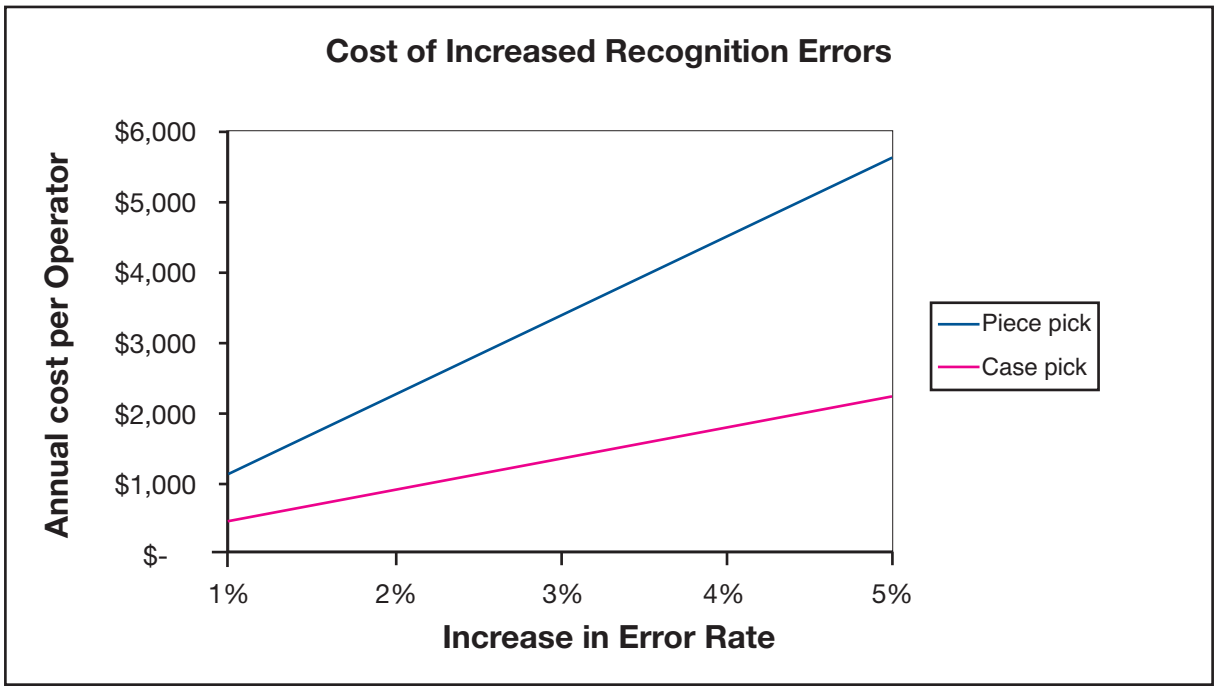
La période de rentabilisation de la formation pré-utilisation sera inférieure à 6 jours.

Avec à la clé des économies de 450 dollars par opérateur et par an.

Comme nous allons le voir ci-dessous, le taux de 1 % est une estimation très prudente de la différence entre un système multi-locuteurs et un système mono-locuteur.

Mesure des performances

On sait qu'il est difficile de mesurer de façon pertinente le taux d'erreurs d'un système de reconnaissance vocale. Certaines entreprises affirment que leur système est « précis à 99,7 % ». Il est sans aucun doute possible de concevoir un test qui permet d'atteindre un taux d'exactitude de 99,7 % (taux d'erreur de 0,3 %) pour n'importe quel module de reconnaissance. Mais il est également possible de concevoir un test différent qui indiquerait des taux d'erreur de 10 % ou plus pour le même système (soit des performances trente fois moins bonnes), même pour une tâche apparemment aussi simple que reconnaître des mots comme "oui" ou "non". Les taux de précision des modules de reconnaissance vocale doivent par conséquent être considérés avec une certaine prudence. Le mieux qu'un fabricant puisse faire est de mesurer, à l'aide d'un important volume de données, les résultats qu'obtiendra l'utilisateur du système — ce qui est à la fois fastidieux et onéreux. Ensuite, il est nécessaire de créer un environnement de test reproduisant aussi fidèlement que possible les résultats obtenus en conditions réelles. On peut ensuite espérer que si des modifications sont apportées au module de reconnaissance vocale et qu'une amélioration est enregistrée au niveau des résultats des tests, l'utilisateur bénéficiera sur le terrain d'améliorations comparables. Il est nécessaire de procéder régulièrement à un ré-étalonnage en recueillant de nouvelles informations issues de conditions « réelles », en utilisant la dernière version du module de reconnaissance vocale et en comparant à nouveau les résultats par rapport aux résultats fournis par l'environnement de test. Même en prenant des mesures aussi rigoureuses, il reste très difficile d'annoncer un niveau de précision et d'exactitude crédible. Une amélioration substantielle de la suppression du bruit de fond, par exemple, n'aura aucun impact sur un environnement qui n'est pas sujet à ce type de phénomène.



Hausse du coût annuel par opérateur en fonction de l'augmentation du nombre d'erreurs (sur la base de 500 préparations à l'unité par heure (courbe Piece pick en bleu) contre 200 pour la préparation de colis (courbe Case pick en fuchsia).

Conclusions

En ce qui concerne les entrepôts, un système de reconnaissance vocale multi-locuteurs présente l'avantage de ne pas nécessiter d'investissement initial pour former les opérateurs avant de pouvoir être utilisé. En revanche, un système mono-locuteur générera un retour sur investissement nettement supérieur à long terme. Les caractéristiques de l'application permettent d'utiliser un système mono-locuteur en entrepôt tout en constituant un choix évident pour quiconque conçoit un module de reconnaissance spécifiquement destiné à ce type d'environnement. En premier lieu, ce système garantit une plus grande précision parce qu'il est capable de différencier et de reconnaître la façon dont chaque utilisateur prononce chaque mot — il n'est pas nécessaire de prévoir les variations de prononciation associées à une région ou une langue. Cette spécialisation permet par ailleurs de rejeter les sons qui ne doivent pas être reconnus, ce qui évite les onéreuses erreurs d'insertion.

De plus, les modifications enregistrées par les modèles vocaux au fil du temps font de la reconnaissance adaptative un choix évident pour les applications en entrepôt. Si les modules de reconnaissance mono- et multi-locuteurs peuvent être adaptatifs, les premiers, en utilisant des modèles de mots complets, assurent un niveau de précision supérieur aux seconds qui utilisent des modèles basés sur les phonèmes, c'est-à-dire des sons individuels dont l'association forme les mots.

L'adoption de Vocollect Voice par plus de 300 000 utilisateurs dans des dizaines de pays s'exprimant dans différentes langues doublées de spécificités et accents locaux encore plus nombreux, constitue la meilleure preuve du succès de l'approche mono-locuteur suivie par Vocollect.

L'acceptation globale de Vocollect Voice par plus de 300,000 utilisateurs dans des douzaines de pays, parlant multiples langues et encore plus de dialectes et accents régionaux, démontre le succès du système mono-locuteur de Vocollect.

En résumé:

1. Les caractéristiques des applications de la technologie vocale en entrepôts plaident en faveur des modules de reconnaissance vocale mono-locuteur au détriment des systèmes multi-locuteurs.
2. Dans une application en entrepôt, le coût minime de formation d'un module de reconnaissance vocale est largement compensé par les gains de performances qu'apporte cette formation.
3. Les économies opérationnelles assurées par un module de reconnaissance mono-locuteur par rapport à un module multi-locuteurs varient entre plusieurs centaines et plusieurs milliers de dollars par opérateur et par an, en fonction des exigences de l'application et des performances relatives du module de reconnaissance vocale.
4. Les modules de reconnaissance mono-locuteur offrent des avantages supplémentaires significatifs dans la mesure où ils peuvent être utilisés par des équipes parlant différentes langues.

À propos de Vocollect

Vocollect, premier fournisseur mondial de systèmes vocaux pour utilisateurs professionnels mobiles, accompagne ses clients dans l'amélioration de leurs performances grâce à la voix. Chaque jour, Vocollect permet à plus de 300 000 utilisateurs à travers le monde de transférer l'équivalent de plus de 3 milliards de dollars de marchandises entre leurs centres de distribution et entrepôts vers les sites de leurs clients. Une équipe mondiale comptant plus de 2 000 partenaires spécialisés dans la gestion de la chaîne logistique (distributeurs et revendeurs) assure la commercialisation et l'assistance technique des solutions Vocollect Voice dans plus de 60 pays et dans 36 langues.

Vocollect VoiceWorld Suite s'intègre en toute transparence aux principaux systèmes de gestion d'entrepôts (WMS) et aux progiciels de gestion intégrés (ERP) dont SAP et supporte de nombreuses solutions informatiques mobiles.

Pour tout complément d'information, visitez le site www.vocollect.com/fr

Vocollect Amérique du Nord:
info@vocollect.com
+1.412.829.8145

Vocollect EMEA:
voc_emea@vocollect.com
+44 (0) 1628.55.2900

Vocollect APAC:
apac@vocollect.com
+852.9612.3708

Vocollect Amérique latine:
latin_america@vocollect.com
+1.412.349.2675

Vocollect Japon:
japan@vocollect.com
+813.3769.5601

Vocollect
#1 in the Business of Voice™

Publié par Vocollect, Inc.
703 Rodi Road, Pittsburgh, PA 15235
(412) 829-8145, Fax (412) 829-0972, http://www.vocollect.com